A Context-Aware Data Mining Method Recommender for Enhanced Insight Relevance: A Case study Approach

PhD Thesis

Constintine Nhundu

Supervisor I: Professor Mo SARAEE

Supervisor II: Surbhi Khan

School of Science Engineering and Environment

January 2025

i. Abstract

Classical data mining approach struggle to produce insights that relates to a practical situation being analysed. Determining an appropriate data mining method is resource intensive and more of a trial-and-error strategy. The thesis proposed method recommender Method Assistant Buddy for data mining method recommendation. It also develops Context-aware Method and Algorithm Selecting (CMAS) approach to integrate contextual factors into Data Mining.

Data Mining is moving towards Context-Aware to relate dataset to situation surrounding it. *Main Objective: Improve relevance of insight in the practical setting in which a dataset is found. Other objective: Aid recommendation of data mining method based on expert knowledge on both subject and capabilities of specific data mining methods.*

Data mining and analysis in general are focused more on the figures or values than the context in which datasets are found. The major issue with the current Data Mining approach is that Analyst struggle to interpret outcomes. Issues also arise when analysis is divorced from the world it relates to, since it will be difficult to scope and design situation appropriate for a model. Established Data Mining approach produce results that are not linked to physical activities in a straightforward way. Novel ways to appropriately scope effort for a situation needs to be developed.

Context-Aware data mining can improve model usability and reusability. This research proposes, Context-aware Method and Algorithm Selecting Framework (CMAS), applying software decision-making strength. The strength of software is based on knowledge sharing, processing speed and continuous improvements. The thesis develops a Method Assistant Buddy to help novice analysts during model design.

CMAS could assist both novice Analysts and Systems End Users. Academics aim to simplify the data mining process. Industry gains from focused effort given limited

resources. This research explores a proposed framework, three use cases and analysis of improvements demonstrated by applying the novel approach.

Three real life case studies were performed to illustrate the proposed solution. CMAS can be applied to medical records where background knowledge is informative, a highlight could be PIMA India Diabetes context. It can also be applied to India Traffic data where lots of stakeholders might have varying context. A look at World Pollution challenges is complex, therefore CMAS can be applied. The domain knowledge and semantics guide context-based data analysis.

The thesis creates a framework that applies contextual factors and uses accumulated expert information to select data mining methods. Evaluation of the proposal has been performed on three real life datasets (PIMA India Diabetes, India Traffic and World Pollution dataset). Analysis seeks to develop an approach, illustrate, and discuss achieved improvements.

PIMA Indian Diabetes was analyzed using Association Rules data mining whose performance was measured based on confidence and conviction metrics. CMAS was applied to India traffic data in which the Method Buddy selected Clustering Data Mining method. The quality of clustered was measured using Silhouette score and Davies Bouldin Index. Last use case World Air and Water pollution was processed using prediction tool Regression. Regression performance was measured by Mean square error (MSE) and Root Mean Square error (RMSE). CMAS has shown that it reduces the complexity of the whole data analysis and improves performance of selected models.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Meaning	Page
CMAS	Context-Aware Method and Algorithm Selecting Framework	5
RMSE	Root Mean Square Error	150
CNN	Convolutional Neural Network	2
TbnISA	Text based Intelligent User Assistant	14
CRISP-DM	Cross Industry Standard for Data Mining	21
DMME	Data Mining Methodology for Engineering	27
DMC	Decision-Making Centre	28
KDD	Knowledge Discovery Database	28
SEMMA	Sample Explore Modify Model Assess	28
VFT	Value Focused Thinking	29
GQM	Goal Question Metrics	29
SP-CCADM	Scenario Platform – Collaboration and Context-Aware DM	30
CASP-DM	Context-Aware Standard Processing for Data Mining	28
DL	Deep Learning	18
CANloc	Context-Aware Next Location predictors	34
CAA	Context-Aware Analyzer	40
DSR	Design Science Research	41

CHAPTER 1 Introduction

1.1 Introduction

Context-Aware Data Mining has attracted academic research effort over the decades based on huge benefits generated. Benefits like insights that have clarity on the subject matter and domain terms. Each activity in the world is viewed in varying context, these factors need to be included in data mining solution design. Context-Aware Data Mining has huge possibilities to be applied on current and future devices. Produced applications also turn to make data analysis more focused and nearer to real-world cases.

By looking at the dataset, experts can build context to use in an analysis process. (Dey, 2000), gave a more generic definition of context as information aiding to typify a situation of an entity. Entity is an object focused by analysis for example a customer or health related images. The biggest challenge being analysis outcomes that are too abstract, needing effort to translate into actionable insight. There are major challenges in data mining related to results that cannot be easily related to the real physical activities.

Data mining (DM) discovers interesting patterns from large datasets. Evolution over the decades has led to more applicable insights. In today's advanced technical world, data and analysis techniques are created exponentially resulting in new challenges.
The advances on methodologies and computer processing power have led research on Context-Aware Data Mining. In this research a holistic view has been taken of the data process journey. Further emphasis has been put on application of Semi-Automated approaches to model development. Data Mining needs to benefit from information technology's ability to store and retrieve information.

Context can help in determining distribution of resources. Relationships between contexts need to be measured to decide to include or exclude them in any given analysis. Context-aware system (Malik, Mahmud, & Javed, 2007) understanding situations through collecting pointers, gathering them, and finding a way to represent the findings before initiating some adaptation.

Real life situations present varying levels of complexity, working on medical images, (Tuget, Binh, Quoc, & Khare, 2021) extracted the context local features presented in code-word using (CNN), Convolutional Neural Network. Environmental factors may influence feature behaviour, example of context is location, weather, time season to form situation. Therefore, analysis may look at relationships between Location Context, Weather Context to Season Context. Another analysis may look at the whole dataset in the view of Time Context. Based on this perspective, context can work as a subject of analysis or a base. The context-aware domain brings dynamic perspective to data analysis.

Data mining uses several methods that are grouped in classes like Classification, Association, Clustering. During model design analyst select a method to apply. Information or knowledge on data mining methods can be put in a database. Using software one can match a situation, data features to a narrow range of methods. Once the range is reduced human intuition can be used to pick the final method to apply.

At higher level, an analyst needs to select data mining method most appropriate for issue at hand. Each method is designed to resolve issues of specific characteristics. The model needs to match dataset characteristics, objective of analysis and situation in which the dataset is found. When these considerations are not made the results can be both inaccurate and not in line with the physical world being investigated.

Context-Aware data mining starts from understanding an issue moving all the way to interpretation of results. Interest has also been gathering on selection of optimised machine learning algorithms for a given context mix. Decision on best algorithm entails use of performance metrics, which will be explored in this research. Outcome of these model is linked to the real-world processes which aid interpretation. Using Ontological approach (Singh, Vajirkar, & Lee, 2003), was of the view, context increases effectiveness by reducing input going into data mining model. The data reduction originates from pruning to address a focused situation. Development of innovative frameworks also considers data models performance.

The Data Mining process is resource intensive therefore, academics need to find methodologies to develop lean, appropriately scoped and analyses focused for every project. There is an interest in conducting research on practises and logic used in manual systems. This research should resolve a chain of key decisions by teaching machines to mimic human actions. Solution will be helpful in decision making made by the Data Scientists on applying Context-Aware concept. Semi-automation of this knowledge-based step should improve both the model quality and results that are getting generated. Text based Intelligent User Assistant (TbnIAS), developed around Natural Language analysis of problem description (Zschech, Horn, Janiesch, & Heinrich, 2020) had challenges around processing of changing domain knowledge. Looking at key decisions like data mining method selection being assisted by a database of related knowledge. The knowledge database can be adapted into an open-source repository.

Dataset can have high dimension which reduces algorithm performance. High dimension also results in overwhelming volumes of insights which makes interpretation challenging. Feature selection techniques rank features on relevance taking top few (Remeseiro & Bolon-Canedo, 2019), and feature extraction combines features as a way of dimension reduction. Context-aware feature selection is like a traditional approach, save for it focusing on the context-features. (Chen & Xia, 2021) created feature selection around contextual data by ranking on relevance and a weighing system. While removing context redundancy the approach also resolved the data sparsity challenge.

Development of Data mining solutions are characterised by strategy competitions between rigour and relevance. Relevance (Pechenizkiy, Pauronea, & Tsymbal, 2008) being the utility of resulting insight on organisation or domain.

1.2 Problem Statement

Problem is selecting appropriate data mining method for analysis challenge at hand. The current state of the art Classical Data Mining approach using trial and error approach based on limited human mind referencing of past solutions. The other problem is production of data mining results that are not easily linked to the situation in which the dataset is found. This increases cost of effort to interpret results in a way that will allow to translate to practical solution. The link to environment has a scoping effect reducing resources required throughout the analysis. Domain thematic are likely to increase insight usefulness to end user or stakeholders. Of interest to academic research could be whether the appropriateness of the method and relevance of results to situation be measured. Can a comparison be performed between Traditional / Classical approach and Context-Aware data mining method Recommender be performed.

Data analysis need not produce knowledge abstracted from the real world. Analysis is meant to produce insights that can be easily applied to resolve issues. Making data mining solutions that can easily relate to the real world. The main problem is that established data mining approaches do not consider the environment where datasets are found. Leaving out the influence of the environment turns to make resulting insight too abstract thereby distancing itself from the real world. In instances where the environment or context affect the dataset, these effects are lost to the analysis.

Data mining needs to incorporate environmental factors (Context) without increasing process complexity. Context perspective encourages combining objects or attributes to create a view related to the physical operations of the world. In business activities, Customer Recruitment Context focuses on the reduced aspect of a wide business field of Income Generation.

1.3 Motivation

1) Need for situation focused data analysis that produces insights aligned to specific real-life aspects. Data mining must avoid abstract and overwhelming results since that reduces insight relevance.

2) Simplified, technology and expert knowledge-based recommendation of appropriate data mining methods. The approach attempts to replace the current resource intensive trial and error approach.

1.4 Research Question

Question one

Can novice Data Analysts increase insight domain relevance, quality, accuracy, domain informed and right relationship with real world by designing data mining model premised on context-aware approach? Could this model be appropriately scoped based on expert knowledge and easy to apply to real life challenges?

Question two

Are there ways to better manage data mining resources? Can we trim and focus analysis effort based on expert understanding of the context surrounding a dataset?

Question three

To what extent can we optimise data mining process and use most appropriate methods by using domain expertise based Semi-automated methods selection software?

1.5 Aim and Objectives

1.5.0 Aim

Research aims to improve quality of data mining insights through incorporating contextual factors in the analysis solution. Also explore the possibility of improving the quality of data mining insights produced by using Context-Aware principles. Insight quality being subject area relevance, focus and any process simplifications.

The thesis also intended to investigate potential applications of software components to select a suitable data mining method to apply on a particular challenge. Effort is made to investigate use of accumulated knowledge in the selection of Data Mining methods.

1.5.1 Objectives

1) Develop a Context-aware Method and Algorithm Selecting Framework (CMAS)
    a) Outline the framework.
    b) Summarise its key characteristics and differences from another framework.
    c) Details its benefits (Model quality, insight domain relevance and effort optimization)

a) Challenges with the new framework.

2) Create Semi-automated Data Mining Method Selector

a) Develop logic to select appropriate Data Mining Method for given analysis challenge.
b) Illustrate the logic by pseudocode, pictures, or diagrams.

3) Evaluate CMAS and Method Selector based on three use cases

a) Search for three real life challenges (Indian PIMA diabetes, Indian Traffic, and World Pollution analysis. The cases cover health, social and climate change issues, these being important to human life and affected by context. They should allow handy testing of the proposal.
b) Describe each of these datasets in domain term and context considerations.
c) User Method Selector to suggest the method most appropriate for each situation.
d) Perform end to end data analysis using the CMAS framework.

4) Validate CMAS and Data Mining Method Selector based on three use cases

a) Validate increased level of domain sensitivity of output from the three use cases

b) Validate resource savings due to focussed effort resulting from the framework

c) Validate easy of interpretation of insights from the CMAS solution

1.6 Hypothesis

Implementing Data Mining with considerations of contextual factors improve the quality of the resulting insights. Quality taken as to include relevance to the physical world, focus, ease of interpretations and level of insight applicability to solution development. Therefore, predict that Data Mining performed in a Context-aware approach will result in high quality insights.

The Data Mining process could be aided by software applications that select appropriate Data Mining methods for a given situation. Therefore, predict that a Data Scientist using

software to select a Data Mining method will find the process simpler and produce a model that reflects years of expert's knowledge.

1.7 Context-Aware CMAS

According to Oxford dictionary context detail circumstance informing or increasing understanding of the setting. Context could be taken to be the same as surrounding factors, conditions, environment, atmosphere, situation in which a subject is found. In data mining subject is the dataset, more specifically attributes and features of the dataset.

Cambridge.org defined context as situation in which a subject exists or happens it explains the subject. Common to these definitions is where the context relates to subject and that it gives more information about the subject. Working to develop complex information-based manufacturing models (Scholze & Barata, 2017) defined context as 'information that can be used to characterize the situation of an entity.

For this thesis context is in relation to a data mining dataset as subject. Thesis to explores effects of environment in which the subject is found. In this case the effects of environment to the patterns or insights produced by the Data Mining effort. Could understanding or results of analysis have been different if the environment was to be changed?

Context-aware expand to analysis that factor in environment or situational information to allow better understanding of the dataset. Over the years emerged Context-aware computing due to devices being portable and able process data from the surroundings. GPS network is a good example taking in locations etc.

Let us look at context as it relates to the three case studies in this thesis. Looking at the PIMA Indian diabetes dataset the context is the two views that covers the World Health Organisation understanding of diabetes and established biological attributes of PIMA Indian as it relates to diabetes. Insights from the Context-Aware Data Mining approach should be more informative and accurate for the PIMA population. Is medical advice more helpful when factoring attributes of a unique population?

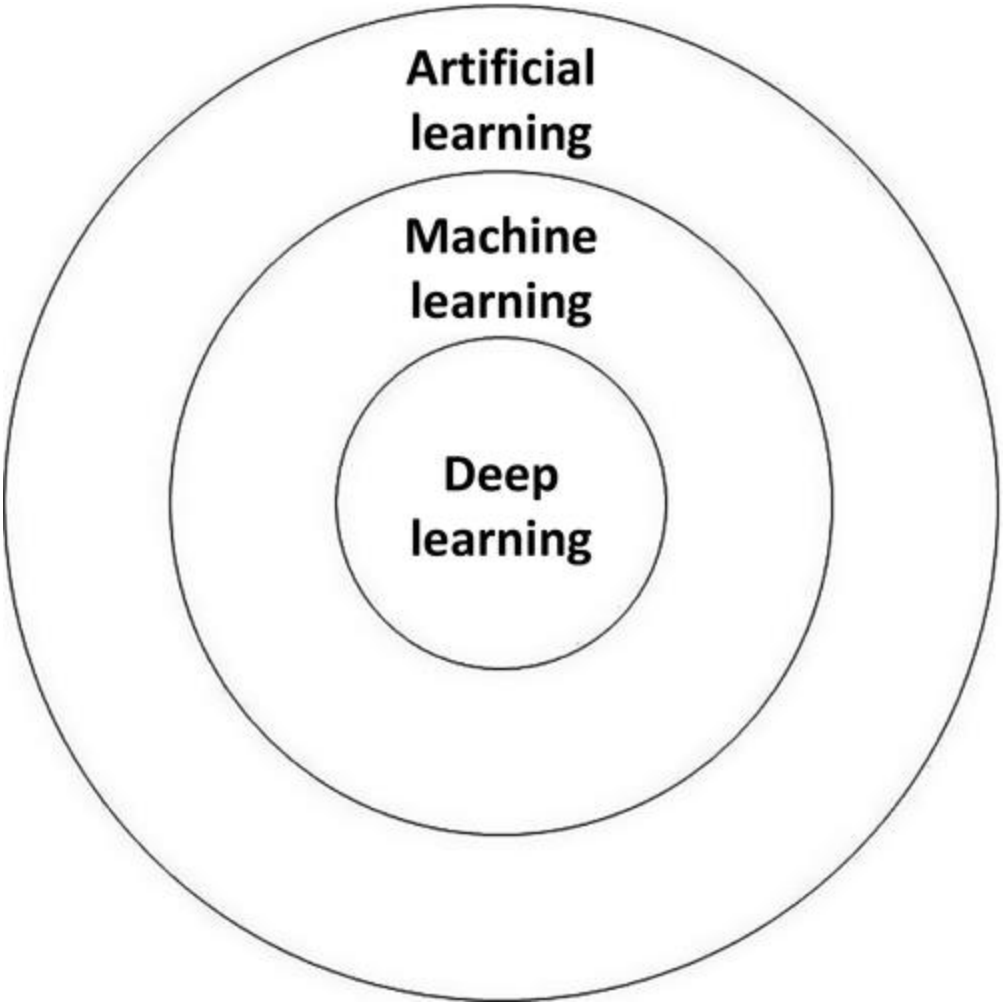
Analysis of Indian Traffic was analysed in the context of complex Indian population density and infrastructure development. It was also looked on the perspective of many possible stakeholder. Context answers questions like given data with 32 attributes would one sub-scope when trying to answer worries of different stakeholders. Context-aware

analysis may be resource conservative if the analyst bases their work in perspective of say one stakeholder e.g. a government team responsible for road infrastructure.

On the World Pollution case study three data sources were used to give a better understanding of the pollution. The subject was pollution, but full context was types of pollution and effects being death figures linked to pollution. Data mining could be performed on a given dataset or bring a given dataset at other environmental factors like death rates related to pollution.

1.8 Deep Learning Method

The proposed work is Context-Aware data mining using a Data Mining method recommender. Deep learning is a Data Mining method that may be recommended by the solution on a given problem. Context-aware relates to the analyst accepting that the local, main datasets or entities exists in a surrounding that may affect behaviours and relationships. Whereas Context-Awareness is at a high level of attempting to resolve a given challenge, Deep learning is a model that given input and output can learn to process future inputs. If during mapping of a solution Deep Learning is selected by the Recommender, it is then likely to have contextual features as part of the input used to train the network.



Deep learning as a subset of Machine Learning which is a subset of Artificial Learning

Figure 1:Deep learning as a subset of Machine Learning which is a subset of Artificial Learning

Deep learning is Artificial Neural Network ANN (Jakhar & Kaur, 2020) solution with many hidden layers. Design following the pattern learning of biological brain the more hidden layers the more complex problems the network can resolve.

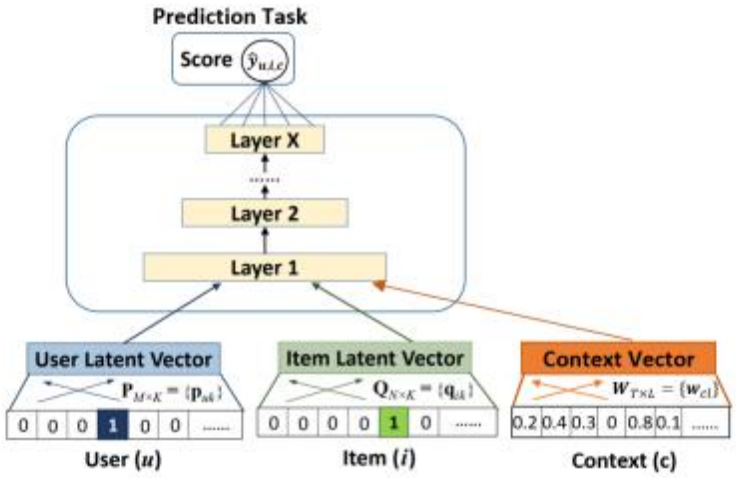


Figure 2: Basic Logic of Deep Learning

Context-awareness is usually implemented by adding contextual features (Unger, Tuzhillin, & Livne, 2020) added time, user activity and location, Deep Learning is one of possible models to analysis the assembled data. As said before the deep learning capabilities are increased by number of processing layers. As shown on (fig.2) the three datasets user, item and context become input to the deep learning network.

Context in context-aware data mining is premised on context information being not the same as the standard features of a data set due for analysis. Context is situation defining information that may even be collected in a manner that is variant to the standard input data collection one.

Deep learning is static without any environment characterizing data, context (Mijnsbrugge, Ongenae, & Van Hoecke, 2023) integrated Context to deep Learning networks to produce more dynamic and informative insights. Context-aware Data mining can apply any model including Deep Learning, main distinguishing characteristic being the incorporation of environmental factors in the analysis.

Table of differences between Context-Awareness and Deep Learning

Aspect	***Context-Awareness and method Recommendations***	***Deep Learning***	***Comment***
Breath	*Covers the full analysis cycle from getting an opportunity to generating insights*	*One stage of the life cycle that os model development and application*	*Deep Learning could be one of selected DM methods for an analysis*
Main principle	*Consider surrounding in designing an analysis solution. Also consider characteristic of Data Mining Methods to recommend the one matching the challenge defined*	*Given an input and output learn the functions to allow for prediction of future input's outputs. The predictions is done by logic in layers that mimic biological brain was of processing information*	*One if an option within the other. When we say Deep Learning with Context-Awareness, it's more of saying context is being treated as one of the inputs to the logic*
Data Mining Method Selection	*YES*	*NO*	
Context Consideration	*YES*	*NO*	
Logic	*Selection and Context applied throughout the cycle*	*Use a network with layers to predict output from input*	*Deep Learning has advanced software development towards Artificial Intelligence*

Table 1:Table of differences between Context-Awareness and Deep Learning

1.9 Real life application of the proposed CMAS solution

Three datasets were looked at with view of resolving technical challenges and illustrating logic behind the proposed approach. Easy, quick, and expert knowledge-based selection of appropriate Data Mining Method. Method selection need to be knowledge-based decision is best left to software and computers given their ability to store and search knowledge at high speeds. The human practitioner is best able to deal with unexpected situations. Method selection may be best handled by combing software help and human input.

Data mining analysis that factors in the Context information. The resulting insights are likely to be easy to transform into actionable position given the alignment of analysis to the domain thematic.

PIMA Indian diabetes case study

Medical background and expertise is needed to relate figures, value of attributes in a medical disease data analysis. Knowledge specific to a population at hand PIMA Indian as they relate to the diabetes conditions. Further breakdown of features or attributes is necessary to understand medical meaning of range of values. These categories allow Medical Practitioners to formulate advice to patients.

Data Mining insights turn to be divorced from the real-world in which a dataset is found. The thesis attempts to resolve the gap by incorporating Contextual data into the Data Mining process. Insights maybe too abstract, with no clear relation to Environmental factors. Insights should be personalised or situation referring to be more relevant to stakeholders. In this case study Medical Practitioners understand diabetic readings categorization, using them to develop advice to patients.

Main Challenge: To create an analysis model that considers the characteristic difference of a PIMA Indian Women as they relate to the diabetes disease. Finding ways to incorporate World Health Organisation accepted knowledge on the worldwide population and a varying expertise knowledge specifically for PIMA Indian Women. The second challenge is in making selection of Data Mining Method to use which have abilities that align with the problem at hand and expected analysis.

Given methods like Association, Clustering etc which can use the available dataset to produce usable, sensible results. Looking at types of data each method takes, logic it

uses to retrieve patterns and form of results it produces. For the PIMA dataset one needs to produce insights that will inform a person in the medical domain.

Data mining method accepts particular type of input, process it and produce results that can be visualised to provide insights. In this case one needs a Data Mining Method that takes in the eight features and produce medically informative insights.

The challenges is that if the above decision is made by a human being, they need to dig into their knowledge of methods and ways they have resolved similar problems. Proposed solution is to store historical knowledge of abilities and results that each Data Mining method are best on and have a script that matches problems to methods. This approach is both fast and knowledge based without the human experience limits especial for new analyst and end users.

Indian Traffic Case Study

Indian Traffic dataset is overwhelming looking at the large number, possible groupings, possible interactions of features. The wide range of data and characteristics can complicate understanding of issues at hand. Indian traffic is unique from other countries because of population volume, favoured mode of transport, accepted expertise knowledge and requirement from a variety of stakeholders. Analysts need to decide on standalone Data Mining methods or possible combinations the have increased chances of mining useful insights. Researchers should develop ways of matching characteristics of Indian Traffic Data to capabilities of Data Mining Methods.

Appling Traditional data mining approach would mean undertaking trying many methods before selecting on to apply. Based on the proposed approach the Method Select Assistant would reduce candidate to a few. Manual selection from the few methods would likely require less human effort. Environmental considerations would include cultural ethos, levels of safety Assistant would reduce they campaigns etc which need to be integrated with local data. Analysis should produce insights that can be translated into usable knowledge for specific stakeholders for example Local Government Officers. Each set of stake holders is unique in their focus and insight of interest.

Indian traffic is a integrated interaction of many factors, entities, relationships that are complex. Understanding various contexts and creating practical insights is challenging.

Decisions on the best data mining Method to produce appropriate results is overwhelming. How to produce insight that is at necessary level for interested stakeholders like government departments and safety knowledge campaigners.

Data Mining Method selection need understanding of all 32 features of the dataset, clear knowledge of aspects that once analysed create results that make sense. The process also needs to match known abilities of the methods to handle features and situation at hand. In a traditional setting on could search past analysis that has been perform on similar dataset to select previously used methods. This approach is time consuming and limited in volumes of past historical cases that a human being can refer to before making design decision.

Indian Traffic dataset was extracted from a special country with country specific characteristics that affect the features in different ways. In each Context is considered how much of background on each context is required to make appropriate use of data on each feature. Contextual considerations include finding what is important for a given stakeholder. Could it be more informative to perform analysis using a subset of features that affect area of interest to a given stakeholder.

Context based analysis result in focused Sub model that encourages deeper understanding and insights that are tailor made for a given audiance. In case of the Indian Traffic case, what informs Local Government Officer does not equally inform a Road Engineer as their interests vary. Context or situation considerations pushing to investigate whether a similar dataset from another country would bring back different insights. Are there any cultural behaviour issues? Could insight be affected by awareness investments made by a country in which dataset is found. Context-Aware Data Mining attempts to go beyond feature, attributes, and statistical calculations.

World Air and Water Pollution

Main challenge is to know if a supplied dataset tells a full story on pollution. In this case search for more data was preformed to create a picture of full context. It is not always the case that an original dataset covers the whole situations. Pollution issue, characteristic of data guide selection of Data Mining method that has chances of producing meaningful insights.

Academic researchers are focussing on linking datasets, data mining methods, patterns and insight to produce rich practical decisions. Knowing levels of pollution related death by location may encourage localized interventions. Water and Air pollution could be linked to population density, Industrialization, and water treatment infrastructure. Can action be taken to rectify an identified negative situation.

1.10 Contribution to knowledge

The thesis builds on the general world standard CRISP approach to data mining and specifically to development of Context-Aware Data Mining (CADM), Context-Aware Standard Processing for Data Mining and Context-aware Attention (CARE) which focused on Context-Aware data mining. The knowledge gap remains in applying Context to any stage of the Data Mining Lifecycle and use of technology to recommend Data Mining method for a given situation. The thesis attempts to add flexibility and automation element to the incorporation of contextual factors.

Context-aware data mining can constrain analysis making it more focused. Other researchers have emphasised on resulting insights having improved domain semantic thereby easy of human understanding. Issue: Most insight from Data Mining cannot be directly linked to real-world situations. Considerable effort is required to make the insight usable by domain experts. The effort is usually to give insights thematic of domain of a given dataset. Environment turns to affect features or their relationships. Data mining process needs to incorporate these effects to be more useful and more relevant to those that may need to infer for understanding and application to practical solutions.

Selection of an appropriate data mining method for a given challenge is difficult as it is currently based on human being searching for historical applications in each situation. A method recommender based on key aspects of a dataset and established abilities of methods could both streamline and simplify this process. Data Mining insight relevant to situations are more usable. Design of the solution that incorporate environmental factors reflect a real world in which dataset do not exist in isolation from surroundings.

Can data mining process and outcomes be improved by using software to select appropriate method and considering effects of the surrounding to the patterns found from the dataset. Classical/ Traditional Data Mining has challenges of producing overwhelming, abstract, unfocused and insights that are not algin to situations, it also has a cumbersome way of selecting appropriate method that suit a situation at hand. Theses develop a Context-Aware Data Mining approach that use a software-based method recommender.

Improve quality of insights

Data mining over the years has concentrated on finding patterns in a dataset without consideration pf the environment in which dataset exist. The problem is that this approach produces results that are too abstract. Data being the new oil, many novel Analyst and users are contributing to building data models. Any assistance from software to streamline and simplify the Data Mining art will improve and speedup entrance into this field by more professional.

Work needs to be undertaken to relate results to real-world situation. Context-aware data mining encourage including contextual features from understanding issue at hand to interpretation of insights. Development of Context-Aware and method recommender could produce situation specific insights from data mining effort. Environment affects features in a dataset. Data Mining method selection is currently a challenge since one might end up tail and error a few times before completing analysis with one of the methods. Selection of methods is time consuming and may lead to wrong reduction of quality of insights.

Insight incorporating situation are likely to be more meaningful to domain experts. Physical world is made-up of interacting eco-system, this reality needs to be reflected in the data analysis. Develop insights. Costly guess work in selection of appropriate data mining method to apply to a given dataset in a specific environment.

Patterns can be too abstract necessity costly effort to related them to a human level situation that can help deriving a practical solution. Selection of appropriate Data Mining method for a given situation are resource costly and currently a hit and miss undertaking. Insights not clearly linked to the environment or domain principles in which the dataset is

found. Interpretation challenges and huge volumes of results from which one must. Data Mining has challenges of not producing results directly linked to situation, environment or domain where the dataset if found. Whenever it undertakes a data mining implementation, they perform a trial-and-error act to select appropriate Data Mining method.

Thesis attempts to find ways to integrate environmental factors and create Data Minig Method knowledge in deciding appropriate method to a given challenge. This will apply Method Recommender and Context-aware application to Data mining process to produce domain appropriate models. Theses apply a Recommender to three real life study cases PIMA Indian Diabetes, Indian Traffic and World Pollution. We then scientifically measure level of insight importance and alignment to the Domain.

Will the insights from analysis of PIMA Indian Diabetes be meaningful and easy to interpret for a Health Experts? Are the insight in the right level of medical terms. Does the insight highlight or align to the two perspectives on disease and local population. Results interpretation and volume relevance to the situation and domain or expert perspective. Analysis relationship with real-physical world from which a dataset is extracted. Concept of the immediate environment surroundings the dataset. How can data be analysed considering effects of the immediate environmental factors.

Target audience of the thesis

Thesis would help student of data mining and junior analyst by assisting in evidence-based Method Recommendations. It would also assist in producing Context sensitive results which are low in volumes and high in situation relevance. For established data Scientist the thesis application would save resources by being focused on a specified situation. For users of insight the thesis would help by providing insights that are in perspective of a particular challenge. Thematic will be in line with experts of a given field. For example, the medical scenario (PIMA Indian) results are in diabetes terms and attention to a given patient cohort.

The Indian Traffic analysis results are focused on several contexts. They would allow a stakeholder to further analyse their context of interest. The uniqueness of the Indian Traffic ecosystem makes the inherent patterns vary from traffic analysis of other countries.

Once a method Recommender and Context-Aware approach.

1) There is a knowledge gap on applying Contextual factors to analysis of datasets. This research contributes to the use of environmental or contextual factors as the base for analysis of data. Past research has concentrated on finding patterns in dataset without considering the context. That approach usual produce overwhelming results which require high level of resources to produce sensible interpretations.
2) Second contribution is on using technology to assist in the selection of Data Mining Methods appropriate for specific challenges. Over the last decade academics research has focussed on automate machine learning covering selection algorithms not the methods. The thesis contributes to knowledge on implementation of data mining solutions that use the context to increase real-world relevance of resulting insights. There is limited research done on combining Context concepts and semi-automated method selection.

Research papers	Context-Aware Model Design	Context / Feature Engineering	Algorithms Class Selection	Algorithm Performance Analysis	Process Automation
Tuget etal, 2021				X	X
Bhadane, C & Shah, 2021	X	X			
Van Houd, 2020	X		X	X	
Martinez-Plumed etal, 2019	X			X	X
Huag etal, 2018		X	X	X	
Lange, 2017	X				X
Nallaperum etal, 2017	X	X	X	X	
Amir etal, 2016		X		X	
Dhanesshwar & Patil, 2016	X	X	X		X
Tomar & Ayarwal, 2013				X	X
Osei-Bryson, 2012	X		X	X	
Gilbert etal, 2012		X		X	
Chitguakar etal, 2005	X	X	X		
Avram etal, 2000	X	X		X	
This Research	X	X	X	X	X

Figure 3: Contribution to knowledge, past research work.

As shown in fig.3, past research has ranged from incorporating context knowledge to applications, data mining processes and selection of machine learning classes. In this research contribution by investigations will be done on simplification and automation of multitude of decisions made in Context-Aware designing models. The work includes performance, quality, and lots of other measurements to select components of these dynamic models. The research will contribute through exploring decision making in model building.

Thesis has developed novel framework CMAS, applied it to three use cases PIMA Indian Diabetes, Indian Accident Severity and World Pollution datasets. Showing the integration of data, Context, data mining method characteristics to produce domain scoped insights.

Research structure.

The thesis is structured as follows:

Chapter 1, ‘Introduction’ discusses challenges facing novice Data Analysts. It covers motivation of the project, problem, aims, and objectives. It positions the research within the wide search for knowledge.

Chapter 2, ‘Literature Review’ examines development in research on created data mining related frameworks. Chapters detailing previous knowledge researched on Data mining process, context-aware concepts, decision making systems and algorithm selection. Theory of comparing data mining frameworks is also reviewed. Detailed look at the likely challenges and background of the sample data used for illustration.

Chapter 3. Research Methodologies. Academic work has agreed methods of research, this one being a mix of solution development and knowledge accumulation takes a modified approach. Here we highlight and connect the dots to explain the approach used.

Chapter 4 ‘Proposed Framework/Model’ details the proposed functions from coding, gathering of parameters and validation of use cases. This will also expand on the logic of making design decisions.

Chapter 5 ’Experimentation/Case Study’ illustrates the model using a few datasets. Logic and data mining of PIMA Indian, Indian Traffic and World Pollution problems.

Chapter 6 ‘Concept Analysis, Evaluation and Discussions’, evaluates the proposal using several approaches and use cases and finally. A look at outcomes from the solution suggested in Chapter 4 and implemented three times in chapter 5. It attempts to evaluate if the research has met its objectives.

Chapter 7 ‘Conclusions’, is the conclusion and a pointer to future research areas as they relate to work covered in this report.

CHAPTER 2 Literature Review

Chapter 2 Introduction

The chapter explores, critic past research papers on Context-Aware Data Mining from varying perspectives. These perspectives include increasing insights quality, ways of incorporating situational factors and making method selection techniques efficient from options including Association, Clustering etc. The following past research review will follow history of the development of data mining concepts up to the current interesting and rewarding topic of Context-Aware Data Mining frameworks. This research focuses on the domain of Context-Aware Data mining as it relates to solution development.

A survey including previous research on machine learning class comparisons and design logic is undertaken with reference to Context-Aware domains. Past research work is analysed, discussed, and grouped to further knowledge on Context-Awareness. Data mining frameworks developments are explored in the wide domain of Data Mining Science in, 2.1 “Context-Aware Data Mining Evolution.” Subsection 2.1.1 “Current state of the Art on Data Mining Frameworks” is a recap of past literature on established ways of analysing large datasets. 2.1.2 “Context-aware Data mining concepts” explore the growth of these environmental centric analysis. Finally, 2.2 ‘Context-Aware Machine Learning Class Selection” covers the interesting factors on processing, mimicking, or understanding how human beings make design decisions. Decision making investigations are limited to logic around selection of Data Mining methods.

2.1 Context-Aware Data Mining Evolution

The following review looks at building of CRISP-DM guidance. It then goes on to look at improvements likely to come out of application of contextual information and automated recommendation of data mining methods. Context happens to give a more complete picture of a given situation. In a Context-Aware Data Mining (CADM) effort looking at soil moisture and interval-based temperature as base dataset (Avram, Matei, Pintea, Pop, & Anton, 2020) included location-based temperature as the context. In a classical data

mining, they only used the base/local dataset with interval-based temperature and soil moisture which produced some patterns but less insight quality. Avram et al went on to investigate the effect of Context data quality on overall results of a Context-Aware analysis.

Academic researchers have focused on many concepts to increase understanding and application of context in data mining. (Zhu, et al., 2013) investigated context independence and dependency proposing that Contextual Feature can be Day, Location and Time range. They noted that added location specific context to Classical data mining approach to improve model efficiency and enhance results. In line with technological advancement over the years allowing machines to generate more informative situation-based data. During a Mobile Phone based study (Vu, Phan, & Phan, 2022) acknowledged that data can allow context-based behaviour, situation and context understanding. Context-aware concepts are being enriched by availability of data on subject surroundings.

Subject specific and localised (Cabri & Nocetti, 2024) data make insights from a Context-aware solution more accurate and applicable to decision making. The development of Context-Aware approach needs ways to gather expert domain knowledge into Data Mining features (Hollmann, Muller, & Hutter, 2023) applied Large Languages Models (LLM) in the Context-Aware Automated Feature Engineering (CAAFE) approach.

Context-Aware Data mining brings together a variety of data sometimes need an ensemble (Cabri & Nocetti, 2024) used Classification and Clustering Data Mining Method to process context like road type and daylight. Of interest could be the amount of effort applied to the selection effort and level of missed opportunities of better performing Data Mining methods. The human effort required to select an algorithm based on trial and error to select Gradient Boosting Classier (GBC), Convolutional Neural Network (CNN) and K-Means is huge, so researchers need to improve this task.

Success of analysis depends on deeper understanding of contextual information (Tong, Zhang, Qiao, Wang, & Wang, 2024) used relation score encoder, weights approach and context labeling in their CARE Context-Aware Attention interest REdistribution. Human based expertise process uncertain situations very well. (Yaghtin & Mero, 2024), they can

complement strength of Machine Learning techniques during development of dynamic personalized solution. Research is working to produce easily applicable insights from Data Mining Activities. Wide ranging expertise and domain knowledge need (Papageorgiou, Sarlis, & Tjortjis, 2024) to be applied in the design of complex Data Mining solutions to produce insights that have practical applications.

(Onim, Thapliyal, & Rhodus, 2024)Established a 0.04 increase in F-1 Score and 4% improvement in accuracy due to adding Contextual information in Machine Learning analysis of Stress Biomarker data. Current technology has enabled the world to collect large size of data with increased complexities of application areas (Theodorakoulos, Thanasas, & Halkiopoulos, 2024) thought cross-functional and tailor designed Data Mining methodologies. Data can be complex and multi-dimensional for example in Climate and Deforestation Data Analysis, (Temenos, et al., 2024) increased insight applicability and usability by applying Context-Aware Adaptive data Cube (C2A-DC). Complex applications like rail maintenance brings lot of data challenges (Steenwinckel, et al., 2021) developed FLAG (Fused-AI Interpretable Anomaly Generation System), using Context-Aware approach to reduce false positives due to high levels of situation-based information.

Sarker is of the opinion that success of machine learning solution depends (Sarker, 2021), on characteristics of data and selection of appropriate algorithms. That study focussed on smart and intelligent real-world application. Data mining process success can be measured on insight quality which includes ease of interpretation. Another insight quality of interest includes ease of interpretation. Below (Fig.4: Data Mining Framework Taxonomy) is a taxonomy of model building blocks researched in the past.

Proposed Novel Data Mining Framework Taxonomy

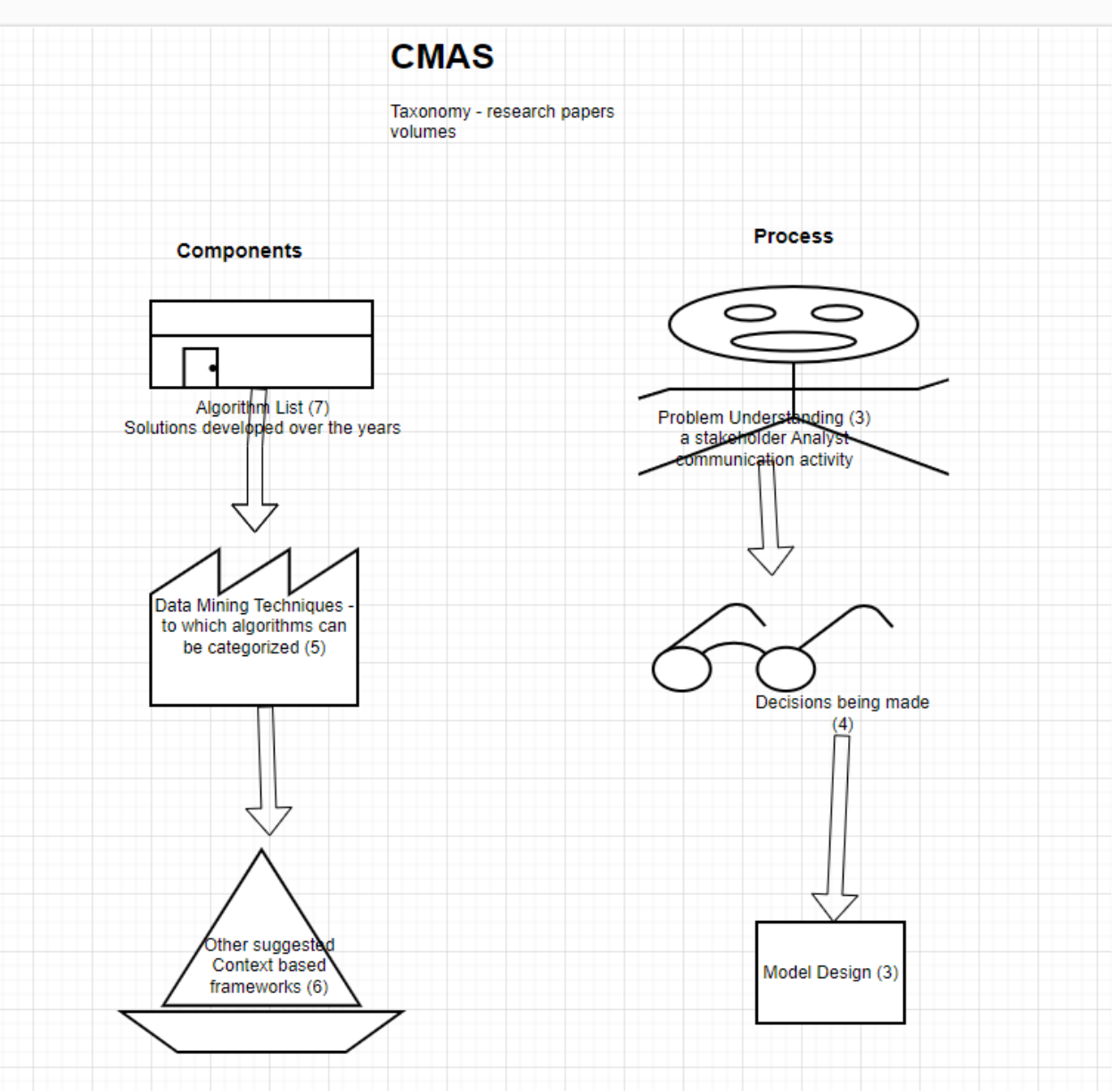


Figure 4: Data Mining framework taxonomy.

The design of the proposed data mining framework (CMAS) is informed by a range of past articles from a wide range of topics. The research can be looked at as having processes and components that build the framework as shown in (Fig.4), taxonomy.

A framework can be looked at from process perspectives, like Problem Definition, Artificial decision making and Model Design. It is constructed by putting together a few components like a Collection of Algorithms and Data Mining techniques. A framework is a holistic assembling of interacting activities. The literature of these processes and components informs the design of CMAS frameworks.

2.1.1 Current State of the art on Data Mining frameworks

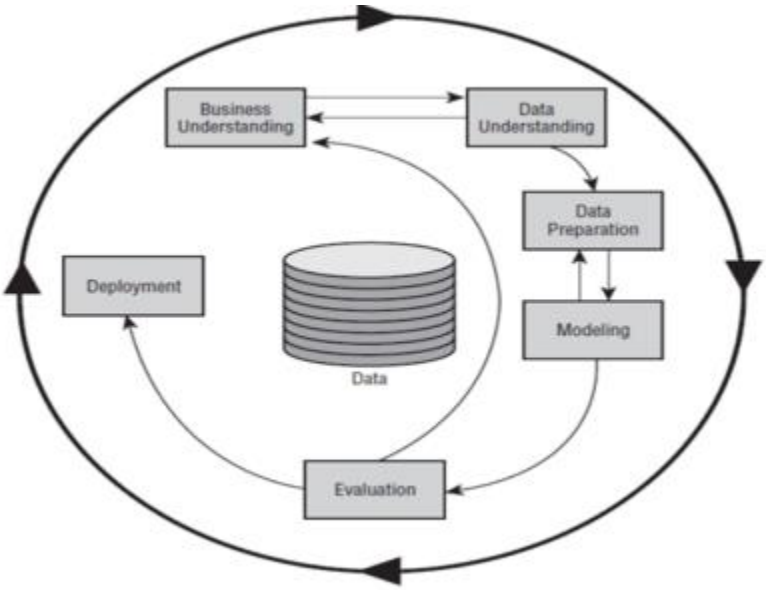


Figure 5:Steps followed in most data processing projects.

Most Data Mining projects attempt to cover the fundamental processes above (fig.5). This steps-based cycle called CRISP-DM (Hotz, 2023), progress in knowledge has foundation in understanding problems. The six basic processes were agreed through extensive studies by industry and academia. Above basic approach enables application on a wider range of situations. This research is advancing or contributing to ascertained Data Mining models through a highly focused approach that accepts environments to be of immense influence patterns in the dataset.

Complexity of data mining solutions design has increased over the last three decades due to volumes of data and wider application of analysis. Cross Industry Standard Process for Data Mining (CRISP-DM), illustrated in fig.5, provides framework of phases, tasks with checklist and high-level guide (Wirth & Hipp, 2000) on ways of completing each of these data mining components. CRISP-DM is at an extremely high industrial level forcing developers to implement situation tuned versions of the principal methodology.

A framework is the structure built to perform a given task. Building Context-Aware or Adaptive framework, (Sarker, 2021) may be around contextual data like environmental, spatial temporal or social context making it relevant to the domain and issue at hand.

(Silva, Saraee, & Saraee, 2019) promoted a Visual Data Mining framework to increase data analysis competence on mental health data, as a concept sharing vehicle. Adaptation of a conceptual framework affords a clear direction for the Data Scientist effort. Visual Data Mining becomes a communication tool positioning data analysis factor, work scope, output evaluation measures and any accepted limitations to drive the analysis effort.

CRISP-DM guide analysis process (Think Insights, 2023) from understanding a given challenge, data wrangling, model design, implementation, and evaluation. CRISP-DM also has a looping logic to encourage modifications during development. The proposed framework adds a context perspective to every analysis stage in CRISP-MD. The proposal evaluation should focus on increasing performance of a modified CRISP-MD.

Data mining can be applied to a wide range of domains. The domains vary in complexity, hence there exist issues with a one-size fit all concept in data mining. Development of domain focused data mining solutions has enjoyed increased industrial and academic research effort. Data Mining Methodology for Engineering (DMME), (Huber, Wiemer, Schneider, & Ihlenfeldt, 2018) added to CRISP-DM Technical Understanding, Realisation, and Implementation phases to modify towards an engineering problem. DMME helps in documentation and creating a clearer workflow.

(Escobar, Espinosa, Espinosa, Monroy, & Solar, 2019) mixed approaches with focus on process reverse engineering in mind combined Decision-Making Centres (DMC) generic,

Arizona State University Decision Theatre process and CRISP-DM to improve data mining process design.

Three data mining models have dominated academic and industrial research namely.

- KDD (knowledge Discovery Database), (Shafique & Qaiser, 2014)
- CRISP-DM developed Daimler Benzes, SPSS, and NCR around 1996
- SEMMA (sample, explore, modify, model, assess) from SAS institute.

The strategies KDD, CRISP-DM and SEMMA have nine, six and five steps, respectively. The study covers the two KDD phases, Data Mining task selection and choosing algorithms. These two steps, tool selection make-up the start of the modelling phase of SEMMA and CRISP-DM.

2.1.2 Context-Aware Data mining concepts

Integrating business context information into the design of a data mining solution has lots of challenges, (Osei-Bryson, 2012) applied Value Focused Thinking (VFT) and Goal Question Metrics (GQM) approaches to evaluate data mining algorithms performance on a specific problem. VFT and GQM link a given data mining situation to standard algorithm performance measures like accuracy etc. The result approach is a cumulative processing of knowledge on past experiences to guide data mining model design effort. Osei-Bryson accepts the importance of context and inherent challenges of integrating it into the data mining process.

Researchers have been improving context approaches from varying angles. These angles, including reusable model design or past knowledge. (Avram A. , Matei, Pintea, & Anton, 2020) improved classical method through combination of past similar, context and current data in the design of situation sensitivity data mining model. They called it, Scenarios Platform – Collaborative and Context-Aware Data Mining (SP-CCADM) approach. CASP-DM, Context-Aware Standard Process for Data Mining (Martinez-Plumed, et al., 2019) added context related tasks and decision points to the standard CRISP-DM methodology. The decisions centred around whether the current context can be resolved by the original versatile model, need the model to be reframed or if a new model must be built. The CRISP-DM, CASP-DM and SP-CCADM approaches improved

on accepted data mining procedure to add the dynamic environment in which a local dataset can exist.

Benefits of using contextual information in Data Mining has led to development of innovative context centred methodologies. (Martinez-Plumed, et al., 2017) extended CRISP-DM by adding context handling functions to all processing steps producing (CASP-DM), Context Aware Standard Process of Data Mining. They noted that CASP-DM increases model reuse, by covering situation changes and handling in a better way than earlier methods. Amalgamation of situation makes the outcome more relevant to the challenges at hand. Improvements are moving away from one solution fit all situations approaches of earlier decades.

There are variations of combinations of approaches to Context-Aware Data Mining. Research and technique development aiming at data mining results need to mirror real life complexity has continued over the last decade. The direction of research interest has been moving from Classic data mining to Context-aware data mining and various combinations. Predicting changing situations (Avram A. , Matei, Pintea, & Anton, 2020) found (k-NN) k-Nearest Neighbour and (DL) Deep Learning, performed highly in a technique combining (CADM) Context-Aware Data Mining and (CDM) Collaborative Data mining approaches. They performed comparison of algorithms in the domain of context sensitive solutions.

Context Framework and Contribution to Data Mining Domain

Perspective	Framework	Reference	Contribution

Environmental, Spatial, Temporal, and social context	Context-Aware or Adaptive Framework	Saker, 2021	Breadth of features
Clarity of Analysis	Visual Data Mining Framework	Silva, Saraee, Saraee,2021	Understanding of analysis outcome
Importance of Context, Awareness of inherent challenges, focus of Data Mining challenges	Value focused Thinking (VFT) and Goal Question Metrics (GQM)	Osei-Bryson, 2012	Focused analysis effort and insights
Past knowledge, past similar, context	Scenarios Platform-Collaborative and Context-aware Data Mining SP-CCADM	Avram, Matel, Pinte & Anton, 2020	Situation sensitive Model design
Added Context handling functions to the CRISP-DM methodology	Context-Aware Standard Process for Data Mining CASP-DM	Martime-Plumed, etal, 2019	Better handling of situation changes
Rank Context features and only work on the high ranked once	CAFÉ-map	Amir, Minhas, Asif & Arif, 2016	Trim elements for processing
Context-aware and DM method selection	CMAS Context-aware Method and Algorithm Selection	This Thesis	Most of above contributions and method selection

Figure 6: Context Frameworks and Contributions

Looking at (fig 6. Context Framework and Contributions) researchers have been adding elements to the standard CRISP-DM, with a summary of their contributions. These advanced approaches have been applied to many domains including, medical information and biological factors are mixed by experts to make life saving decisions. CAFÉ-Map (Amir, Minhas, Asif, & Arif, 2016) was developed as a Context-Aware feature ranking tool. Experiments done on prostate cancer managed to reveal relationships with genes that

match results from other research papers. Ranking of context features helps in trimming elements to be considered during a Data Mining solution.

When Analyst decide to use Context-Aware approach they need to find appropriate stage to apply the principle. Points of applying context information can be prefiltering, post filtering or (Dhaneshwar & Patil, 2016) contextual modelling. Prefiltering being at the point of selecting attributes, Postfiltering being later stages of information analyses and contextual modelling is use of context information throughout the project. Dhaneshwar and Patil reflect the dynamic and multiple stages in solution implementation and application of context.

The need for applications to run on handheld devices results in developers facing both time and computational resource challenges. Concept of Context Space (CS) models designed (Haghighi, Zaslavsky, Wang, Gaber, & Loke, 2009) using Fuzzy Situation Inference (FSI) and reasoning approach was used in mobile applications to create a lightweight solution. CS has both focus traits and solutions for specific applications.

Context-awareness encourages investigations into the surrounding. The resolution uses knowledge on both issues and past solutions to other problems with similar characteristics (Moyle & Jorge, 2001). There is also the formation of new hypotheses on the subject. There is a need to state an evaluation criterion for the model from the commencement. The model can prove, disapprove, or contribute knowledge on the hypothesis. Understanding and resolving a situation is the key contribution of the project.

Researchers' views on data mining undertaking turn out to affect usefulness of resulting insights. (Treffinger, Selby, & Isaksen, 2008) accentuated on problem as an opportunity to achieve an outcome. The journey of achieving these context sensitive results is the central theme of Context-Aware Data Mining. Lean dataset produced by this approach assists in focusing the analysis.

Knowledge research is based on a shared range of backgrounds to create novel views of the world. There is a need to borrow approaches from other domains, which have looked at focusing investigation effort. Translation of a problem into elements that for analysis is a key challenge in any data mining project. (Rauch, 2005) have explored the GUHA (General Unary Hypotheses Automation) procedure which produces prime hypotheses.

Input into the GUHA are key data matrices and parameters used in creating interesting hypotheses. The output is a prime hypothesis, which is true and not the output of another hypotheses. The project will create a story or hypotheses from a subject. Data analysis will confirm or disapprove the hypotheses.

. Feature ranking is one way to improve output quality. (Mao, Mitra, & Swaminathan, 2017) created feature indices in a binary space, mapped it to original space to create a feature ranking. Context attention drives the framework, but even features used for context formation should be ranked.

Feature understanding and processing are key to any quality data mining effort. Methods also vary, like combination of selection search (Ramezani, Moradi, & Tab, 2013) floating, backward and forward feature analysis approaches. Context-aware feature selection is complex, sometimes combinations of processes and a variety of algorithms. All feature processing aims to reduce data dimensionality.

Solutions need to be able to overcome challenges like managing nonlinear feature space. Relevant feature selection algorithms include ReliefF, (CFS) Completely Fair Scheduler and MI Mutual Information. Real-Time Feature Extraction algorithm (RFE), proposed by (Guo, Wang, Liu, Guo, & Lu, 2015) combines a feature selector (e.g., ReleifF) and a modified (PCA), Principal Component Analysis process to produce situation relevant features. Context-aware concepts do not any data mining tasks, rather just modify perspectives.

Research has focused on feature selection types including graph-based, self-training based, co-training based and SVM-based. Academics also worked on the merits of selection algorithms like filters, wrappers, or embedded methods. Through a survey (Sheikhpour, Sarram, Gharaghani, & Chahooki, 2017) concluded that graph-based methods are computationally efficient and good generalization ability.

Problems at hand may have data coming from a range of sources, so they will be need for normalisation and fusion of data. An example of this could be an attempt to analyse IT security risk of a large network. Correlating security sensors, network traffic and network vulnerability databases (Lange, 2017) managed to create context-aware source file for further analysis.

Development of data-based systems could be split into (IE) Intelligent Environments and (AI) Artificial Intelligence domains. Intelligent Environments being more reactive actioning where the application performs defined actions when predefined points are reached by environmental factors for example (GPS), Global Positioning System. Artificial Intelligence is more focused on amassing knowledge over periods of time. While accepting that over the years IE and AI have developed more synergies, (Wrede, Visa, Alerge, & Aztriria, 2018) distinguished machine learning as attempting to apply a single tool to multiple different problems.

When applying data solutions to the real-world the issue of noisy subject data arises. A graph-mining strategy based on approximation was used by (Jia, Zhang, & Huan, 2011)in (APGM) Approximate Graph Mining) to handle protein graphs since they are not well defined in real-life. They found that APGM highly managed to handle noise data. In this case noise was extra considerations in solution design. Algorithm ranking process is complex, (Liu, Wang, & Bhuiyan, 2021) in model pCAR which used greed approach in an iterative loop integrating contextual information and user preferences in creating a personalised Recommender system.

Data preparation, cleaning or removing noise can be costly in both effort and resources, hence the need to consider its necessity during data mining. Context noise and missing data below 20% (Avram A. , Matei, Pintea, Pop, & Anton, 2020) had the same effect in CADM as clean context data.

Context-Awareness concept brings with it a few additional tasks so may need cost benefit analysis before being applied. Context-aware solutions are complex in nature, sometimes needing multitudes of algorithms in a single data analysis project. In a Recommender system (Suppa & Zimeo, 2016) used tags to bring together personal preferences and contextual information. Selection of data mining techniques to apply in Context-aware projects faces challenges (Thabet, Ganouni, Ghannouchi, & Hajjami, 2019) since the situation is changing and tools perform varyingly to different situations. They promoted use of an algorithm and case past performance database as reference base for a selection tool.

The need for situation specific analysis has led to the development of widowing step of times. Time series or temporal data plays a key role in context-aware mining. Extra work might be necessary to find frequent itemset over a specific period. The division of datasets into periods are key to context integration. Behaviours of subjects may overlap these time divisions (Saleen & Masseglia, 2011) defined Solid Itemset as being compact and coherent behaviours in a specific period.

Expert knowledge can scope, inform, or guide a data mining process. Context can be traced from various angles including subject experts. Subject experts inform the design of some data analysis projects. They can also bring other factors like time, location, or business context to understanding of the subject. 'Wisdom Mining, coined by (Khan & Shaheen, 2021) advances data mining by attempting to mimic an expert contribution, automatically informing the process to produce lifetime facts.

In a study of stand-alone and context-aware (Matei, Rusu, Bozga, Sitar, & Anton, 2017) focused on context data quality, with the perspective of availability. They found that Context data improves accuracy of analysis when its availability is above the 79.61% threshold. After finding that k-Nearest and Neural Network algorithms provided highly accurate prediction, (Avram A. , Matei, Pintea, Pop, & Anton, 2019) also concluded that by widening situation influencing factors, Context-Aware process gave a more realistic view of a situation being analysed.

Dynamic contexts need to be aligned with local dataset. There are several challenges not yet resolved by researchers in the paradigm of Context-Aware Data Mining. In a study of (IoT (Internet of Things)) Internet of things, data streaming (Nallaperuma, De Silva, Alahakoon, & Yu, 2017) designed an algorithm to detect concept drift of the context-aware data.

Tourist attractions have attributes that are seasonal, they affect visitors' decision-making process. (Huang, Wang, Yang, & Xu, 2018) expanded the standard (LDA) Latent Dirichlet Allocation to form (STLDA) Seasonal Topic Latent Dirichlet Allocation by incorporating seasonal characteristics. Their empirical comparison found the STLDA being comprehensive and much closer to real-life behaviour of this domain.

Context knowledge has always been used for decision making and as part of production systems to determine actions. Using China as a backdrop (Majid, et al., 2013) predicted recommendation in a new city using geotagged photos collected on a tourist past travelling history.

(Bhadane & Shah, Context-aware next location prediction using data mining and metaheuristics, 2021) proposed CANLoc, Context-Aware Next Location Prediction, which automated assigning of location tags to subjects using context data based on metaheuristic methods to improve accuracy by between 12-15%. Metaheuristic techniques perform wide ranging data mining to gain patterns, they include Genetic algorithm, Neural Networks, Naïve Bayes, Hidden Markov Model etc.

Data analysis attempts to resolve real life business issues. This presents a challenge of data being in too abstract levels compared to description of the business issues. (Van Houdt, 2020) suggested feature engineering to create root cause analysis that are informed by the context.

Context-Aware systems are complex to design, (Sherif & Alesheikh, 2018) split it into presentation, provisioning, and processing layers. This being interaction with situation, processing context and modelling data, respectively.

2.2 Context-Aware Machine Learning Class Selection

Data Mining Techniques

Data mining methodologies design depend on the nature of the issue at hand the opportunity can either be classified as supervised, semi-supervised or unsupervised. Over the years supervised methods perform highly (Van Amasterdam, Jackson, & Stoyanov, 2021), as compared to Unsupervised and semi-supervised. Unsupervised and Semi-Supervised datasets occur in more occurrences than Supervised ones. Below is an overview and history of the main methods being association, classification, and clustering methods.

2.2.1 Association Mining

Association methods which find items that appear together or have some relationship which can help a business decision. The project will progress from selection of basic methods to discussing advanced variants and their distinct characteristics.

The Fuzzy rules oversee quantitative rules, boundary problems and generation of vague rules. On the other hand, the classic Association Rule deals easily with the binary situation of presence or absence of items in a set. Neuromorphic Association Rule improved by (Abdel-Basset, Mohamed, Smarandache, & Chang, 2018) attempts to use truth, indeterminacy, and falsity function to manage wider indeterminacy rules. This is an approach deemed to increase the real-life domain that Association Rules can be applied to.

2.2.2 Classification Mining

Classification which maps attribute X to Class Y based on their characteristics, making it a typical Supervised analysis. Classification tree the leaves are allocated values of the target variable to label a set of records. There are many methods including C-5. Algorithms return varying performance rankings on key mathematical measures posing an extra challenge to any selection automation effort. (Sridevi & Prakasha, Comparative Study on Various Clustering Algorithm Review, 2021) had to use the average of (F-measure, ROC area, Recall, accuracy rate, precision) in measuring performance of (Naïve Bayes, decision trees, Neural Network, logistic regression) algorithms on a financial dataset.

Association mining (Santoso, 2021) is the use of Apriori algorithm to explain relationships between items in each itemset, loosely termed shopping basket analysis. Over the last decade the trend has been to mix established approaches in creating more situation sensitive data mining solutions. Based on Association Rules (AR) (Al-Shargabi & Siewe, 2020) considered user environment and physical surroundings as context to predict maximum temperature by creating (LWRCCAR) a Lightweight Association Rules Based Prediction Algorithm.

The context-awareness can also be passed onto algorithm grouping level like Deep learning. Some metrics are method specific like Classification which is measured using

(Accuracy, F-Measures, G-measures, Precision, Misclassification Rate, False Positive Rate, True Positive Rate, Specificity and Relevance).

Reviews of data mining algorithms involve several measuring metrics. (Suragala, Venkateswaralu, & China Raju, 2020) found Support Vector Machine (SVM) having accuracy of 74.82% using 10-fold cross validation method, it outperformed Naïve Bayesian Classification (NBC), K-nearest Neighbour (KNN) and Artificial Network (ANN). Using a large cancer dataset in 10-fold cross validation of three classification algorithms (ANN, C5 decision tree and Logistic Regression, (Delen, Walker, & Kadam, 2005) confirmed that, C5 performed highly at 93.6% accuracy. The experiment also found logistic analysis to be extremely sensitive to changes in attribute values.

2.2.3 Clustering

Clustering that allows inferences by grouping records based on values of characteristics of attributes. Clustering groups records with similar feature values into same the clusters. Below clustering reviewed from its basics, distance between feature values, classes, and performance metrics. Dunn and Bezdek developed functions and algorithms (Bezdek, 1973) to extract whether an object belonged to a group and if the group is a cluster, this has become known as fuzzy c-means or Fuzzy ISODATA.

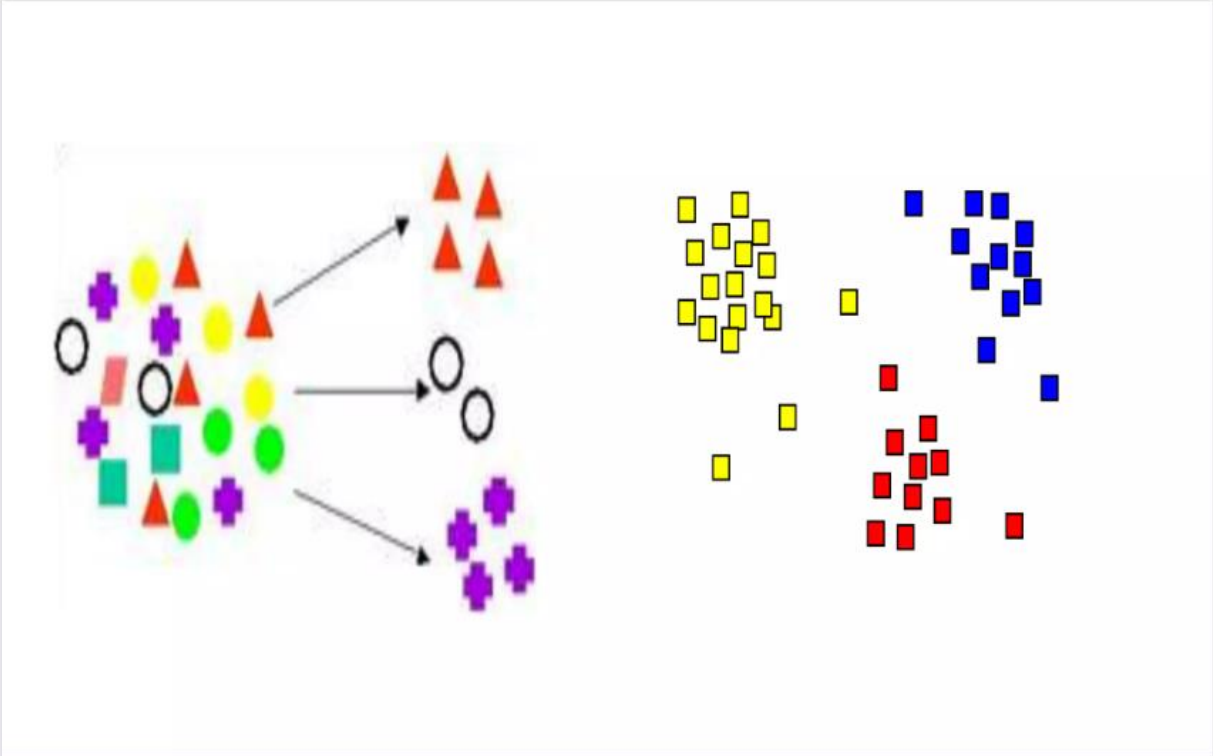


Figure 7: Basics of Cluster Data Mining (S. Archana,2023)

Fig.7 shows grouping of data about the same shapes into clusters based on their features.

The interpretation of the features of each Cluster helps understanding issues (Khanbabaer, Alborzi, Sobhanl, & Rafar, 2019) under investigations. Algorithms like K-mean are used for building these clusters and distance between records is measured using the Euclidean method. Clustering algorithms can split into four classes, namely Partitional, Hierarchical, Density based and Grid. These clustering algorithm classes work best depending on data properties for example data types, (Popat & Emmanueal, 2014) dataset size, noise, and outliers in the data. These classes also manage scalability, cluster shapes, and need for domain parameters differently. (Sridevi & Prakasha, 2021) pointed to Hierarchical class not managing high dimensions, Partitional needing more computational processing while Density based class tends to manage data noise well.

Based on the nested-EM algorithm, spatial context visuals and context trends in features (Yuan & Wu, 2008) created a Clustering framework that reduced errors and increased convergence more than a traditional k-means clustering approach. During model implementation the user may need to tune the algorithm design. These adjustments improve end results. (Lin, Coller, & O'Hare, 2017) noted that most clustering algorithms have poor multihop intra-cluster links and lack user interface. Therefore, the authors encouraged effort on user interface to allow customization, thereby improving performance. In a study measuring model performance, (Pourghasemi, Yousefi, Kornejady, & Cerda, 2017) found an ensembled model of ANN-SVM (Artificial Neural Network) (Support Vector Machine) to highly perform on goodness-of-fit and predictive power. The reasons for the outcome are related to the design of each model and compensating features of both. Comparing 14 Clustering algorithms (Duo, Robinson, & Soneson, 2018) *securat* and *SC3* had best performance on; run-time, stability, and recovery of marked clusters.

2.2.4 Ensembles and other methods

The advancement of technology and complex areas of application has given rise to ensemble learning methods. These are solutions using multiple data mining methods and algorithms. In this case the first method can be used to perform some data mining. The data mining may be to prepare data for further analysis. The further analysis might be using a different method. In this thesis it could be to integrate Context and Data.

Feature selection design complexity is dependent on structure of data, for example semi-supervised data is not easy to process. Semi-supervised data has both labelled and unlabelled subsets. Context-Aware Analyzer (CAA), (Kim & Chung, 2016) follow the main stages of analysis being Input, Process and output it uses HanNanum KO-NLP speech analysis, LSA Latent Semantic Analysis and Fuzzy c_Mean (FCM) Clustering to evaluate transportation situation. They processed unstructured and structured environmental data.

2.2.5 Semi-Automated Decisions Making in selecting Data Mining methods.

Chain of decisions made in designing a context data mining solution are a sum of decision-making process, nature of dataset, ability of data mining method and algorithm performance. Background of decision-making systems must be informed by past literature on the human decision-making process. Understanding of human decision making has been studied in domains from History, Religion, Biology, Psychology to name a few. (Fogil & Guida, 2013) Decision Support Systems DSS should be subject knowledge-centred in design and considering user requirements. Selection of model components will be robust and of wider application.

Data is about real life which has interaction contexts, (Yau, Dickert, & Joy, 2010) used various analysis methods in mCALS mobile Context-aware learning Schedule, to put together time, scheduling and learning context for students. The decision automation of phases of model creation (Duan, Edwards, & Dwivedi, 2019) augments human effort rather than replacing it. The logic is a mix of rule inference and other processing techniques. The decision-making agent should be controllable, adaptable, easy to use (Duan, Ong, Xu, & Mathews, 2012) and reactive to varying operational environments. Knowledge elicitation is key to (Konar, 2000) this project, it involves presentation in structures, refinement and absorbing new pieces of knowledge from the surrounding.

Chapter 2 Summary

The chapter looked at past research by peers on the building blocks of the concept of Context-Aware data mining. It looked at the possible benefits of this novel way of looking at the data mining task. Semi-automating the method selection stage depends on deep understanding of characteristics of existing Data Mining methods.

CHAPTER 3 Research Methodologies

Chapter 3 Introduction

This chapter looks at academic research methodologies established over the last decades matching this thesis. Research methodology works to give academics a way of understanding peers' novel ideas. The chapter addresses evolutionary growth of research highlighting the path followed by this thesis.

3.1 Research Methodology background

The thesis presents a proposal on a new framework, suggesting incorporation of environmental factors during Data Mining processes. The methodology needs to explain the proposal, work through implementation challenges, illustrate implementation and explore performance measuring approaches. Researchers in Information Technology domain lag in following traditional research methodologies. Possible Methodology include Case Study, (Jenkins, 2000) defined Case Study methodology as involving observing subjects without influencing variables be they intervening, dependent or independent.

The thesis analyses implementation of Context-Aware framework on a real-life PIMA diabetic, India Traffic and World Pollution dataset. Data Mining domain principles are then used to evaluate the approach to measure performance against targets like increasing relevance, easy interpretation, and focus. Thesis has loosely used the word Data Mining framework to refer to model design or algorithm design. In this thesis framework is the end-to-end analysis architecture.

Methodology Summary Table

Objective	Methodology applied	Reason for using the method
Develop CMAS Framework	Design Science Research (DSR)	Method designed to communicate development of a product. The steps defined in the Methodology are aligned to stage the move ideas through definition to testing a product.
Semi-Automated method selection	Standard Programming tools (Flowchart, pseudocodes etc}	Research aimed to translate ideas into software form, the tools have become established for this purpose. These tools communicate ideas and ways computers achieve them.
Evaluate the framework	Case Study	Case study illustrate the framework and its application in real-life. Case Study highlights steps and challenges with applying framework to varying contexts.
Analysis of framework	Case Study	Comparing performance of the novel framework with others need a vehicle, Case study suits well since it has a Context and Selection of method.

Table 2 : Objectives and Methodologies Summary

As shown above in table 2, combination of Methodologies covers application of contextual factors and retrieval of knowledge for decision making. Design Science Research (DSR) approach was used to build and communicate the novel framework CMAS. Analysis and evaluation were done using the CMAS framework on PIMA Indian diabetes, Indian Traffic

and World Pollution datasets. This research will propose a way of constructing a solution using building blocks like existing Data Mining methods and algorithms.

Challenges range from optimal levels of knowledge on algorithm characteristics that match a given data situation. The other challenge is to do with the large volume of algorithms that one can apply to similar situations. Simple functions and summary of steps will be used to illustrate the proposed solution. The framework will be a skeleton promoting method rather than a rigid outcome. In future other researchers will be able to apply the steps for varying problems producing a high performing solution. Performance of framework or tool should measure on properties like fault tolerance, scalable, dependable, consistency, accessibility, and robustness.

3.2 From basic academic methodologies to Design Science Research (DSR)

The CMAS will used on several of domains to illustrate and produce metrics on quality improvements obtained from the novel approach. Academic research solution design start with philosophy, in this case Positivism, Realism or Interpretivism being. Three Real-life data analysis is performed to validate the proposed novel solution. Few hypotheses were formulated, and expert opinion was sorted through expert analysis of results to accept or dispel or confirm the prediction.

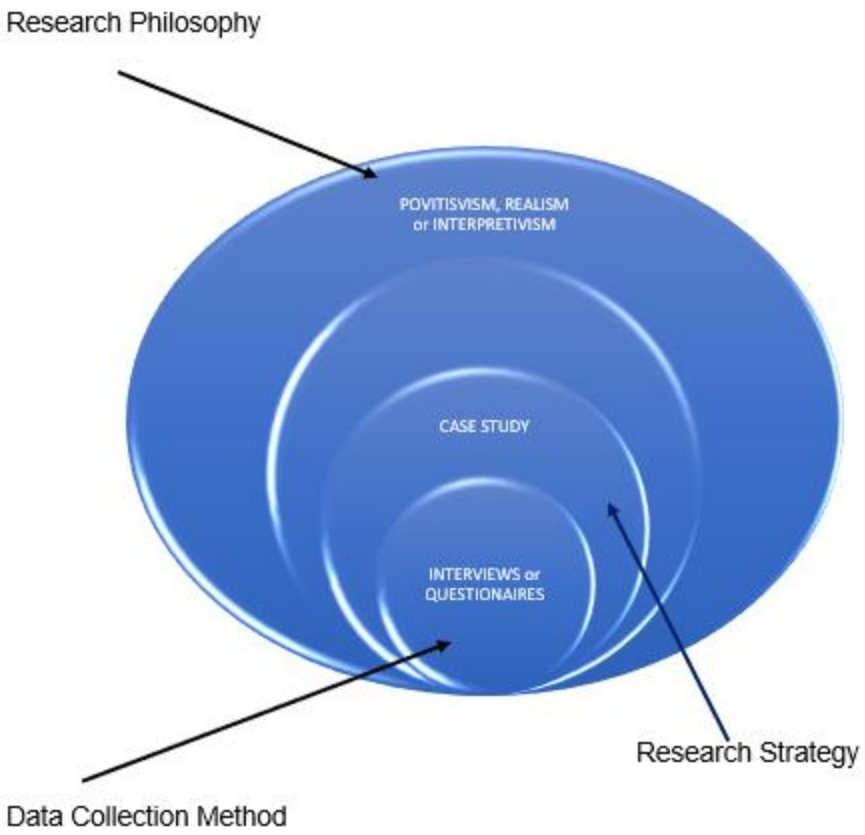


Figure 8: Context-Aware Research Onion

Fig.8 is an outline of Research Philosophy Research Strategy and Data Collection Methods followed in this thesis.

The proposal is illustrated through a case study for analysis by using domain methods. The validation and evaluation sit in Constructivist research philosophy (UK Essays, 2022) allowing experts views collected through questionnaires to give a qualitative response.

The thesis explores use of context at various stages so a Data Mining framework. It also investigates systems decision making as an aid to selection of data mining methods for a given case.

The thesis sits in the Design Science Research (DSR), (Brocke, Hevner, & Maedch, 2020) paradigm, which aims to design a solution based on knowledge, evaluation, and presentation of the design. As shown on (Fig 7. Design Science Research) it has several of steps. Evaluation of the design involves any methods including interviews and surveys. Success is measured by the level of resolving the original challenges. In this case the problem is Data Mining solutions being divorced from the real world in which datasets are found.

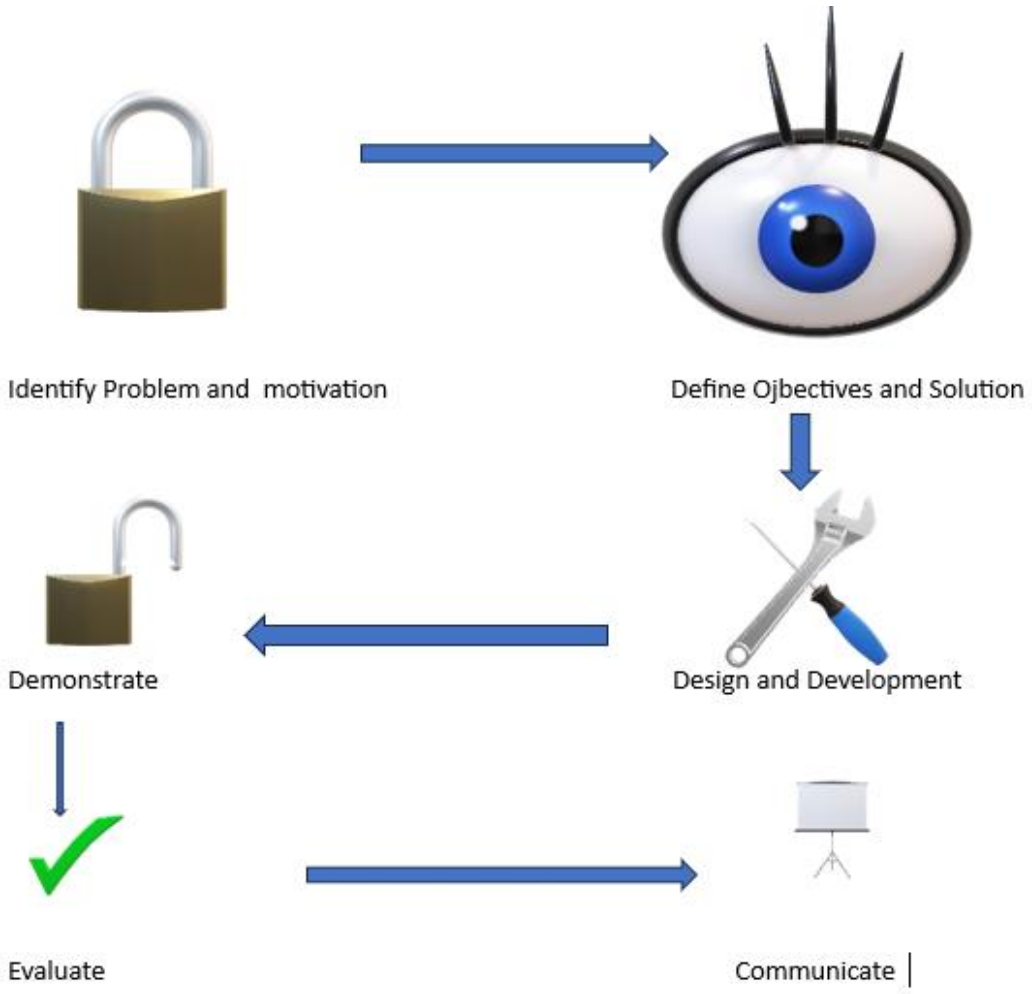


Figure 9 Design Science Research DSR (Brocke, Hevner& Maedch, 2020)

This research has followed the steps illustrated in fig.9.

Academic work may wish to investigate some theorem, follow existing established steps. In case of DSR the research attempts to construct something.

Steps being:

- Identify problem or motivation, this is scoping what the new construct aims to resolve.
- Define Objectives and Solutions, structure a solution. For example, CADM aims to resolve or remove difficulty of interpretation of data mining outcomes.
- Design and develop – put together ideas and construct something new.
- Demonstrate – Once built the new object need to be shown in sue to peers.
- Evaluate – Measure performance against aims and objectives, is it resolving the problem.
- Communicate – Let the world know about the new product. This could include manuals and bulletins.

This research is more about solving an existing problem through design of a novel framework making it sit well in the DSR method.

Context-aware Data Mining was performed on Indian PIMA diabetes, India Severity traffic Accident and World Pollution datasets. Context information played a significant role in the design, implementation, and interpretation of outcome. The CMAS approach uses a Data Mining Method Selection Buddy to assist in model building decisions. Theses perform a comparison analysis to measure benefits accurate from the novel approach.

Chapter 3 Summary

Research methodology ties together all chapters of the thesis. It takes the journey from ideas, definitions of ideas, exploration of implementation ways to evaluation of real-world implementation challenges.

CHAPTER 4 Proposed Novel Context-Aware Data Mining Framework

Chapter 4 Introduction

This Chapter highlights on proposing an overall conceptual framework considering various aspects to mine data with context consideration. Develop a reusable data mining framework that considers the environment surrounding the dataset and uses accumulated data mining methods knowledge. The approach should be an agile, robust, and simple Context-Aware Data Mining framework. It will be robust in adjusting to changing contextual factors. By working with slimmed down focused contextual scope it will be a simple failure fast and retry approach as advanced by agile concepts. The framework takes into consideration the environment where the dataset is found. It should illustrate ways of defining context.

This Chapter highlights actions from gathering context features covering interpretation of insights based on environmental factors of a dataset. It is presented in three-sections.

The first section '4.1 Traditional data mining approaches/ Background provides building blocks of the novel approach by looking into the past development in the field of data analysis.

The second section '4.2 Contextual information integration decisions' emphasis on bringing contextual perspective into the data to be used in the model. This perspective influences most decisions implemented in the analysis process.

The third section '4.3 Context Framework Method and Algorithm Selection' emphasis on understanding requirements of tasks at hand to match them to existing data mining methods. The nature of the input data is also an influence, though data can be transformed to a limited extent.

The fourth and last section '4.4 Context-aware Model Implementation and insight interpretation' runs the model and obtains sights. This step has lots of human effort to

understand the outcome. If the context related trimming works well the step will produce limited but quality insights.

4.1 Traditional Data Mining Approaches/Background

Over the years, the Data Science industry and academics have created approaches like CRISP-DM, KDD, and SEMMA. Most recent data mining methodologies are variations with limited addition of these established DM approaches. CMAS retains processes or stages of established traditional data mining. The added aspect or activities are consideration of context as appropriate for a given case. Logic of the proposed solution uses grounded knowledge of data mining techniques and understanding of the problem domain at hand. Data Mining projects design based on guidance from above contributions in key important considerations that improve automated decision-making applications.

Data Mining has investigated by academics and industry creating basic steps. The full spectrum of processing data follows key five steps from knowing business expectations ending with solution deployment. Frameworks are used to put together tasks and guidance around each task on a given process. Emphasis is now on improving the designs of the established stages in line with gathered and accumulated expertise.

Selecting appropriate Data Mining method for a given problem with fast and fully informed knowledge of features and characteristics of established methods. Human beings have limited knowledge storage and processing capacity when compared to computer systems. Novice Analysts and domain experts face challenges when selecting method to apply on a given data analysis problem.

Academic and industry have gathered a huge amount of knowledge on Data Mining methods. There has also been lots of advancement on computer systems processing power. On the other hand, the research has made too little attempt to process this information to aid analysts in selecting the most suitable methods for a given problem.

Context based computing has been increasingly researched over the last three decades. Business Understanding analyst checks what situation need to be resolved. Expanding on CRISP-DM; Data Understanding these are attributes and characteristics of data; Data preparation data comes in many forms and with varying levels of noise. In the activity data needs to be transformed to align with model requirement; Modelling involves development of tools to analyse data; Evaluation is checking if all components are working as expected and levels of performance are at acceptable range. Deployment related to using the solution for further analysis.

The innovative approach should not increase data mining process complexity. Context being surroundings or expert knowledge on the dataset domain. Produce a framework which explores key surround factors, find ways to integrate them into focused analysis of the local data. The research explores the possibility of semi-automating decision-making steps when creating a Context-Aware data mining solutions. This should aid, assist Data Analyst in the design and implementation of improved environment sensitive data mining solution.

The proposed CMAS approach, define and explore use of subject expert knowledge, environmental factors, and situations around a given dataset. Analyse current Data Mining approaches difference with Context-Aware Data Mining concepts looking at problem definition, feature selection, feature transformation and data preparation. CMAS also explore the possibility of automating the design of Context-aware data mining.

It encourages modification of implementation Machine Learning methods including Deep Learning. A few modifications have been done to main (Association, Clustering, and Classification) data mining approaches. Implementations are adjusted to match situations, characteristics of dataset and purpose of the analysis.

Explore ways of measuring insight quality including relevance and ease of interpretation. Increase understanding of the role of expert background knowledge on the process of getting fresh insights from a dataset. Give a deeper understanding of the Context-Awareness concepts and their implementation.

The thesis will contribute on the use of knowledge databases to make design decisions when designing Data Mining solutions. Research contributes to giving developers

attention to the importance of context, application, and domain to the overall data mining journey. The need to accept that understanding of environment aid in relevance of insight resulting from a data mining solution. Explore integration of contextual factors to design and implementation of data mining procedures. It is hoped that there will be an increase in domain relevance of the insights from the analysis. It is a fact that datasets exist in an environment which tends to influence its features.

Thesis contributes knowledge on simplified application of context-aware principles in data mining. Expand knowledge on interactive model design, feature engineering, method selection and expert knowledge integration in the implementation of Context-Aware Data Mining solutions.

The research attempts to create a knowledge and situation processor that assists with model design decisions in a semi-automatic manner. Data mining and moreover Context-Aware data mining is ladened with lots of decisions. The question is whether domain knowledge can be processed by a script to output key design decisions. Data Science is both a science and an art with human creativity playing a role in its advancement. New model must not reduce human creativity.

Formulating steps and principles to integrate environmental information into Data Mining should improve understanding of more situation linked data processing. Analysis of approaches to situation information integration, relevant decisions and choices makes this a candidate for a software-based solution.

There is limited knowledge on automation of Context-Aware data model designing. The research seeks to explore simplified strategies to create high quality analysis solutions using intelligent strategies. Contribute knowledge on mixing background information processing, analyst knowledge, automation of algorithm, performance measuring and other decisions selection considerations. Explore the use of an interactive application to understand analysis opportunities, pre-process data, and select the best Data Mining algorithms.

The research details the application of Context-aware and Automated Machine Learning (AutoML), principles to everyday Data Mining applications. AutoML is explored more as a standalone application that can be given input to produce some results. It also discusses

and explores solutions to the current challenges around these semi-automated approaches.

Start of any Data Mining effort is having a clear understanding of the project aims. Problem understanding is difficult given the wide spectrum of areas analysts try to apply Data Mining and the human expression challenges. A few practices will need to be explored in this research. Data Analysis is done with a range of resource challenges from time, machine processor capacity to interpretation.

Limited research has been done on the development of reusable data mining frameworks. Exploration of automating decisions in Data Mining solution design have also been very few. Improvements from the Context-Aware approach should range from better understanding of techniques, algorithm classes, algorithms, and some performance measures. The research focuses more on building the Context-Aware solution than its implementation.

Effort to explore the possibility of interactive procedures in construction of these data sensitive models. Main contribution will be improving decision making during design and implementation of context-aware data mining solutions.

4.2 Contextual Information integration decisions

Data mining (DM) is a multi-staged process, thereby increasing the possible points of integrating context information. Context-aware information is used at varying stages of analysis (Dhaneshwar & Patil, 2016) suggested at prefiltering, post filtering or during modelling. Prefiltering is part of data preparation before being input to a model. Post filtering is using context in understanding modelling results.

The Context-Aware Data Mining Framework has five key stages, which are Context-Data preparation, Local Dataset preparation, Integration of the two dataset, Model design, method and algorithm selection, model implementation and insight interpretation. The thesis proposes adding Context Data preparation to the Traditional Local Data collection step.

Data Mining is composed of several stages. Academics are interested to know stages to apply Context-Aware concepts. The process need not to be limited by what has been done, it should open to novel solutions and context application points. Moving to a Context-Aware approach promotes the related changes to be made throughout the design of the whole Data Mining solution.

The Data mining setup is a 'Lego.'

Context-Aware Data Mining utilizes the components or steps that have always been used traditionally. Situation linked data analysis tends to produce actionable knowledge on a subject. Data mining is nowadays producing more usable results, credit to advances in model designs. The research attempts to further knowledge on the Context-Aware Data Mining approach.

Situ

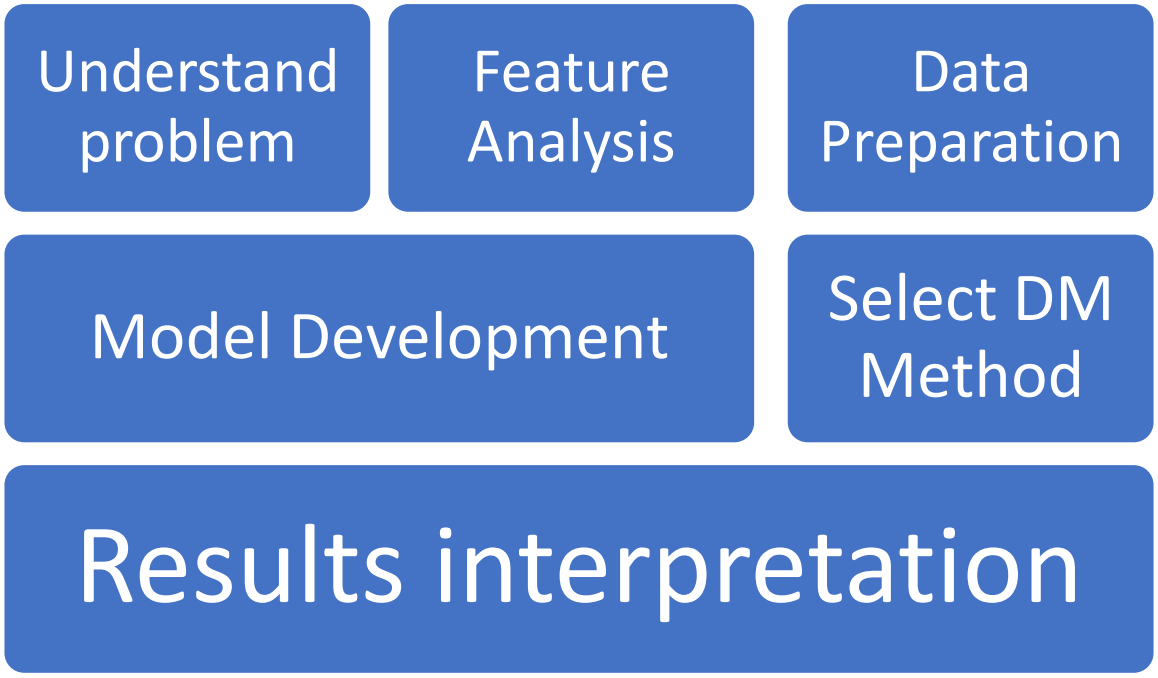


Figure 10 : Data Analysis 'LEGO.'

Various research efforts cover parts of the structure or architecture. Fig. 10 confirms that method selection effort of this research covers the single component 'Select DM Method'

of the above structure. The Context-aware approach covers and considers the whole structure.

There are solution design challenges in keeping it simple but fully integrating situation factors into design of a solution. Data is generated in high volume. As a precondition for processing data from varying sources need to have a common grouping key to aid processing.

CMAS may be measured based on algorithm performance, but it is data preparation and is improved by Context consideration. The emphasis is on understanding key concepts and the semi-automation process. Situation linked data analysis tends to produce actionable knowledge on a subject. Data Mining is nowadays producing more usable results, credit to advances in model designs, this research wishes to contribute to the area.

The proposal seeks to explore promotion of Contextual Features during Data Mining. Research done on Data Mining frameworks, comparison, and design of algorithms, but these tasks have not been integrated into building of applicable solutions. The knowledge gap is around making context a major solution design consideration. The prototype can be reused in varying analytics projects with increased chances of success. This knowledge is used as input to model design effort, with interfaces and level of semi-automation.

Context, inform on the importance of each feature characteristics allowing ranking that help to trim further analysis effort. Benefits of the novel CMAS framework are drawn from an informed scoping that concentrates on end user perspective and domain issues ranking. The focused effort increases likelihood of results being accurate, relevant, and easy to interpret.

Ways need to justify any added processes and complexity between the two procedures. Looking at the use cases, PIMA Indian Diabetes Mellitus analysis effort was concentrated on informed expert views on the subject at hand. The Indian Accident Severity use case was aimed at making high dimensional data simpler to analyse. The final use case World Air and Water Pollution emphasis on effort to supply data covering a reasonable angle of a challenge, based on Context.

4.2.1 Data Sources and considerations

Data Integration is a wide stage with Analyst working out the optimum ways to apply context knowledge. The stage determines the path of solution implementation. Solution will reflect influence of the environment on the local dataset. This integration takes varying forms including modifications in the design of the Data Mining model. to reflect the effects of Contextual factors on understanding of Local Dataset. The design modification includes relationships of features of the Local Dataset and how they are affected by the Context.

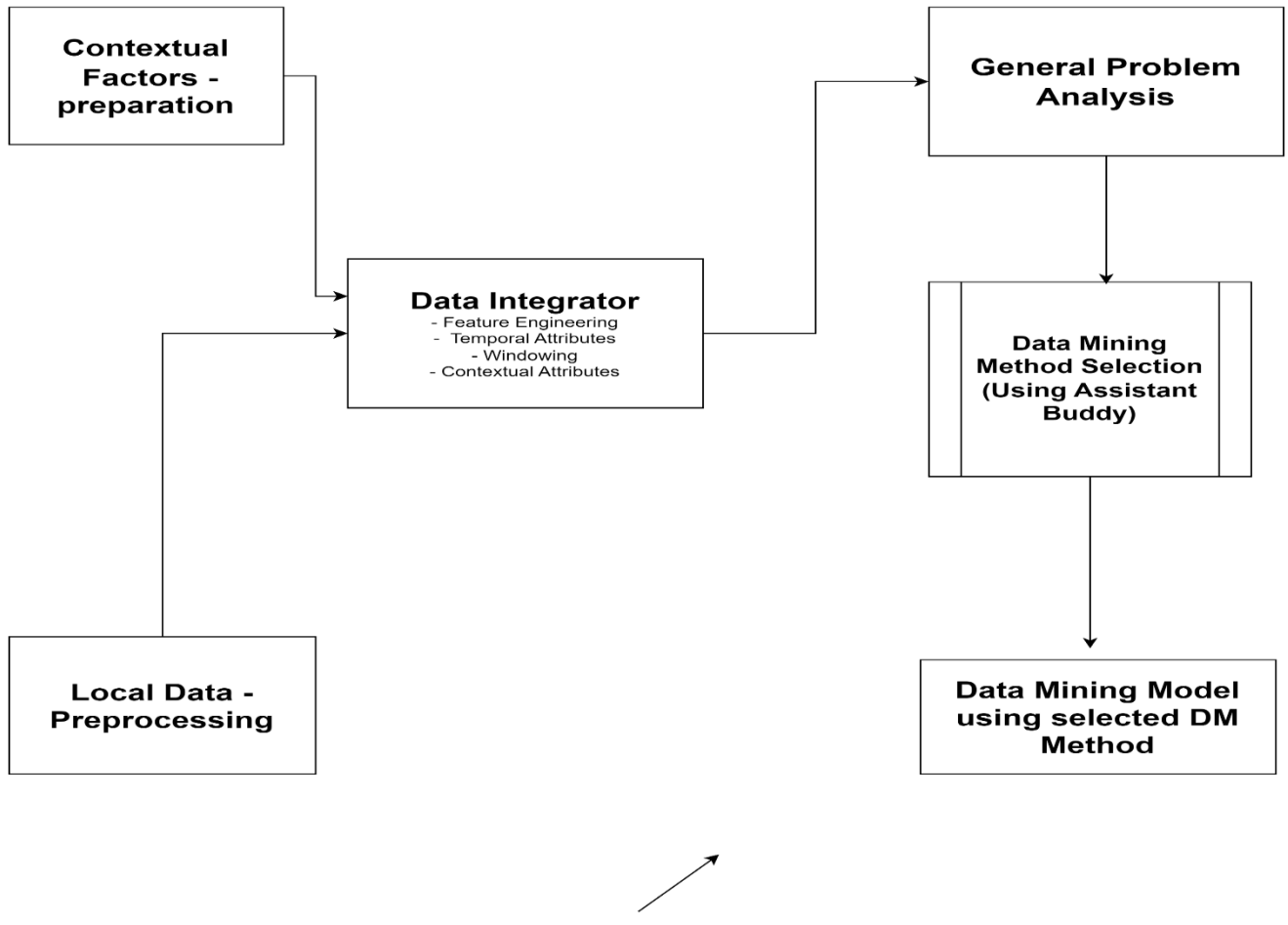


Figure 11:Basic Context-Aware Model

The phases in CMAS are six like illustrated above. Data is about to be processed based on the contextual factors. As reflected above in Fig. 11 there are three main phases being Gathering Local Dataset, Consideration of Environment, Integration of the local and environment datasets and Model Implementation. Data Mining Process being actions that implement the data mining solution. Context-Aware Data Mining follows standard steps importantly including situation influence. Data Mining processes include problem understanding, decision making, data model construction and result visualisation.

Whether as part of a model or results translation some visual outputs are used for better communication. Human understanding of insight is increased by using visuals to illustrate insight. Data mining effort might be processing large dataset that can produce overwhelming results. Decision making is putting together options and selecting the optimum path for a given problem. The proposed framework explores building blocks which are put together to resolve a business challenge. Blocks grouped as algorithms, data mining techniques and the proposed framework. Modification of this framework to make it Context-Aware by making situation consideration during both design and implementation.

4.2.2 Data Mining Method Selection

Proposed solution is a strategy to process grounded knowledge using software functions and interactive information gathering. Some sections of the project cover data gathering and formulation into machine understandable parameters. Context-Aware data mining strategy promotes translation of environmental data to produce insights that are context sensitive. Work on this thesis may concentrate on some stages, but the approach is looking at the whole life cycle of Data Analysis. Illustration should be viewed as the complete cycle, but research emphasis will concentrate on Context integration.

Over the decades machines have proved good at making logical and knowledgeable decisions. The state-of-art around DM method selection decisions are currently being made by humans, based on intuition and historical knowledge. The multi-dimensional perspective of data mining problems like current context, expert attribute importance ranking, and algorithm performance may help in the design of situation sensitive data models.

The first phase of the proposed solution is assistance in deciding when and how to apply context to the data mining process. Implementation of Context-Aware Data Mining follows about five stages. The stages are Input information gathering covers Problem understanding, Environment factors gathering, Local dataset preparation. Varying terms may be used but this stage is making all source data related decisions and activities. The Analyst must make a key decision on the point in the cycle to apply context data.

Logical justification on a particular stage to apply contextual information, need to be produced. The selected stage will allow the highest performance of the analysis. The research will present a logic to make these decisions. Illustrate the number of situations that the model can decide to use as application point. There can be several application points for a single Data Mining solution.

During initial planning of Data Mining solutions environmental factors must be identified and thought must be put on points of application and ways to integrate these factors into the model design.

Datasets exist in an environment, also sub dataset can be identified within larger Datasets. The proposed framework is focussed around considering the environment in which the dataset finds itself. Therefore, the investigation should start by looking at fundamental issues being investigated. Understanding requirements progress from brief statements and likely aspects of interest to high level pointers of what investigation might bring.

There are several theories on ways to understand Information Technology problems. The proposed solution can only be effective based on accurate Data Mining problem definition. It should be noted that Context is defined starting from the issue description. In the same vein Context of interest might vary for the same dataset depending what knowledge is being investigated.

Development towards real-life solutions has forced Data Scientists to mix strategies in a single solution. One method may be used for feature analysis, the another for prediction. Deep learning (Pei, Vidyaratne, Rahman, & Iftekharuddin, 2020) CANet Context-Aware deep Neural Network, use Convolution Neural Network (CNN) to encode context and

select optimal features from a high-dimension situation around cancer tumour segmentation.

The research will perform analysis on a dataset with some contextual information. A new developed script should be able to decide on the appropriate context application point. Actions may be list of Data Mining stages and purpose of contextual information on each stage. Other stages include the following logic needed to decide on applications point expressed both as a flowchart, table, script, and pseudocode. The logic is a processor that takes in problem details and domain knowledge to produce a few key decisions.

The perspective of Context-Aware feature processing is also of interest given the key role they play in quality of insights. Context-Aware concepts transform most stages from problem understanding, feature selection, Data Mining methods and model design.

The research will show outcomes from this logic, a single example will illustrate from a real-world sample dataset and situation. The context analysis will be performed in the 'Experimentation/Case Study' Chapter five.

4.2.3 Context Aware Model Stages

The CMAS framework has five basic stages from User requirements to insight analysis. They are illustrated below as sequential, but they have an iteration approach based on reviews.

Processes in Context-Aware framework (*CONTEXT-AWARE MINING STAGES)*

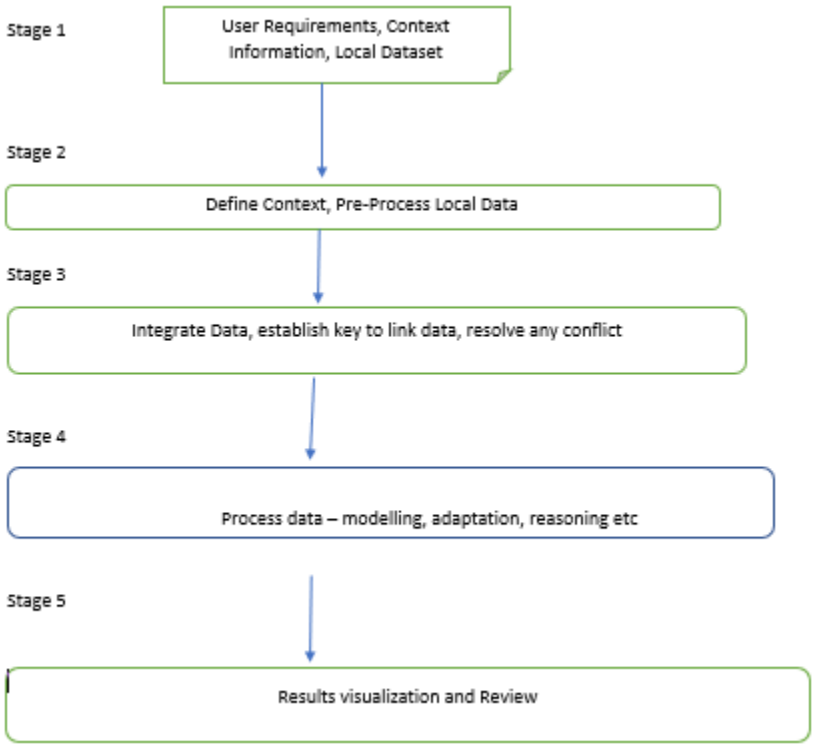


Figure 12: Context-Aware Mining Stages

The stages above on fig. 12 can be looked at as blocks to build Context-Aware models.

Stage 1. Context identification and description.

Creation of context is informed by expert knowledge, data coming from sensor or some features in a dataset. Expert knowledge needs to translate into logical form for example, if features A, D, F are above value 10 then Context C is high. In more complex contexts they might be a function to define or calculate the relationship between features. (BMI) Body Mass Index being a relationship between patient height and weight. There could be labels of low, high, and normal as stipulated by medical literature.

Recognising context occurs at any point of the processing cycle from Requirements definition, Data Pre-processing or during Model Running. Sometimes context is used to aid interpretation of model outcomes. Context can also translate into windows or may change with time. Time series need to be combined with other environmental changes to form a good Context Feature.

Stage 2. Context information mapping to primary or local features

Mapping takes lots of approaches, it is a key part of linking data to the real world. Not all contexts can be considered in a single model, so some rankings are necessary and can be done subjectively or evidence based. Need to find methods to match context to primary data. Approaches include temporal time slots, value ranges resulting from manipulation like averages, medians, etc.

This stage might include pruning source data, by taking that which is of interest to the context. If investigating pregnancy complications, the dataset of a human population can be reduced to female, female of children bearing age, etc. A million rows can be reduced to a few thousands by this pruning. Pruning can be by features; it improves model performance. The procedure can have double benefits of focus and reduction of data and dimensions.

The basic raw dataset usually lacks domain relevance on its own, (Wachowicz & Bogorny, 2009) suggested mixing with application related dimension e.g. aspects of ‘Rush Hours’ when dealing with traffic management or ‘stopovers’ when looking at tourism situations. Context features created from combining different elements of data items or creating windows.

Stage 3. Using context information in the model

Once data and context has been pre-processed it should be formatted in line with the design of the model. This might not be straightforward as either there is an available input data with context integrated or context information plays more of an interaction role. In the later context introduced at the point of creating or running the model without prior changes to input data. This brings the concept of context embed input to model or context modified model design. Context embed input means the local dataset and context integrated before loading into the model. Context modified model design means context used to design the final data model.

Stage 4. Process data – modelling adaptations reasoning.

So, one position is that by now primary data and context information has integrated. Data can input into the model to get results. Another scenario is that context information used to control the model. In this case the model is designed to incorporate context into its logic.

Stage 5. Interpretation of results or use of context to understand obtained outcomes.

Context used to interpret outcome from the data mining model. All above are simplified possibilities, in real life context may be applied in more than one stage of the life cycle.

Context-Aware data mining could be multi-dimensional analysis. Address the question whether a primary or local subject behaved differently when an environmental factor changed. Implementation of Context-Aware Data Mining solutions must resolve many obstacles from matching data from many sources, creating temporal windows to pre-processing a large dataset. This may also include data from sensor and other manipulation that may be necessary before loading into any prediction model.

The research seeks to increase uptake of this Data Mining science without loss of quality of output and allow field expertise to be an integral driving force. Context-Aware solutions can have subject expert knowledge as input or part of situation definition. Construction of

a data model involves lots of decisions, selection of algorithm class and specific one is the most important one out of them.

4.2.4 Context Aware Model Decision Blocks

Context formulation has several decisions to be made for the analyses to be relevant and of high quality. Below the decisions in input to the are shown through a flowchart.

Decision Points

DECISION POINTS IN A CONTEXT-AWARE FRAMEWORK

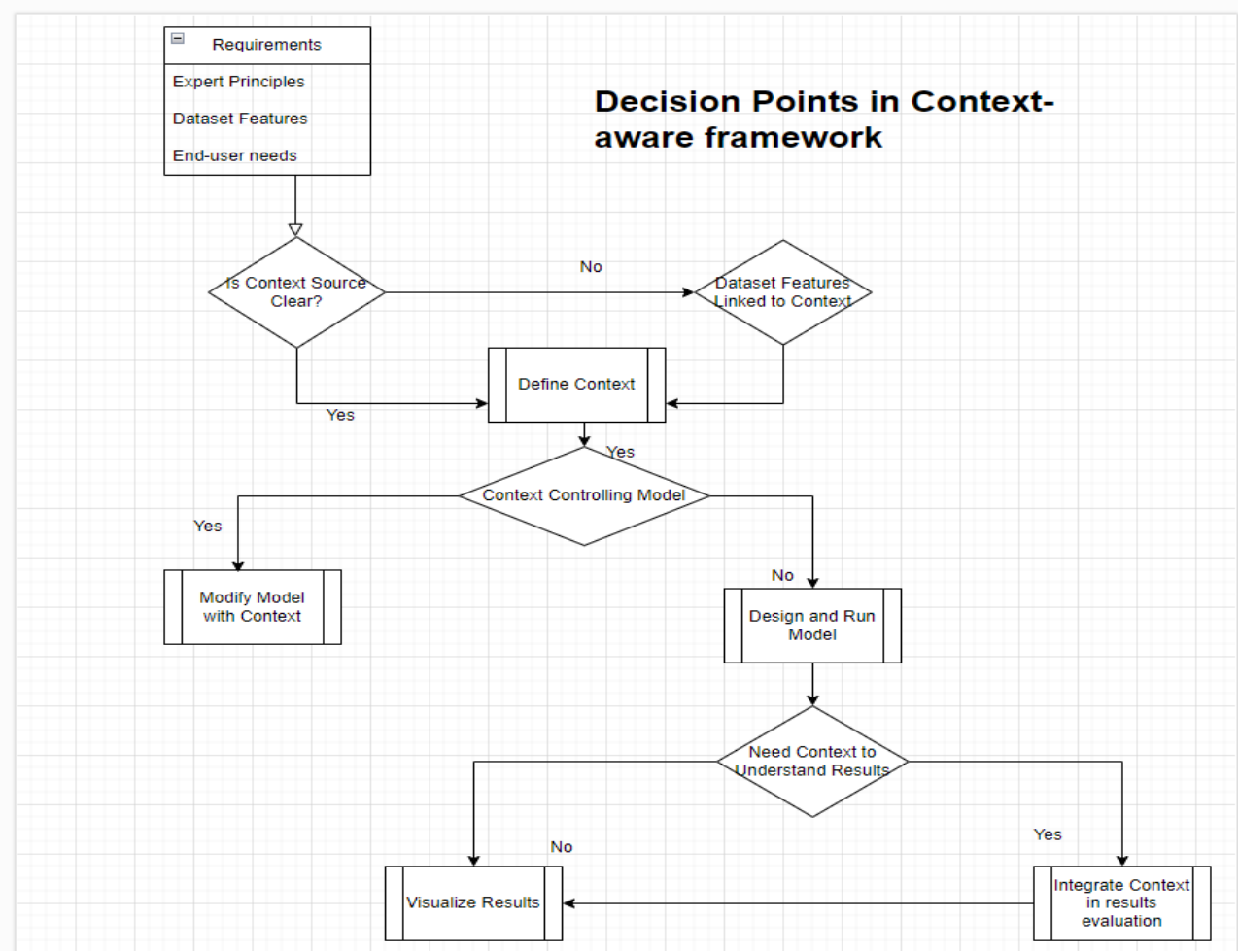


Figure 13: Context-Aware DM Decision Points

The decision made in the fig. 13 are further explained in the follow sections.

Decision description

D1 Decide if the context is complete from expert knowledge and analysis requirements.

D2, should we create Context from some parts of the dataset?

D3 Should Context have applied directly into the model design?

D4 Is it just interpretation of results which already contain context, or should facts of context have applied to the results?

Fig. 13 above shows processes and four decision points. The steps are logical which implements mixing human, programming logic and situation dynamic interactions.

A data model is the actual processing engine that puts together pre-prepared data, algorithm, and initial review of results. Framework refers to a wide view of the solution approach which can be applied in future challenges. Framework used to communicate a proposal rather than a rigid solution. It is the development of methodology and proof of concept. The design of the solution follows a review of work that academics and industry has done on Context and script logic development.

The proposal specifies components and how they joined. These details include the purpose and contribution of each element. Data mining's implementation makes the user an active part of the ecosystem. The user (analyst or data scientist) contributes to tune the solution in a repetitive and interactive process. The project will respect these characteristics in design and evaluation effort with decision points coupled with a few processing loops.

4.2.5 Context Aware Model Data Preparation

The proposed novel framework has contextual features as a driving point for the design and implementation of the solution. Data is looked at as Local Dataset being subject of the analysis. Environmental/ Context features being a faction of outside the scope of

Local Dataset but surrounding and influencing its behaviour. After the first step of understanding the Data Mining problem the second step in the CMAS preparation is the context data and the local data. Third step is integration of the Local and Contextual dataset.

As illustrated below context data comes from varying sources. The mentioned context data does not occur in a single scenario at the same time. It should be noted that not all contexts are important. There is a big dependence on the aims of a data investigation.

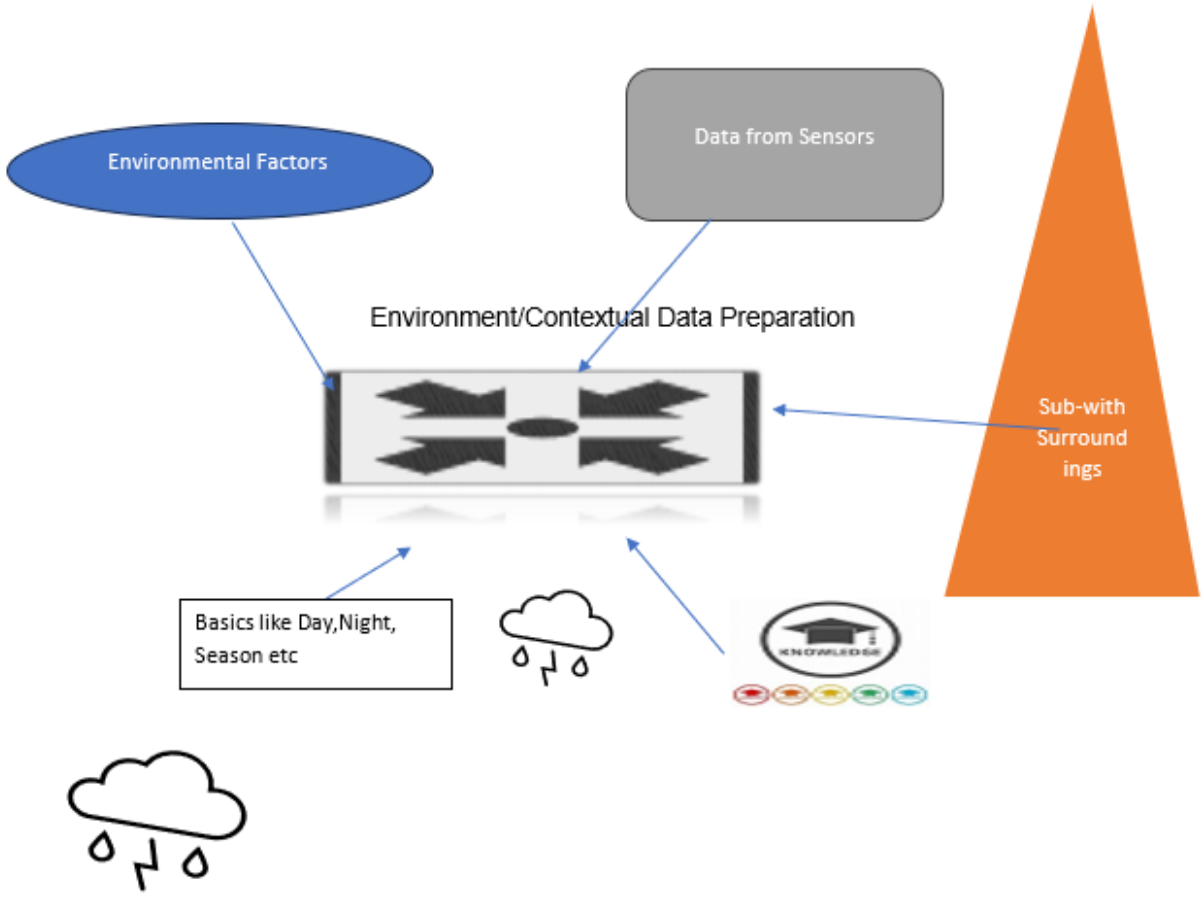


Figure 14: Assembling Contextual Features for Analysis

As per fig. 14 context is in varying forms and is collected using different devices. Combination of sensors collecting data and domain facts can be used in a single data analysis project. The nature of Context in which data is found determines ways of integrating it with local data. Domain knowledge and factors play a pivotal role in the applications and steps in the CMAS process.

In the PIMA medical use case the expert knowledge is used to process incoming data by the novel model. The Indian Traffic dataset split given data into different contexts for further analysis.

4.3 Algorithm Class Selection for Context-Aware Model

This section proposes a context influenced data mining method selection approach. It starts with an overview of the importance of methods in getting interesting insights. A wrong method choice may lead to project failure. The next step moves to high level input, process and output logic of these key decisions further illustrated in flowchart and logic pseudocode. The step ends with revolutionary Automated Machine Learning (AutoML) concepts. Automated machine learning focuses more on selection of appropriate algorithms.

Context-Aware phase of assigning data mining method to a model has its own challenges. Lots of technology can leverage this process like scripting, knowledge retrieval.

A key step in designing a data model is to align the Data Challenge to the right Data Mining method. Until now it has been done by trial and error and use of human intuition. AUTOML limits itself with algorithm selection with a given data mining method. There are volumes of knowledge on types of challenges each method is good at resolving. The following is an interactive solution that asks the Analyst basic characteristics of the challenge at hand then tries to match with expert data mining method knowledge.

Algorithms belong to methods, which are a grouping based on common purpose and features. Data mining methods have developed over the last few decades to extract

insights from data of a given form and with given perspective. As mentioned before AutoML attempts to select algorithms from a specific method.

Context-Aware data mining uses machine learning algorithms to make the final analysis of data. The analysis produces interesting patterns which are interpreted. Data mining covers activities like knowledge exploration purpose, logic, and areas that different methods are good at. At the point of applying these algorithms one must select a method or class (Association, Classification, Clustering to name a few) and a specific algorithm.

Over the years selection of methods and algorithms has been a human based task. Mimicking and tapping from this tough of knowledge and skills is not straightforward. As said before we are looking at ways to enable machines to make decisions. In this section we investigate the history of human decision making, how to teach machines to simulate this and challenges that may arise. Work will also summarise background information on data mining methods which will be fed into the proposed functions.

The exponential growth of both data mining methods and algorithms over the past three decades has created a selection challenge to all levels of data analyst. How does one select the optimum method for a specific dataset? Create a framework to select the appropriate machine learning algorithm that can be applied to a context at hand. Design a Context-Aware Data Mining framework may use for a specific situation for example medical situation and select the best performing Data mining method.

Problem definition is a general term used to refer to what we may be looking at as an opportunity to improve business or life situations. Further exploration of the model building task is detailed at the beginning of this chapter. In chapter “3.0 Literature Review” we discuss research papers on all steps and decisions made during data model design. Research solution, proposed Context-Aware Data Mining Framework will be based on past developments on all components of a data mining model. Context-Aware Data Mining approach proposes making environmental features a foundation of the established mining activities like algorithm selection strategies.

The thesis aims to introduce an extra contextual perspective to the Traditional Data Mining approach. The proposed approach covers frameworks, algorithms, and data

mining strategies. Environmental consideration needs to be part of the understanding and correct application of data mining techniques, called data mining methods.

Background information on methods and areas of application assist in solution design. Libraries of procedures, methods, and algorithms currently available in the Open Source had immensely helped shorten project turnarounds. Research should promote use of Open-Source solutions and tested techniques. The novel proposed framework system attempts to automate method selection in Data Mining projects. Importance of data mining strategies has increased by sharing technical knowledge like codes and scripts in this Internet of Things Global Village.

A script will be prepared to match situations or projects to machine learning algorithm class. The solution developed will be presented through flowcharts and pseudocode. Next step is to select a method to be applied. As said before, this could be done using the decision-making power of modern software.

4.3.1 Method Selection Phase

The first step of the proposal is to select methods to be used, as an automated process we should understand the data mining problem. The following discussion will use one method as reference, note the same consideration should be made for any of the seven main techniques. Other researchers improve development of methods and their implementations. Consider Association rules mining that extract relationships between items in focus.

The function should promote association rule mining if the issues could be resolved by knowing relationships between attributes like items being bought together. (Wang & Liu, 2011) explored improving Apriori algorithm proposed (R. Agrawal) on results, processing time and efficiency. At point of implementation decisions may need to be made using the basic algorithm or improved versions. Fuzzy quantitative constraints (FQC-wed Apriori algorithm) researched by (Lu & Sheng, 2013), tried to enhance Apriori algorithm by giving varying importance to rules and items using a weight system.

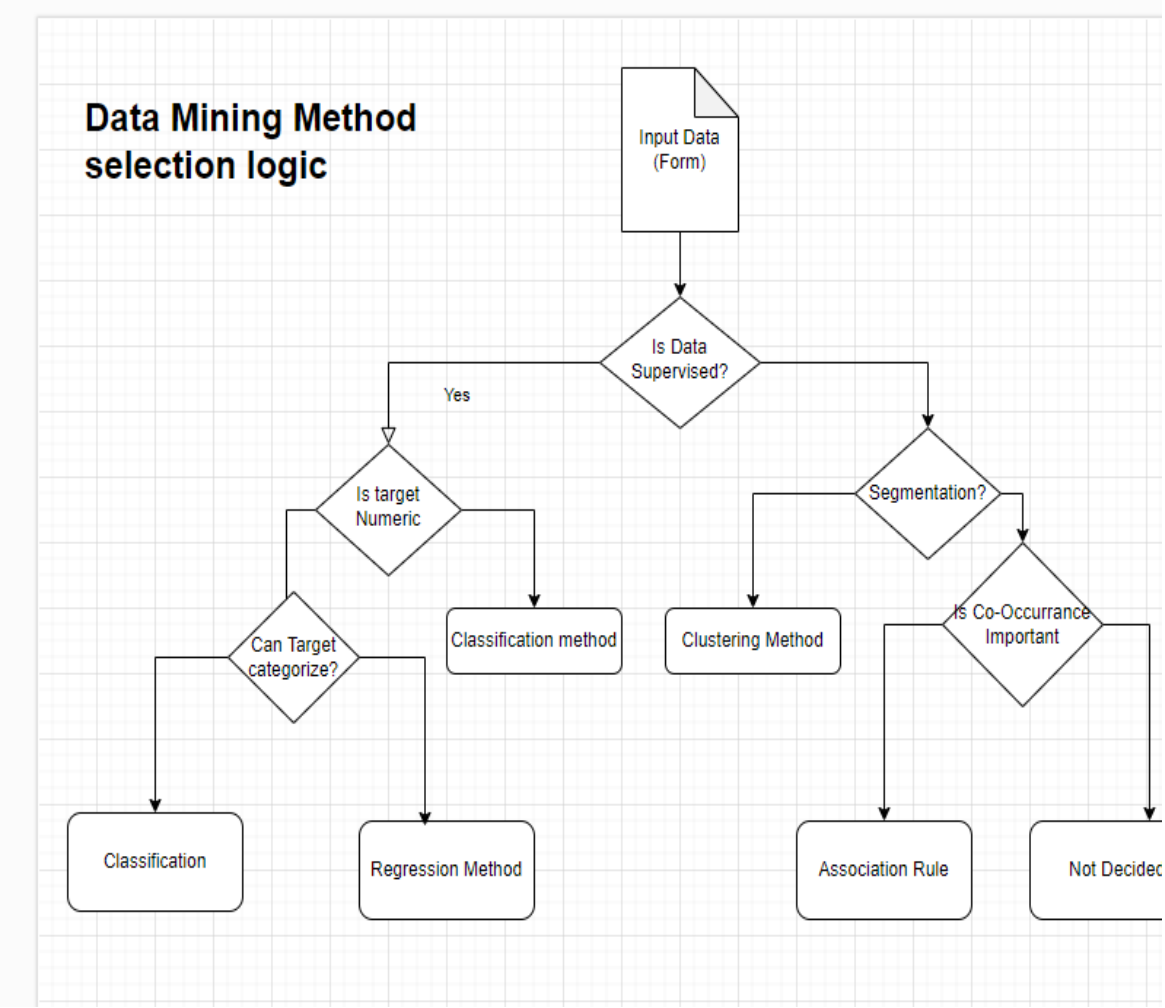


Figure 15:Method Selection flowchart.

Analysts make lots of decisions once an opportunity and data are presented to them. Fig 15. Propose a basic decision sequence to establish the technique likely to fully analyse a given dataset. Analyst considers the nature of classes provided and purpose of the investigation.

The components selection process uses a sample of the dataset, once completed the Data Mining framework will be evaluated on the complete dataset. It is noted that the model design should be context aware in relation to data and purpose of analysis. Nature of problem and size of class of interest might determine selection of a classifier. (Maroco, et al., 2011) assessed that nonparametric classifier like Random Forests and Discriminant analysis outperformed traditional classifiers like Support Vector Machines, Neural Networks Classification trees when class of interest was exceedingly negligible compared to the population size on a binary classification problem.

The knowledge on algorithms provided by other researchers will be loaded into a database with clear information on assumptions since knowledge is a living subject ever growing and changing. The assumptions act as a tool to make our solution future proof.

The phase of selecting a Data Mining Method is controlled by an algorithm as the one below. This algorithm can then be translated into different language like Python or R.

```
If resultsource = predict_record_label                    // Main if
        DMbranch = Supervised              // Supervised data mining issue
        If target type = numeric        // Target type if    // Check if we need Categorization
                If target_structure = Categorizable   // Categorization if
                        DMmethod = Classification
                Else
                        DMmethod = Regression
                End categorizable If
        Else
                DMmethod = Classification
        End target type If
Else
        DMbranch = UNSupervised|                // Further analysis of Unsupervised dataset
        If resultsource = attribute_sergementation
                DMmethod = Clustering
        Else
                If resultsource = Co-occurrence                 // Occurrence if
                        DMmethod = Association
                Else
                        DMmethod = Unknown
                End occurrence If           End Unsupervised If
End main If
```

Figure 16:Pseudocode of a Data mining Method selector function.

The above Fig. 16 logic has a situation ending with an unknown method, has not designed to decide on situations needing ensembled methods. It also does not take care of Semi-Supervised solutions. If time permits, advanced form of this idea may be used to solve these situations. The approach also considers ensembles that use more than one method and algorithm for distinct stages of the analysis.

Next step is to select optimum methods including Clustering, Classification and Associations classes as the main sets from which to select building blocks. Other four methods are allocated supporting roles, also depending on nature of problem and data the tool should be able to get varying results. This is proof-of-concept research so the mentioned blocks can be replaced by different blocks depending on subject and future tools developments.

Construction of the framework encompasses a decision on the purpose of each method, this can vary from time to time. Any developer following this framework may produce different main methods and their purpose. It is a flexible framework that can be applied in future as new methods and data challenges arise. Emphasis should be put on the logic of these new functions more than the outcome of the functions.

For each block we perform Phase 1 comparison to select the class likely to return favourable results. In phase 2 further comparison of members of the selected class are evaluated to produce the most performing algorithm. Effort will be made to fine turn this final algorithm for improved results. All together at least nine comparisons will be performed. Data mining techniques can be split into supervised and unsupervised learning. This focuses on whether the range of selections is provided (supervised) or unlimited (unsupervised). There are other techniques which are an intersection of the two splits.

Looking at some popular methods will inform the broader area of focus of this research. The subject of data mining algorithms is extensive so the project will discuss them at an important level only. (Delen, Walker, & Kadam, Predicting Breast Cancer Survivabilty: a Comparison of Three Data Mining Methods, 2005) compared Artificial Neural Networks (ANNs), Decision tree(C5) and Logistic Regression using a 10-fold cross-validation process. This involves ten subsets of the dataset and averaging of the classier accuracy found at each cycle. Delen's experiment produced 93.6% accuracy for the decision tree but importantly illustrate the use of data as part of method evaluation.

4.3.2 Algorithm Class Selection Phase

In simple terms once key information on the challenge at hand and characteristics of the method are accessed a logic can be used to perform a matching process.

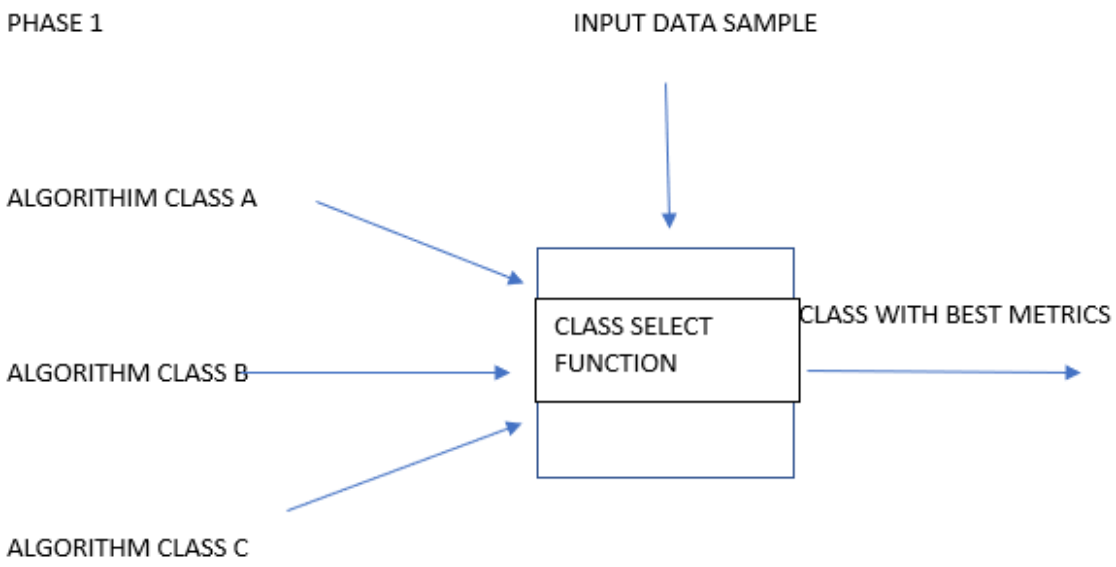


Figure 17: Algorithm class selection function.

Algorithm Class A represents all key information needed to know what type of data and situations it can resolve. This can even be detailed as to what output the algorithms in the class can produce. Input Data Sample also stands for all information pertaining to the data. Given the volume of algorithms per method, the class selection step (Fig. 17) is used to reduce input of the next phase. The process is staged, getting to be nearer to solution after each step, performing the above for Clustering, Classification and Associated to identify classes that are likely to contain algorithms to be added on the Data mining framework. The functions will be created for each method, tested, and only joined once they work well.

This is followed by use of a function (function CS), to select a probable class of algorithms. Selected class is likely to produce best results based on both sample data and metric performance of representative algorithms. Function CS accepts that each method has several classes that are composed of five or more algorithms. The assumption is that an algorithm standing for a class will perform to the level of other members of the group. This algorithm works as a pointer to the class that is likely to have the best performing one. The final phase will use a function (function AS) to select individual algorithms to put into the framework.

4.3.3 Algorithm Selection Phase

At the algorithm level, selection options have been reduced to a set under a given class. These solutions have been previously created by other researchers. This thesis accepts that magnificent work has been done so it should be enough to just mention this aspect of algorithm search.

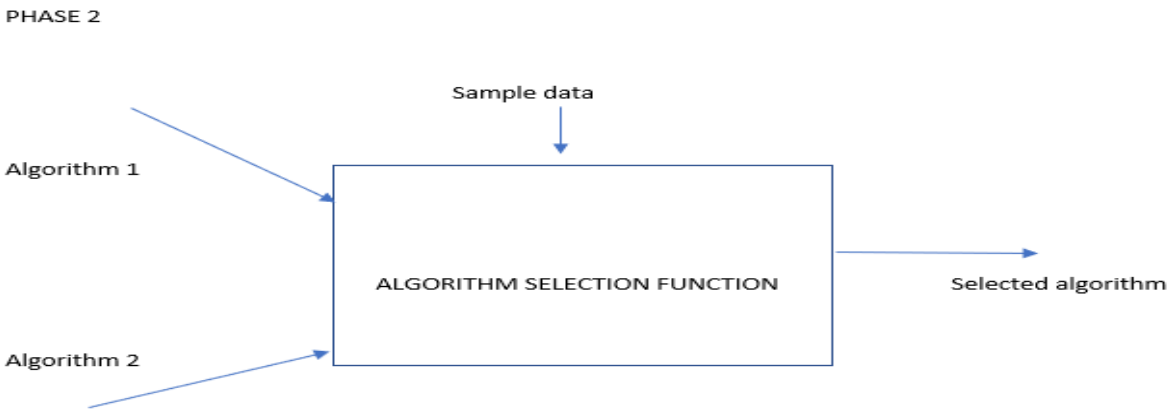


Figure 18:Final algorithm selection function from class.

The above function (Fig. 18) takes the sample data, algorithms in the selected class to make a performance comparison rank. The top performing algorithm is promoted to be used on the whole dataset for further analysis.

Most data mining frameworks use known subjective characteristics to select algorithms. The proposal will promote on applying functions to improve this selection process. History of simulation systems promote the notion of streamlining algorithms and grouping them based on grounded knowledge on their design and characteristics. This should be enhanced by some comparison functions.

Note this is a proof of concept, future application can produce varying components. The prototype should be able apply new comparison parameters as they are developed. The complete Data Mining Framework with placeholder classes and algorithms. The same growth will also happen with algorithm comparison metrics. Many frameworks have been developed where others are just collections of algorithms such as ELKI and WEKA.

Data Preparation may include visualisation, statistical analysis, data cleaning, removal of noise and other Data Mining methods that can help focus like outlier detection. Key to this proposal is a correct sampling method, so a limited introduction to sampling will be included. Clustering, Association, and classification features selected classes and algorithms.

The research success depends on understanding the breakdown of data mining strategies. A few definitions and exploration of design considerations of data mining techniques will be conducted to inform the decision-making functions. Below is an outline of the established techniques.

There are several Data mining techniques with the top three being Classification, Clustering and Association. Other than these three there are four more main methods with possibility of more coming on board in future. The research will also explore varying purposes of these methods in each model.

There is a need to have a basic overview understanding of the highest-level split of data mining strategies. It should note that they serve different purposes at varying instances

depending on the state of input data and the problem at hand. A method might be used to prepare data for further processing by another technique. In cases, algorithms used in an ensemble. Ensemble is a set of Data mining Methods being used together to extract patterns.

4.4 Application of Contextual concepts to model creation and insight interpretation.

After data has been prepared, methods and algorithms selected the ultimate step is running their model to get insights. Gathered environmental knowledge influences the running of the created model. The results insights are also interpreted in view of the context in which a dataset was found. The use of context may range from terms related to fields that are used to describe insights.

A medical dataset will interpreting medical metrics. A financial dataset insight will use financial themes and thematic. Data analysis based on numeric values is divorced from subject area terms so will be too basic.

Final Process may include visualisation, other interpretation Data Mining algorithms like Regression and Prediction. Once selected, effort should be made to improve the algorithm. The way methods work is dependent on context under investigation and nature of data. Raw data may need pre-modification before feeding into the framework.

The use of contextual knowledge throughout the data analysis will reflect in the direction and content of insights. There might be a need to define any assumptions made around the subject of analysis to increase the usefulness of resulting insights.

There are solution design challenges in keeping it simple but fully integrating situation factors into design of a solution given their volume and applications points. As a precondition for processing data from varying sources need to have a common grouping key to aid processing.

Algorithm performance including accuracy, computation resource consumption and evaluation approaches like 10-fold validations were explored. The emphasis is on understanding key concepts and the semi-automation process. Situation linked data analysis tends to produce actionable knowledge on a subject. Data mining is nowadays producing more usable results, credit to advances in model designs, this research wishes to contribute to the area.

The proposal seeks to explore promotion of Contextual Features during Data Mining. Research done on data mining frameworks, comparison, and design of algorithms, but these tasks have not been integrated into building of applicable solutions. The knowledge gap is around making context a major solution design consideration. The prototype can be reused in varying analytics projects with increased chances of success. This knowledge is used as input to model design effort, with interfaces and level of semi-automation.

The project will explain design, logic, and modifications of algorithms. Project should be clear on features of the dataset and aims of the experiment to justify the implemented alterations. The new applications are evaluated by experts and novice data scientists. Technique comparison (Lakshmi, Kumar, & Krishma, 2013) process may include accuracy, computing time, specificity, cross validation error and bootstrap validation error. There are more than ten comparison metrics, but the project will have to guide on their importance. More work may need to pursue this challenge.

There is also the formation of new hypotheses on the subject. An evaluation criterion for the model should be specified from the commencement. The model can prove, disapprove, or contribute knowledge on the hypothesis. Context-awareness approach encourages one to start by analysing the environment in which the dataset resides. Compare that human expertise influence stops before analysis whether there are.

Investigate benefits of context-aware concepts by comparing performance with established frameworks. High level recap of approaches like Deep learning to show cause why the context awareness effort justified. Summary of ways in which Context-Aware approaches increase output relevance to dataset subject matter. Measure the increase

of focus and relevance caused by use of Context-Aware Data Mining. Measure amount of trimming and process streamlining emanating from application of context concepts. Illustrate improved quality by measuring how output is related to the world around the dataset.

Measure level of relevance from results of implementing the two approaches. Measure possibility of trimming when using Context-Aware concepts. Justify an extra effort required by the novel framework.

Record level of pruning resulting from using context-aware principles. What is the original dataset size? What portion of the dataset will be considered for analysis after pruning based on CMAS methodology. Compare relevance and ease of interpretation of insights from the CMAS based analysis.

Accurate description of specific action to achieve aim in an infinitive sentence. Analyse the contribution of focusing on context during data mining by comparing different frameworks.

Computer processing advancement and software decision making logic should enhance model building and design. We borrow key features of existing applications to build user interaction and user interfaces. The research should not reduce data scientist's creativity, but free their effort to focus on other parts of the knowledge extraction journey. It will complement human data processing skills rather than curtailing it.

4.4.1 Benefits of Contextual concepts

Thesis is motivated by creation of data mining solutions that are focused and reflect real-life. In real day to day life where minor ecosystems are influenced by multiple large ecosystems. Opportunity to develop a framework that defines local data and context surrounding the local situation. There has developed the need to create a more dynamic data model, which reflects real-life interactions for multiple factors to any event. A proposed framework can simplify the basic context-aware data mining process to allow for wider application of the approach.

Context-Aware data mining improves the quality of resulting insights. Quality of insights range from relevance, focus, applicability, accuracy, and ease of interpretation. Context

linked data analysis tends to produce actionable knowledge on a subject based on analysis focus. Data mining projects result quality depends on the right levels of focus of effort being applied on a clear scope. Context is a human based definition of a given real-life challenge. Context can, for example, be taken to be the surrounding environment in which the local data occur.

The possibility of creating an application that can select data mining techniques and algorithms, using the context of data and problem area is interesting. Semi-Automated selection of data mining methods simplify and create more appropriate data mining models. The selection software can access accumulated advanced data mining techniques related knowledge.

Research is motivated by the opportunity of using computer processing power to take volumes of principles and problem details to produce key parameters for Context-Aware Data Mining. Computers are capable of crunching lots of logic at high speed. In future these can even be self-learning tools that improve the application over time.

Research is motivated by benefits to novice analyst if overwhelming volumes of redundant insights can be avoided. The overwhelming volumes insights are both time consuming to go through and may lead to wrong interpretations. The wide range of both Data Mining methods and algorithms increase the possibility of overwhelming both domain specialist and novice Analysts. Research in this domain, in combination with technological advancement is likely to yield benefit by suggesting better solutions. Current methods are Association rules, Classification, regression Clustering etc or combinations. Solutions are supposed to match key information on an issue to the known strength of methods as input to a deciding script.

Chapter 4 Summary

The novel CMAS is built on the Traditional Data Mining fundamentals with additional perspective of contextual features. Every step of analysis adds assessment and implementation of environmental related features.

CHAPTER 5 Experimentation / Case Study

Chapter 5 Introduction

We have used three case studies to illustrate the proposed CMAS and will be explained in detail in this chapter. The framework follows six key stages from understanding an

issue to communicating insights. There is an automated data mining method within the framework. Chapter also expands on logic and application in real life settings.

5.1 Context-Aware concepts implementation

The proposed CMAS detailed in Chapter 5 is applied to a medical use case PIMA Indian Diabetes dataset, traffic use case India Accident Severity dataset and finally a humankind crisis World Air and Water pollution data set. Data Mining is centred on techniques like Classification, Clustering, Regression, Association etc.

The first step for the IT Specialist is to understand the issues at hand. What is the situation? What are key contextual factors? Who are the key audiences of the analysis effort? and what are their main expectations. Answers to these pertinent questions guide development of a solution with the situation in mind.

5.1.1 Contextual factors transformation

Context-aware approaches require extra decision on points to apply context in the data mining life cycle. Gather knowledge to allow one to decide on the phases appropriate to apply contextual factors for a given problem. Contextual factors may be used at every data mining stage as illustrated in CMAS or in a single stage of the cycle. In the three use cases detailed later emphasis is put on these points and their justifications.

Define processes to transform expert knowledge, features in a dataset or from sensors into a context that can be used in the proposed model. Find methods to define context into applicable form. Discuss difficulties and challenges encountered during simplification of these complex concepts. develop logic that puts together information from various sources to make key Context-Aware Data Mining solutions decisions.

Data mining produces interesting insights from large datasets. The proposed framework accepts that the dataset exists in a large environment. Context consideration may involve transformation of dataset features or other external information to define the context. The

following is application of the proposed framework in a real-life setting. A full cycle of data analysis done using Context-Aware Data Mining Framework to illustrate the methodology.

5.1.2 Challenges with CMAS

Issues arise from various sources like commercial transaction, science or medical based. Issues described from the perspective of stakeholders. Context-awareness concept is a specialised sub-domain of Data Mining. Context-Aware is a data mining methodology which integrates the situation (environment), around the subject of investigation. Data mining attempts to extract interesting insights from datasets. Full picture of the context-awareness concepts can only build through definition, history, discussion of components of data mining. It illustrates all stages of integrating environment (Context) into appropriate data analysis stages.

Main objectives of pruning, focus and increase output relevance demonstrated with clarification of outcome and modifications of standard analysis tasks. The research promotes a qualitative analysis based on framework evaluation by Data Mining experts.

Medical research and understanding of disease are as old as humankind. The issue is around understanding of biomedical metrics from both diabetic and non-diabetic populations. So, these metrics present a local dataset. The context is expert based and around a specific population factor. In defining the context, a worldwide view is clearly documented, and expert population specific context is also defined. Once Context is defined the proposed framework is used to analyses the dataset. A table of results is produced with analysis of the effective approach.

5.2 Case Studies

The thesis applied CMAS approach to a PIMA Indian diabetes dataset, health challenges in the world need to be resolved by further situation understanding. The pattern investigation can leverage data mining technologies. The health challenge and that Indian population’s unique characterises were pulling factors for this thesis.

The second dataset India Accident Severity problem had many context interactions in varying ways. We were also interested in ways to serve a wide range of stakeholders that had varying requirements. It is interesting to check ways to process data of large dimensions. One approach is to tailor make analysis for given stakeholders and drop features not related to the trimmed requirements.

The Third dataset World Air and Water Pollution, humankind needs to understand issues that are contributing to loss of life. From a data analysis perspective this study handled the challenge of building context in line with experts' views of an issue. Data had to be sourced and pre-processed to assist in the investigations of key patterns.

Below we analyse each case in detail.

5.2.1 PIMA Diabetes Dataset (Association Rules)

5.2.1.1 Contextual Information integration decisions

PIMA Indian diabetes dataset background

With an estimated (Standl, 2021) 463 million living diabetes in the world, also dealing with complications ranging from blindness, kidney or an increased 70% chance of death from CVD (heart problems). Its occurrence is surrounded by many biometrics measurements that have been widely researched over the past decades. Inspiration to analyse is based on the nature of the disease and the of knowledge around it. The experiment was performed around this subject context of the disease in a particular population. This is a dataset of medical information of a particular population, Pima Indians.

Interest is acceptance of context linked to both disease and this population aiming to get insight informed by expert knowledge. It is hoped that mixing accepted expert knowledge, advanced data mining techniques and focused analysis will yield relevant insights. The datasets consist of several medical predictor variables and one target variable, outcome.

Predictor variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

Source from the Kaggle data repository (Smith, Everhart, Dickson, Knowler, & Johannes, 1988). Dataset is on the 'Pima Indians Diabetes Database' collected by the National Institute of Diabetes and Digestive and Kidney Diseases. With diabetes outcome it can be used to predict status using diagnostic measurements in 8 columns. In this analysis a different angle is taken around factors that occur together for given context. Note the population is of Indian Pima heritage.

Research aims to explore basic principles on Context-Aware Data Mining to apply to several real-life scenarios. Effort needs to be on furthering knowledge on key challenging new processes like (windowing, time series, data summarization) to improve relevance and interpretation of analysis results.

The research attempts to develop a data model that can analyse several situations, data taking into consideration influencing events from surrounding. It is likely to help novice and professional data scientists to apply context-awareness principles with improved clarity. This aim is difficult to prove but any further discussions in a subject is likely to yield growth in the pool of knowledge. Aim achieved by meeting the below objectives.

This research is concerned with simplification of multi decision points in data model building. Complex knowledge processing might need some level of automation and translation of theorems into functions. Needs to situate or point to the area it focuses on. Product could be design of models by further understanding decisions construction and possibility of automation openings.

Actionable insights focused Process-based Domain-driven Data Mining – Actionable Knowledge Discovery (PD3M-AKD), developed by (Fatima, Talib, Muhammad, & Awais, 2020) by applying multiple environmental factors in the framework.

Data mining process has matured to a point where the framework and outcomes can improve by considering the context in which the subject dataset exists. Context, environment, or situation is taken to mean the surroundings of a dataset. Note these surrounding may be dynamic, changing during the lifetime of the dataset.

The question is whether these context changes need to be incorporated into the analysis. Context makes analysis more linked to the real world in which data exist. It could even be an analysis improvement to state the context in which a given analysis is.

PIMA Indian diabetes dataset

The PIMA data has nine features on data surveyed on a limited population from a given group of people. One feature is later dropped since it is not of any use to our current analysis. We look at the features in the table below:

Feature	Medical definition	Value Range

Number of pregnancies	There might be a of volume of pregnancies and having diabetes	0 to 8
Plasma glucose concentration	Plasma glucose concentration tested more a two-hour oral glucose tolerance test. Between 70 and 130 mg/dl milligrams per decilitre normal range	70 to 200 mg/dl
Triceps skinfold thickness (mm)	Amount of fat reserves in the body	18.7 (+/- 8.5) mm
Age (years)	Age of participants given most diseases kick in with age	0 to 90
Diabetes pedigree function	Family line affect the likelihood of developing diabetes in life	1 to 0
2 – hour serum insulin (mm U/m)	High level signal insulin resistance low point to diabetes or pancreatitis	2 to 20 mcU/mL
Body mass index	That is weight in kg divided by height in m2.	18.5 to 24.9
Blood pressure	Normal being systolic less than 120 and diastolic 80 mm/Hg	120/80
Diabetes Status	Whether person is diabetic or not	0 or 1

Table 3: PIMA Features Summary

Table 3 summarises the features of the PIMA Indian Dataset before any context related transformations.

Sample Dataset

	A	B	C	D	E	F	G	H	I	J
1	Key	Pregnanci	Glucose	BloodPres	SkinThickr	Insulin	BMI	DiabetesP	Age	Outcome
2	1	6	148	72	35	0	33.6	0.627	50	1
3	2	1	85	66	29	0	26.6	0.351	31	0
4	3	8	183	64	0	0	23.3	0.672	32	1
5	4	1	89	66	23	94	28.1	0.167	21	0
6	5	0	137	40	35	168	43.1	2.288	33	1
7	6	5	116	74	0	0	25.6	0.201	30	0
8	7	3	78	50	32	88	31	0.248	26	1
9	8	10	115	0	0	0	35.3	0.134	29	0
10	9	2	197	70	45	543	30.5	0.158	53	1
11	10	8	125	96	0	0	0	0.232	54	1
12	11	4	110	92	0	0	37.6	0.191	30	0
13	12	10	168	74	0	0	38	0.537	34	1

Figure 19: Pima Indian Dataset

Sample of the local dataset in Fig 19. Data has the above eight health indicators (Pregnancies, Blood Pressure, Skin Thickness, BMI, Diabetes Pri Degree, Age), A key to refer row and Outcome which is 1 when diabetic and 0 will patient in non-diabetic.

Could be looked at as a binary classification issue, but this experiment will try a new twist where the classification column may be used in context definition.768 Observations,8 Input variables and one output variable.

Missing value encoded with values. The data is in numeric values, which might change into ranges or other medical meaning after the initial analysis and expert knowledge research.

Predict diabetes in the Indian population given basic data. Find relationships of features values for diabetic or nondiabetic subjects. What environmental information or data analyst should use to inform the model design process? Gather and list documents of professionally agreed facts on each feature in relation to diabetic status. This may cover medically defining each feature and linking it to the disease. Explore an Indian specific knowledge on each feature.

With reference to chapter 4. CMAS Framework, a solution was implemented in three phases: Context-Aware concepts application points, data mining method selection assistant and sample context-aware implementation.

Data and context used is diabetic dataset for Pima Indian population as sourced from (National Institute of Diabetes and Digestive and Kidney Diseases). Restricted to 768 females of child birthing age of data number of rows, main column. Dataset available on Kaggle site and UCI machine learning repository (Smith, Everhart, Dickson, Knowler, & Johannes, 1988). The main context being developed from diabetic diseases, expert knowledge around it, special knowledge on Indian diabetes metrics and the international (WHO) disease categories.

Data and context data integration

First main stage in the CMAS is 'Contextual Information integration,' in other word match breakdown level of local data and any context features. As shown here this is also a multi-step stage.

Define features, simplify context with expert facts on Indians and diabetes disease. Define scope and direction of project based on context and experts' knowledge. Translate features to context. Use context to predict or inform a situation, in our case predict diabetes in a specific population. Check contexts for each of the nine features to create investigations, guiding statements, or some form of analysis scoping. It may even be relationships between features helping understanding of the main subject e.g., diabetes. For example: Given BMI in Indian ladies need to understand or scale differently from international categories as follows.

Skin thickness relates to diabetics in the following way: -This could lead to transforming values of a feature A into just 5 blocks each with medical meaning helping to speed analysis and bring results near to real world activities. You may find the transformed data works well with analysis methods, life decision tree or some association analysis.

Recorded with gender and age in mind BMI weight KGs divide height in metres squared. Other lifestyle factors need to be considered like fat and muscle for sportspersons who

might have BMI values linked to Obese when they are healthy. Table below gives the currently agreed breakdown of MBI measurements.

Categories	WHO international ranges kg/m2	Indian Women kg/m2
Normal/ healthy weight	18 to 22	18.5 to 21
At Risk weight	23	21.1 to 24
Obese/overweight	23 to 24.9	24.1 to 30
Severely Obese/Obese range	>25	>30

Figure 20: Context of BMI categories.

Fig 20 show breakdown which tone medical practitioners' advice. Obesity is inferred from levels of BMI as reflected on the above table. WHO has internationally adopted range but for South Asia female's lifestyle and genetic based body frame necessitate a different scale (Mohan, Mariamichael, & Kishore, 2011)?

Looking at the fourth row above, 'Severely Obese' patients have BMI above 25 kg/m2 according to WHO experts, but for the PIMA experts it must be above 30 25 kg/m2. In this case the range from 25 to 30 is put in distinct categories affecting any Data mining effort unless context ware concepts are applied.

Dataset definition is based on a use case, named as local dataset to differentiate it from the context considerations. The context can be sourced from inside the local data or from outside. If there are multiple contexts a ranking and selection may be needed to limit solution complexity.

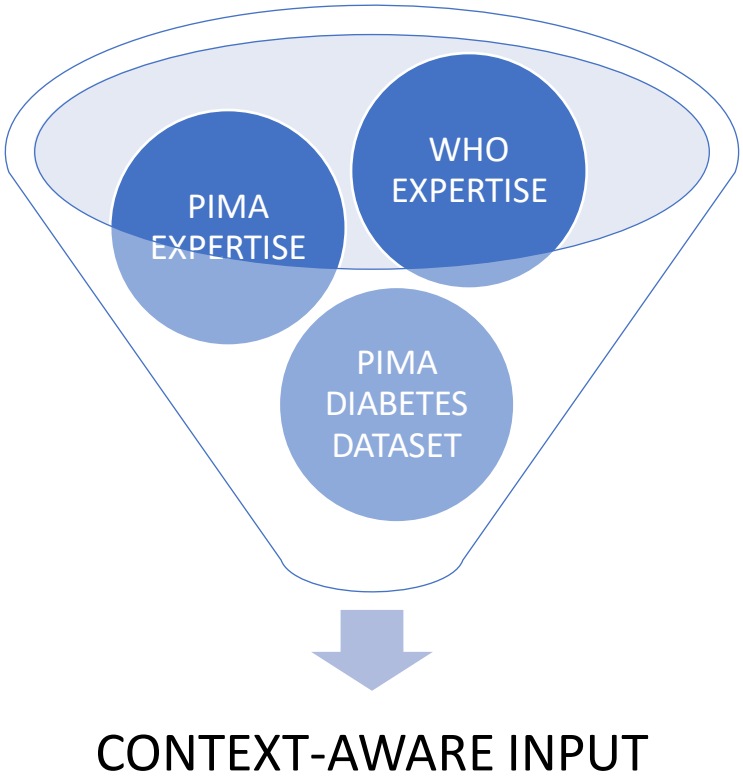


Figure 21 : Context and Local Data integration.

The above fig.21 illustrates that the input to Context-aware model integrates Context, in this case expert knowledge with Local Data (PIMA dataset) to create environment sensitive data analysis. PIMA expertise data is a view that people of this sub-population have characteristics linked to diabetes that vary from the world ones. The WHO Expertise data related to the medically accepted diabetes categories that can be attributed to all people in the world.

5.2.1.2 Context-Aware concepts application points

The data from India population can now be linked to the research medical facts. Once the context has analysed and linked to the local dataset, the framework proceeds to make sets of decisions. Further approaches to using the context in the model need to be made. CMAS framework is of the view that context is embedded on the input data, considered during decision of model, or used during interpretation of analysis results. Not the context concepts can also apply to all three stages, single stage, or various combinations. Key to the proposal is that application review as part of solution design.

Context application points selection assistant script aims to provide analyst with expert knowledge through leading questions. With analysts having knowledge of the situation at hand, the script accepts this knowledge from the given answers. With flexibility the computer attempts to match these choices to agreed routes outputting design suggestions.

Script Based Context Integration

The incorporation of context may also vary in approach. The thesis will introduce few logics allowing for expansion or project specific approaches. The solution is at a high framework level for the PIMA Indian diabetic dataset used in the experiment. The angle or points of applying context-aware concepts are aimed at effectiveness and suitability to opportunity at hand.

The python script (Context.py) has a user-friendly informative interface. It asks the user to confirm information on both context, local data, and project direction of interest. It then uses these pieces of information to suggest context application points. It attempts to give an implementation of the flowchart (fig. 20 Context-Aware DM Decision Points)

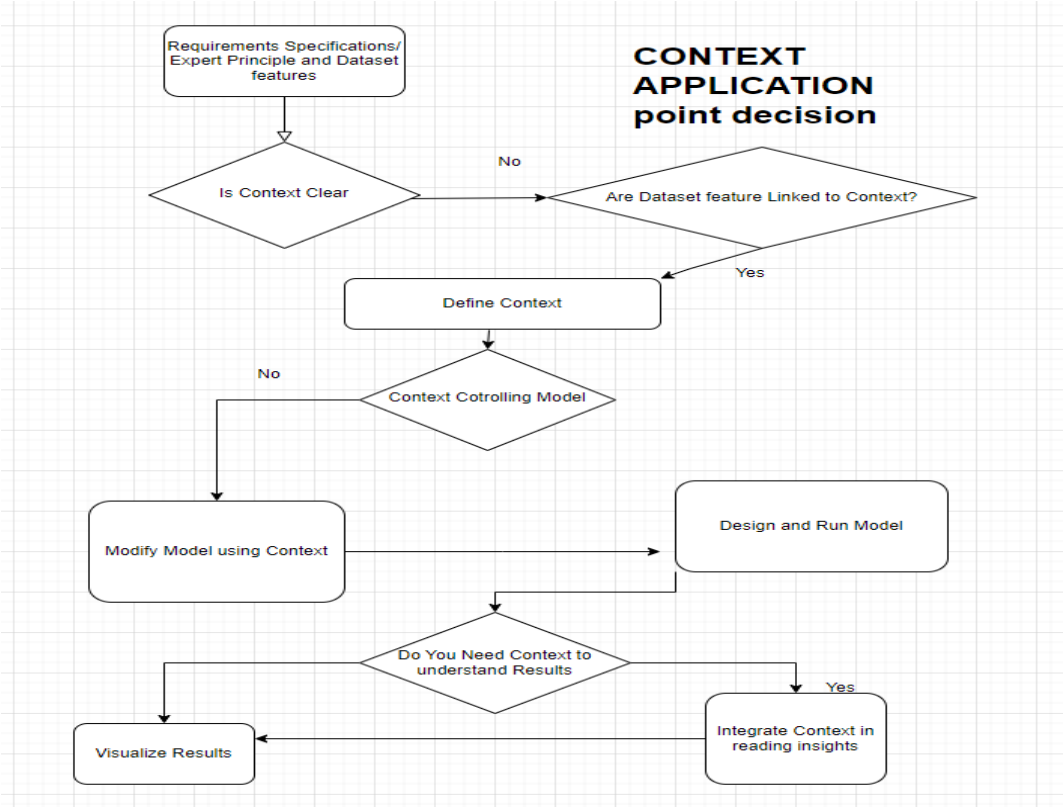
The application user checks boxes and information displayed on the screen to interact with the users. Based on selections made the systems suggest context application points. Data science is an act allowing for immense human based creativity, so the script is innovative, allowing expert knowledge to be shared with the user who has the best view of both the environment (context) and local data.

Above is an example of information provided by the system, options taken by users and suggestions. Information: User Selection: System context application suggestion: These are only one route illustration depending on user selection a variety of suggestions can be reached. The approach is to allow expert knowledge.

- The Context-Aware application point assistant helps in them key decision-making process. It informs on issues to consider when making a context aware solution. The human being process, given knowledge, informs the application on a set of choices. The system in turn uses the input and expert knowledge to map a route through the flowchart reaching an action point. The action points are a guide to the data scientist in designing an environment sensitive data mining solution.
    - This step could be iterated. Make first attempt on problem or issues description.
    - Data mining will attempt to answer two questions. One, how does the BMI category of an individual increase chances of diabetes. Two, is more information to perform BMI analysis having expert background more informative?
    - Other dimensions include:
- With categorised features perform an association analysis to find prevalent combinations
- For context of diabetic patients find prevalent combinations
- For non-diabetic patients find prevalent feature combinations
- Use results of the above to infer biometrics values on the diabetic disease.
    - Summarise all context and expert knowledge.
    - Convert BMI values into Categories based on WHO International Standards and Indian Specific Standards. Perform analysis without context, based on WHO International Standards. Perform context-based analysis, based on Indian specific context.
    - Use results from 1 and 2 to create a context of interest. This context may give scope of the data analysis, reduce features of interest, and create a clear project statement of intent.

- Redefine the problem or issue at hand. Context to be applied and scope as defined by 1,2 and 3.
- Run the input into the Context Application point select script.

The flowchart illustrates the detailed logical process of information being analysed on the Indian Diabetes situations. Basic logic can always be improved to handle more complex settings.



DECISION POINTS IN A CONTEXT-AWARE FRAMEWORK

Figure 22: Context-Aware Application Point Logic for PIMA solution.

Following the double lines and the coloured processes, the applications confirmed that Context is combined with local data, it also affected the model design and results will be defined with context in mind. The solution, designed in three parts, is an interface allowing

analysts to input their view of the dataset and environment. The python script makes decisions that are returned to the product owner.

D1 Decide if the context is complete from expert knowledge and analysis requirements.

D2, should we create Context from some parts of the dataset?

D3 Should Context apply directly into the model design?

D4 Is it just interpretation of results which already contain context or should facts of context apply to the results.

Table below summarises interaction used to decide when to apply conceptual actions in the analysis development cycle. The decision is a combination of knowledge data and query to retrieve it. Discussion might not fully illustrate this.

Input for the Context Application Point Script

Action/ Decision	Yes	No	Sequential Process
1. Summarise expert knowledge, problem story			
2. Is context well defined?	Yes		Define Context
3. Link Features to context		No	Define Context
4. Should context be in model logic?		No	Design and Run Model
5. Modify model Context	Yes		Design and Run Model
6. Need context to understand results?		No	Visualise Results
7. Integrate context into results interpretation	Yes		Visualise Results

Figure 23: Context Application decisions.

As fig. 23 shows the YES or NO answers point to specific action to be followed.

With three process that must be followed no matter input

Define Context, Design & Run Model and Visualise Results

The other three processes only depend on decisions.

Link features to context, modify model logic with context and apply context to results interpretation.

INPUT INTO PIMA DIABETIC ANALYSIS A

Is context well defined? - No

Output – link features to context.

Define Context

Should context be in model logic? - Yes

Adapt model with Context.

Design and run Context.

Should Context be used in results interpretation? - Yes

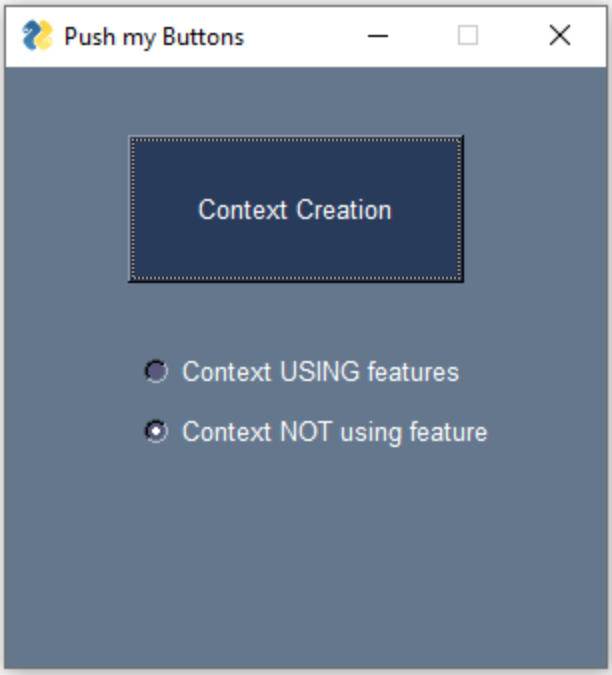
Visualisation of model results

5.2.1.3 Algorithm Class Selection for Context-Aware Data mining solution

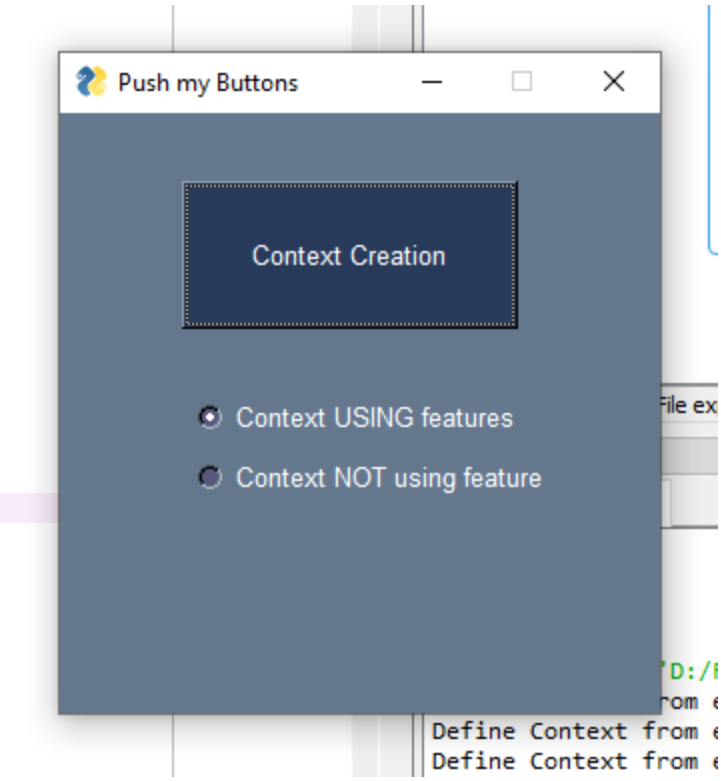
CMAS stage of 'Context framework Method and Algorithm Selection is a technology and human interaction-based phase. Phase is built around knowledge accumulation and accurate retrieval in modern settings.

Basic Python Based Interfaces

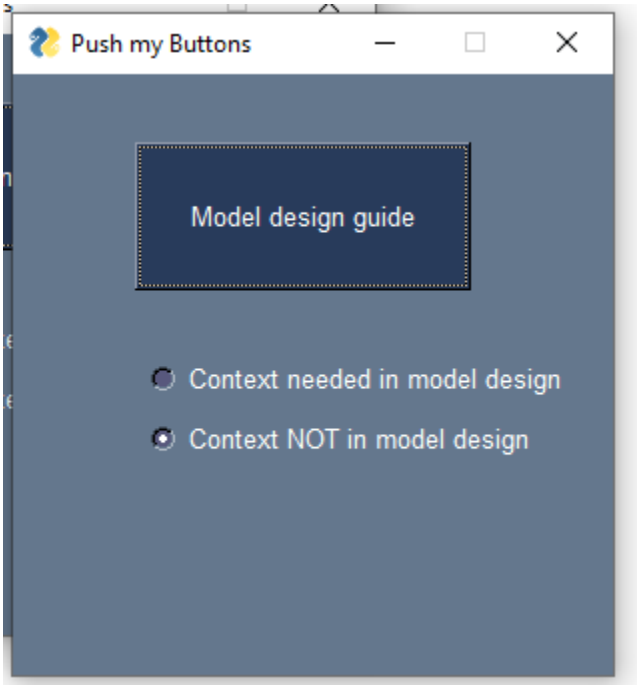
Below is a user-friendly interface to pass data understanding from Analyst to the system. Analyst has a view of the dataset and environment so can answer leading questions. A decision engine is created in the python scripts.



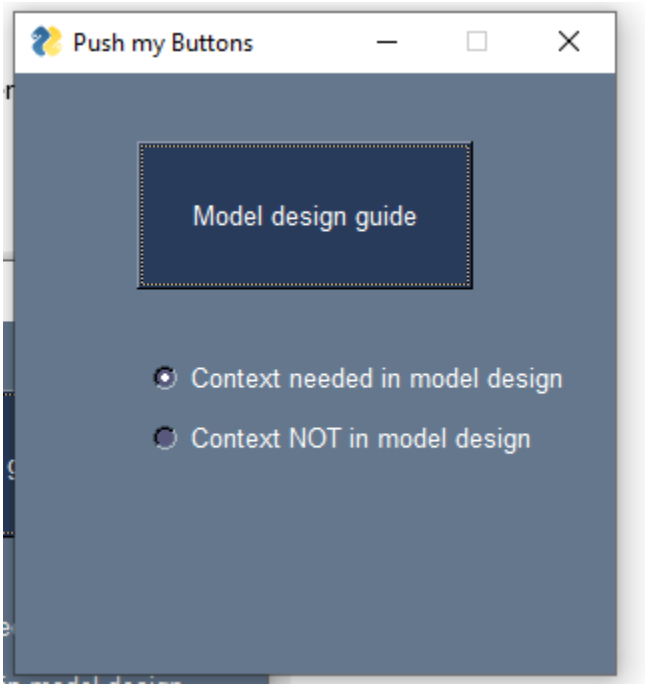
Decide if data features are enough to define Context. Above we have decided that the context is not defined features but by other knowledge including expert knowledge.



Selected Features are enough to define the context.



Decide if the Context should be used in the model design?



Decided that the Context can be used to design the model.

The PIMA Indian data challenge can be put through a user-friendly interface that asks key questions, goes to refer to a knowledge database and feeds back some suggestions. The thesis just scraps in the idea as a direction of development effort.

Next stage is to align the gathered Indian PIMA knowledge to an appropriate data mining technique. Once the nature, characteristics and direction of investigation is fully defined, the problem can classify as Supervised, Semi-supervised or Unsupervised. The Indian Diabetic data issue could be classified classify as suitable for Association, Classification or Clustering data mining method. If the aim is to predict a patient diabetic status, then using the given labels one can use the Classification method.

If on the other hand investigation is thrust to numerous groupings of characteristic, attribute values to explain diabetic conditions, then Clustering methods may be pursued. Lastly Association method may follow with some data transformation if the relationship

between attributes is able to give more information on the population and the diabetics disease.

Following logic detailed in Fig. 22: Method selection Flowchart and Fig. 14: Pseudocode of a data mining method selector function, the python script helped to make the following decisions.

Is data supervised?

First decision is on the problem and the dataset to decide if data analysis produce interesting knowledge. A script will ask analysts to confirm if labels given and informative to be further investigated? Data could be defined as Supervised since it has labels. The Actor, using on initiative, felt more knowledge could be found by ignoring the labels. Note the strength of the framework is mixing expert knowledge and human real time input.

Given the target is numeric?

The target is not to be numeric. It was decided that the aim is not predicting whether a patient was diabetic (1) or not (0). Therefore, this problem was not suitable for the Classification data mining method.

Is segregation important?

If segregation of the diabetic patients by attribute is important, then the analysis should follow the Clustering data mining method.

Can Categorise targets?

Diabetic data target checked if they could categorise, which could have led to use of Classification Method or Non Categorizable leading to use of Regression methods. Given the earlier decision of not using labels, these options were not considered.

Method Selection Buddy logic

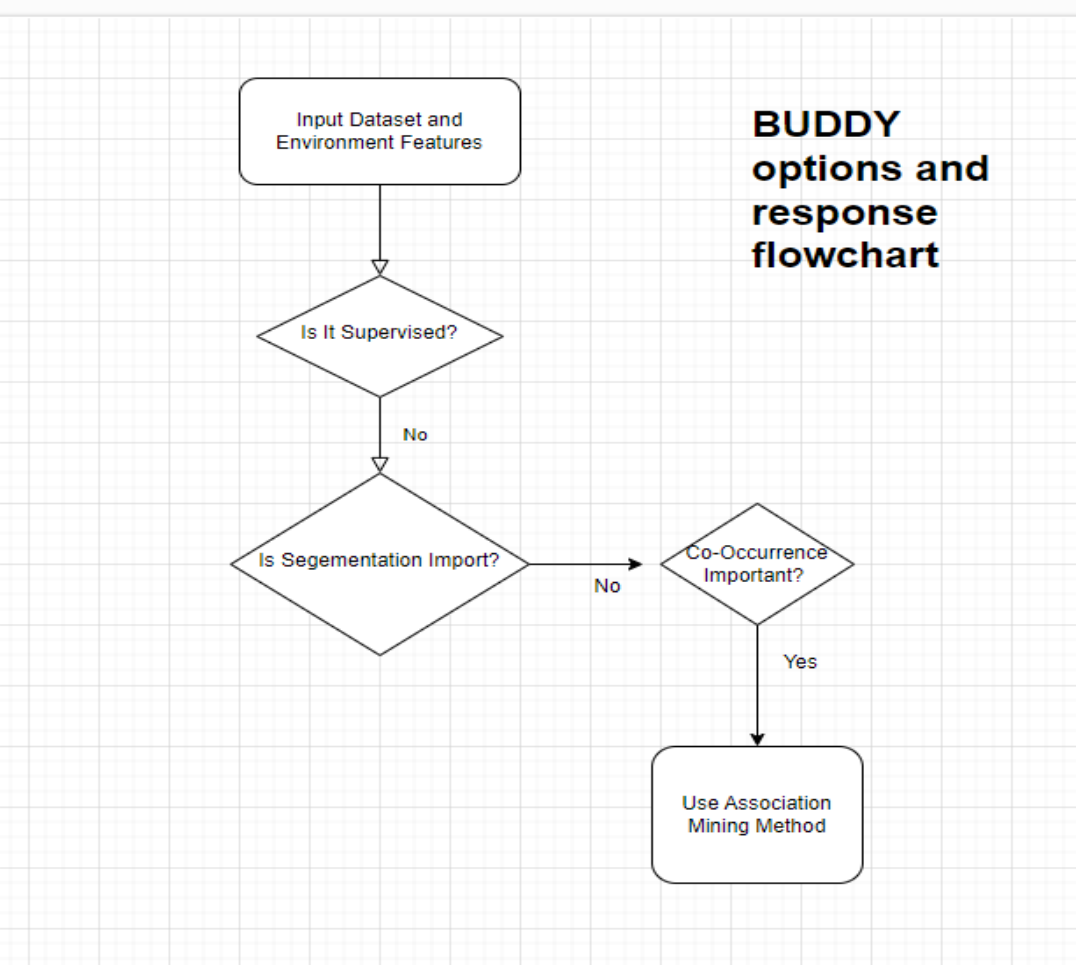


Figure 24: Data Mining Method path.

PIMA DM method selector

As reflected above the analyst has given information suggesting that the issue has no result labels, separating and grouping attributes does not help. The logic has also concluded that insight might be extracted from knowing categories that occur together.

Is Co-Occurrence of Attributes Important?

Questions on the problem at hand

Question	Decision	Method
Is Issue Supervised?	NO	
Segmentation Important?	NO	
Co-Occurrence Important?	YES	Association Method

Table 4 : Script Question and Answers

As reflected in the above table 4, the script managed to decide on Association Method, by marching expert knowledge to Data Analyst answers to some leading Questions on the problem at hand, Diabetic analysis for an Indian population.

Finally analysing whether co-occurrence of some attribute values being informative on the diabetic disease in an Indian population, the team (Analyst and Analyst assistant application) decided to use the Association data mining method.

5.2.1.4 Association rule mining method

System developed in a Scientific Python Development Environment, Spyder Ver 3.3.3. Python scripting languages used based on many machine learning packages already added to it. Scripts show various data mining approaches once Context is considered. As mentioned before, local data, or focus of an analysis, exists in an environment, which when ignored reduces the significance of the process outcomes.

This script illustrates the power Context-Aware approach when applied to Data Mining. The main context is around ways WHO World Health Organization categories BMI values and that specifically used of PIMA Indian populations. Note the categories are used for the whole worldwide population. Illustration should show improvements in relevance by referring to expert knowledge on a specific population background and grounded scientific knowledge.

Second context co-occurrences of biometrics for populations that are either diabetic or non-diabetic, then comparing the two to gain more knowledge on each grouping. The findings may help understand attributes related to either of these datasets.

In the next section is a summary of model design and development. Scripts and detailed code documentation can be accessed in the appendix.

5.2.1.5 Context-Aware Data Mining Model Development (Prototype)

Once the decisions above have been taken by the framework, implementation moves on to development of the data mining model. In this section key design and implementation steps will be detailed. The solutions are designed as four models, two for the WHO context and the other two for the PIMA INDIANs expert-based context. These could have been done as a single model with added complexity.

Machine Learning Packages

```
from mlxtend. frequent_patterns import apriori
from mlxtend.frequent_patterns import association_rules
from mlxtend.preprocessing import TransactionEncoder
```

With a bit more theory, Apriori was developed by Anwar. Used for shopping baskets, it says in medical terms looking at situation S these health attributes A1, A2, A3 may have some relationship given how usually appear together. This research proposes a time framed integration of context data and primary diabetic population. The domain of interest is medical data analysis where both lifestyle and biological factors are in play.

In the past decade integration of environmental factors to primary subject data has increased usability of data mining results. The world wide web, millions of medical scientists working on the population, environment and diseases producing large volumes of multiplicity data. Primary data has focused on key medical measures and what they may imply on patient illness. Context or situation changes were looked at around expert knowledge on a given population which is still a subset of the entire world. In the experiment

an attempt is made to address challenges from widowing, accuracy, data mining algorithm selection and handling of noise in both primary and context data sets.

Transformation of numeric values to categories

Indexing

```
diabetesDF = diabetesDF.set_index('Key') # Apply set_index
function
```

To allow for categorization Python needed an index, so a column of row identifiers called 'Key' was added to the data. The key has no patient identifiable data to be inline confidentiality.

Categorization (WHO Context)

The source dataset is in numeric form, but most medical information is given in description format. Medical professionals even advise patients based on the categories. Change numeric values to categories based on medical world labels. The categorization effort also has context concepts as a central driver. The code below transforms Age and Glucose numeric data into medical categories.

```
AgeDF = pd.cut(diabetesDF.Age,bins = [2,9,19,29,39,59,70,80],
labels =
['Todder','Child','Earlyadult','Adult','Middleage','Earlyelderly
','Elderly'])

GlucoseDF = pd.cut(diabetesDF.Glucose,bins = [70,100,190,230],
labels = ['PrePrandial','PostPrandial','PreBed'])
```

```
       Glucose        BloodPressure         Age
Key
1    PostPrandial      Prehypertension   Middleage
3    PostPrandial  NormalBloodPressure       Adult
5    PostPrandial  NormalBloodPressure       Adult
```

Transform numeric values, or ranges into categories, medical terms. These tell a story of the level of health attributes. The amount of subject expertise used at this stage affects the relevance of output from the machine learning process. Final knowledge and insights are dependent on background information used in processing data.

Categorization transformed attributes into labels based on value filling within a given range. This is a key part of the Context since the WHO categorization differed from the Pima Indian one. The source data is the same, but categorization varies. Note from the Data Mining point Context was applied to the Data Preparation stage or Data Transformation Stage refer to the proposal section.

BMI

This is the key categorization based on WHO BMI measurements breakdown with ranges of BMI. As said before this is a key context feature. We take WHO to be one expert view generalized for the world population.

```
BMIDF = pd.cut(diabetesDF.BMI,bins = [17,22,23.9,24.9,100], labels =
['HealthBMI','RiskBMI','OverWeightBMI','SeverelyObeseBMI'])
```

KEY GLUCOSE BLOODPRESSURE ... INSULIN BMI

5	PostPrandial	NormalBloodPressure	...	MediumInsulin	SeverelyObeseBMI
7	PrePrandial	NormalBloodPressure	...	MediumInsulin	SeverelyObeseBMI
9	PreBed	NormalBloodPressure	...	HighInsulin	SeverelyObeseBMI
20	PostPrandial	NormalBloodPressure	...	MediumInsulin	SeverelyObeseBMI
25	PostPrandial	HighBloodPressure	...	MediumInsulin	SeverelyObeseBMI
26	PostPrandial	NormalBloodPressure	...	MediumInsulin	SeverelyObeseBMI
32	PostPrandial	Prehypertension	...	HighInsulin	SeverelyObeseBMI

Categorised data without null valued cells.

Transaction Encoder

```
te = TransactionEncoder ()

te_ary = te.fit(biometrics_list). transform(biometrics_list)
```

Above is an effort to create an array from a list by using TransactionEncoder, fit and transform functions. TransactionEncoder prepares data to be used frequently for data itemset mining. The tool from preprocessing libraries. It changes data into a numPy array from a Python list of lists.

	ADULT	EARLYADULT	...	PREHYPERTENSION	SEVERELYOBESEBMI
0	True	False	...	False	True
1	False	True	...	False	True
2	False	False	...	False	True
3	False	False	...	False	True
4	False	False	...	True	True

Note once transformed the data is now in ‘false’ and ‘true’ values with each category standing for what could be called products in a shopping basket analogue. So, for each patient, (transaction) the system checks if one is ‘SeverlyObeseMBI’ or not.

Fit method stores the unique labels in the dataset.

Transform creates one-hot encoded array, which are Boolean which are memory efficient.

Column Attribute

The unique column details corresponding array values can be addressed using the column logic.

```
df = pd.DataFrame(te_ary, columns=te.columns_)
```

Apriori

Being the main algorithm to put a data frame df into itemsets. Itemsets in this case represent a state of health metrics that occur together for a given health context. Note overall context is whether analysis is being done based on WHO categories, then inner context being whether the sub-population is diabetic or not.

The building of items requires specification of computation metrics in this example min_support, giving control to the level of relevance of outcomes from the function. A range of metrics can now be extracted from the Association Rules. The thesis will just give a few as focus is more on Context integration than general model building.

	SUPPORT	ITEMSETS
0	0.5000	(HIGHINSULIN)
1	0.7500	(HIGHSKINTHICKNESS)
2	0.5000	(MEDIUMINSULIN)
3	0.5625	(MIDDLEAGE)
4	0.5625	(NORMALBLOODPRESSURE)

Support is the number of patients in which a specific range of health attributes occurred. As we progress a health attribute becomes a set of attributes occurring together, itemset. So, members of an itemset start from one to many, also called set length.

Due to result interpretation challenges the analyst needs to decide on a practical minimum value known as support threshold. Any processing after this step will prune those not within the limit.

```
frequent_itemsets1=apriori          (df,          min_support=0.5,
use_colnames=True)

print(frequent_itemsets1)
```

==== Basic output from the Association workout ===

```
0.7500          (SeverelyObeseBMI, HighSkinThickness)
```

We could say with 75% minimum confidence that for patients with diabetes.

Those with HighSkinThickness are also 75% likely to be SeverelyObese based.

on WHO scaling (WHO Context)

```
=====end ========================================
```

There are many terms that are key to implementation of apriori analysis. Filter products that are below a set frequency are known as non-frequency itemsets. On further processing the system removes itemsets with non-frequent itemsets.

Each machine learning algorithm class has varying performance metrics. The project lists these metrics for each class and expands the situation they used. Will discuss the logic behind most of these metrics. The research plans to give more details on performance measuring steps for one of the selected classes that will have been selected for the sample dataset.

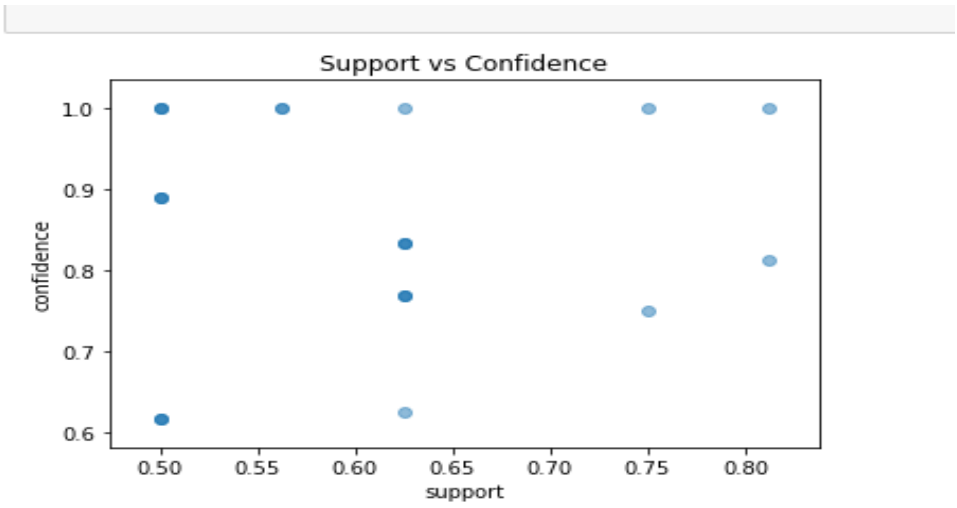
For the selected dataset algorithms performance comparison will be performed and any challenges discussed. The research will illustrate the proposal by using the created script on a dataset. Algorithm performance will also be performed for the selected dataset.

	ANTECEDENTS	...	CONVICTION
0	(HighInsulin)	...	inf
1	(HighSkinThickness)	...	1.125000
2	(PostPrandial)	...	1.083333
3	(HighSkinThickness)	...	inf
4	(SeverelyObeseBMI)	...	1.000000
5	(MediumInsulin)	...	inf
6	(Middleage)	...	1.687500
7	(PostPrandial)	...	1.137500
8	(Middleage)	...	inf

Table 5: Antecedents and Conviction for Indian PIMA dataset

Each medical feature is further analysed to get their Conviction values as shown in table 5.

The medical diabetes relationships or cooccurrence are measured by metrics known as Support and Confidence.



4

Figure 25: Support metric relation with Confidence.

There fig. 25 is some relationship between these two metrics with interesting relations around support of 0.63 link to four confidence levels.

Support and Lift

Fig below shows changes in Lift with increase in Support. Most of Support values link to a Lift of one. Further understanding of PIMA Indian diabetic measures comes from understanding a disease and interaction of its health indicators.

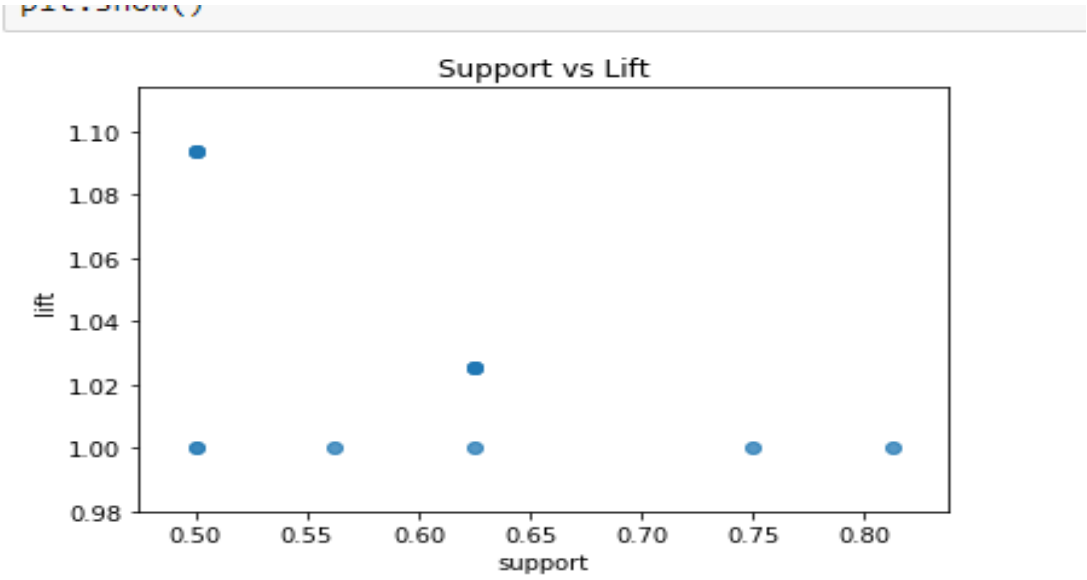


Figure 26: Diabetes Support and Lift

5.2.2 Indian Traffic Accidents Dataset (Cluster Analysis) USE CASE TWO

World over life changing events need to be analysed to find if behaviours can be modified to reduce death. The scenario of the Indian road jungle provides a complex mix of interacting factors which is a rich patterns base. Thesis is aimed at finding ways to analyse this data through a simple defined context approach.

5.2.2.1 Contextual Information integration decisions

The Indian Traffic Accidents data is key to country policies, individual life decisions and affects citizens quality of life. Data analysis needs to be well informed on the culture and

situation in India. Thesis aims to investigate design of data mining solutions that gives equal importance to local dataset and contextual factors. The solution is likely to be both focused and easier to interpret in relation to real-life situations. Results can be explained with reference to the environment, giving an easy to interpret picture of reality. It is known that most local data behave differently with changing environments. This variation in subject behaviour needs to be researched academically.

Context-Awareness implementation starts from understanding the data mining issue and surroundings of the dataset. Transformation of the recognized environmental factors into forms serviceable by a data model is necessary. This can be simply putting together ranges of values like months in the context of seasons. Further decisions will be on the data mining stage to implement context related approaches. Ways of integrating context factors to the design of the model may vary.

Working with stakeholders through either talking to them or referencing their prepared manuals, analyst link knowledge to solution design. In the case of the Indian Accident dataset government officials will be interested in reducing volumes of accidents. Being informed about an environment includes exploring existing articles covering Data mining, decision-making process, knowledge extraction and Indian background. Special focus was made on the key data model building tasks, namely method and algorithm selection. Analyst research should be informed by past work on DM processes, investigation of challenges in DM, background of techniques design and principles to measure algorithm performance.

The Indian Traffic dataset is complex, if we add requirements of clustering approach developers may force to develop innovative approaches. Context-aware effort can increase analysis complexity, (Fung, Jianxin, Xueguan, Milan, & Zhaoquan, 2023) devised (CA-CGL) Context-Aware contrastive graph learning, which was used to class relationships and boundaries of clusters by mixing IGVs (Influential Graph Views) and (TGV) Topological Graph Views approaches.

Data Exploratory Analysis

The traffic data has features covering lots of contexts as summarised in below fig.27 The first step should be to understand the characteristics of the data to be investigated. Data exploration can cover Data Types, Box analysis, Data description and correlations.

Column, rows, and non-null entries

#	Column	Non-Null Count	Dtype
0	Time	12316 non-null	object
1	Day_of_week	12316 non-null	object
2	Age_band_of_driver	12316 non-null	object
3	Sex_of_driver	12316 non-null	object
4	Educational_level	11575 non-null	object
5	Vehicle_driver_relation	11737 non-null	object
6	Driving_experience	11487 non-null	object
7	Type_of_vehicle	11366 non-null	object
8	Owner_of_vehicle	11834 non-null	object
9	Service_year_of_vehicle	8388 non-null	object
10	Defect_of_vehicle	7889 non-null	object
11	Area_accident_occured	12077 non-null	object
12	Lanes_or_Medians	11931 non-null	object
13	Road_allignment	12174 non-null	object
14	Types_of_Junction	11429 non-null	object
15	Road_surface_type	12144 non-null	object
16	Road_surface_conditions	12316 non-null	object
17	Light_conditions	12316 non-null	object
18	Weather_conditions	12316 non-null	object
19	Type_of_collision	12161 non-null	object
20	Number_of_vehicles_involved	12316 non-null	int64
21	Number_of_casualties	12316 non-null	int64
22	Vehicle_movement	12008 non-null	object
23	Casualty_class	12316 non-null	object
24	Sex_of_casualty	12316 non-null	object
25	Age_band_of_casualty	12316 non-null	object
26	Casualty_severity	12316 non-null	object
27	Work_of_casuality	9118 non-null	object
28	Fitness_of_casuality	9681 non-null	object
29	Pedestrian_movement	12316 non-null	object
30	Cause_of_accident	12316 non-null	object
31	Accident_severity	12316 non-null	object

dtypes: int64(2), object (30

memory usage: 3.0+ MB

Figure 27: Traffic Data null values list.

There 12316 rows of data, thirty-two attributes in the datasets. The above also confirms that null entries are very low implying that any column can be confidently analysed. Depending on the nature of data analysis sometimes rows with null values are removed or holder values are inserted. Holder values are usually feature mean values.

Data types

Data manipulation logic need to be supplied with input of certain types. The traffic data was explored for information on data types, results are shown below.

Feature name	Data type
Time	object
Day_of_week	object
Age_band_of_driver	object
Sex_of_driver	object
Educational_level	object
Vehicle_driver_relation	object
Driving_experience	object
Type_of_vehicle	object
Owner_of_vehicle	object
Service_year_of_vehicle	object
Defect_of_vehicle	object
Area_accident_occured	object
Lanes_or_Medians	object
Road_allignment	object
Types_of_Junction	object
Road_surface_type	object
Road_surface_conditions	object
Light_conditions	object
Weather_conditions	object
Type_of_collision	object
Number_of_vehicles_involved	int64
Number_of_casualties	int64
Vehicle_movement	object
Casualty_class	object
Sex_of_casualty	object
Age_band_of_casualty	object
Casualty_severity	object
Work_of_casuality	object
Fitness_of_casuality	object
Pedestrian_movement	object
Cause_of_accident	object
Accident_severity	object

Table 6:Traffic feature list and data types

Our exploration Table 6 produced two types: Object and int64.

Only two (`Number_of_vehicles_involved and Number_of_casualities`) out of thirty-two columns are of data type integer, pointing to need for data transformation if there is need for some types of data analysis.

Box analysis of Casualties

Outliers degrade performance of some data mining algorithms, before using data once need to know the nature of value distribution for each feature. Below is on feature distributions analysis.

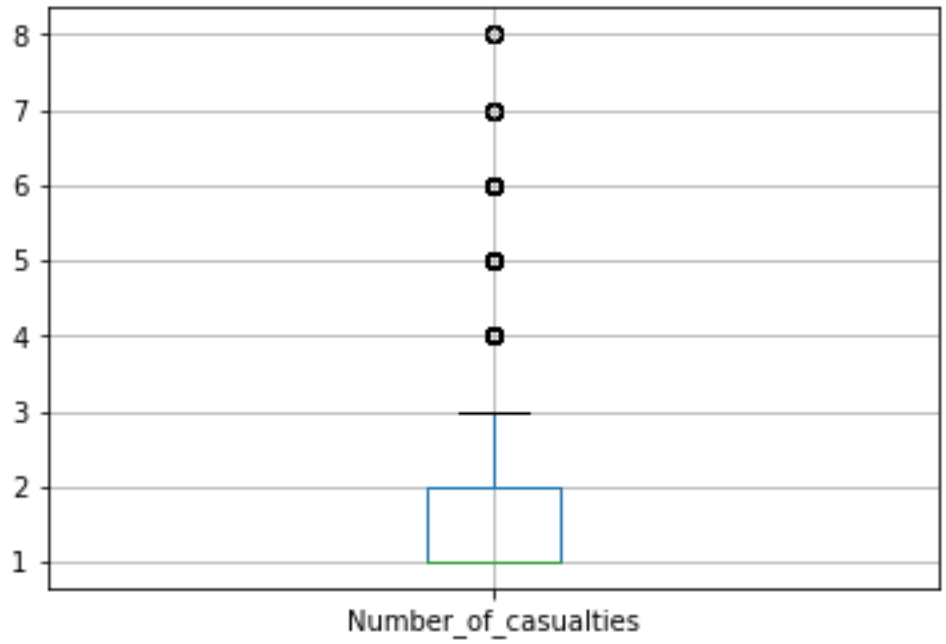


Figure 28: Casualties in an accident Boxplot.

Boxplot for the number of casualties attribute confirms that usually two persons are involved in accidents.

Data Statistical description

Basic information on features like counts, mean, maximum and quartile breakdowns assist in the understanding of data. A gender maximum of two does not give much meaning in real life. Most values supplied in the fig. 29 below make sense.

Out[23]:

	Sex_of_driver	Educational_level	Type_of_vehicle	Area_accident_occured	Road_surface_conditions	Light_conditions	We
count	12316.000000	12316.000000	12316.000000	12316.000000	12316.000000	12316.000000	
mean	0.085823	1.316580	3.845486	2.816174	0.247645	0.307730	
std	0.327676	1.093095	4.243970	2.676400	0.445730	0.513536	
min	0.000000	-1.000000	-1.000000	-1.000000	0.000000	0.000000	
25%	0.000000	1.000000	0.000000	1.000000	0.000000	0.000000	
50%	0.000000	1.000000	2.000000	3.000000	0.000000	0.000000	
75%	0.000000	2.000000	8.000000	4.000000	0.000000	1.000000	
max	2.000000	6.000000	16.000000	13.000000	3.000000	3.000000	

Figure 29: Traffic data exploration figures.

Data description

Data objects were transformed from categories to numeric form, above there a statistic values like count, maximum, mean, and standard deviation. One may need to refer to the categories and real-life context of the attribute for example what does the Maximum 2 for ‘Sex-Of-Driver mean if Male is one and Female is zero. Human data interpretation is key to Data Mining, else a formula just outputs standard results.

Correlations

Traffic features have relationships, correlations can be used to trim the dimensions. A feature is completely related to itself with value 1. Decisions might be made to drop one of any features that are closely related. Below is a Correlations report for the Indian Traffic Dataset.

	Sex_of_driver	Educational_level	Type_of_vehicle	Area_accident_occured	Road_surface_conditions	Li
Sex_of_driver	1.000000	-0.005129	0.007902	-0.011638	0.005137	
Educational_level	-0.005129	1.000000	0.014116	0.018395	-0.017428	
Type_of_vehicle	0.007902	0.014116	1.000000	-0.008585	0.014091	
Area_accident_occured	-0.011638	0.018395	-0.008585	1.000000	0.006648	
Road_surface_conditions	0.005137	-0.017428	0.014091	0.006648	1.000000	
Light_conditions	0.040887	-0.014732	0.010828	0.001223	0.187813	
Weather_conditions	0.020269	0.003579	0.012212	-0.014892	0.290776	
Type_of_collision	-0.005743	0.002046	-0.000044	0.007248	0.010541	
Vehicle_movement	-0.015665	0.003916	0.036073	-0.005078	-0.017172	
Pedestrian_movement	0.005261	-0.008668	-0.012467	-0.005872	-0.014276	
Cause_of_accident	-0.002256	0.003094	0.005219	-0.001370	0.013343	
Accident_severity	-0.013859	-0.005641	0.002881	-0.014009	-0.009377	

Figure 30: Indian Traffic Correlations values.

Correlations of features are shown above in fig 30. There are correlations that are meaningless in real life for example 0.040887 between Sex_of_driver and light conditions. Whereas the -0.009377 might have meaning since Accident severity could easily relate to Road_surface_conditions. These facts increase the need to detail the Context in which the dataset is found.

Correlation Visual

The data above can be mapped into a graph showing feature relationships.

Correlations heatmap

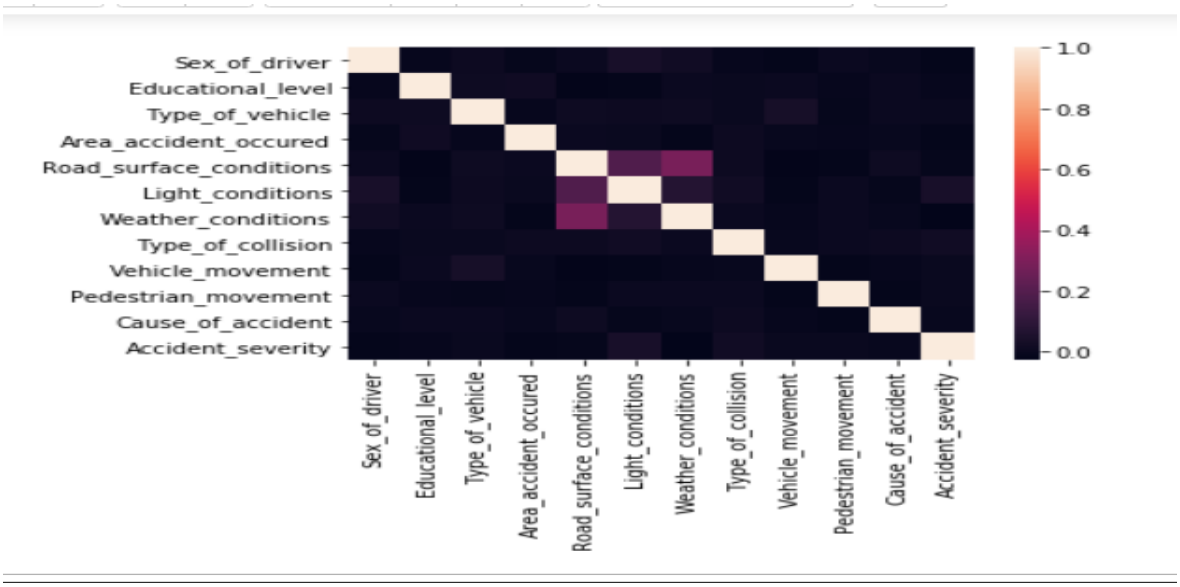


Figure 31: Traffic Correlation Heatmap

The legend on the right indicates values and colour expressions fig. 31 allows for faster decision making.

5.2.2.2 Context-Aware concepts application points

Data science, data analysis or data mining has highlighted challenges with human understanding of the world. Philosophy research has created mathematical approaches to translate relationships between objects and attributes to define Context. (Vinh, Anti, & Siricharoen, 2017) promoted Formal Concept Analysis as an approach to use domain perspectives (context) to a given dataset.

Road Data and situation analysis

Step 1: List Contexts.

- Time Context
- Driver Context
- Vehicle Context
- Road Context
- Environmental Context
- Accident Movement Context
- Causality Context
- Pedestrian Context
- Accident Cause Context
- Severity Context

Software has used to process all sorts of data including varying forms of decision making. Translation of challenges to machine workable and focusing analysis to the correct situation

There now exists lots of source of data including GPS devices, adding contextual facts enhance any analysis (Bhadane & Shah, Context-Aware next location prediction using Data Mining and Metaheuristics, 2020) considered contextual view being time spent on a location, location name and activities associated with the place, they had to design a new model CANLoc to put all factors into the Data mining solution.

. Design of a model comprised of creation of processes, components and decisions being put together to create an adaptable solution. Context-Aware Data Mining follows most of the traditional steps, but with increased emphasis on environment in which the dataset is found.

Step 2: Link Contexts to attributes

There are ten contexts and thirty-two attributes in the Road dataset. Context is of interest to various stakeholders, so analysts need to define them. Models could be developed for a limited number of contexts.

Context	Road Dataset Attributes
Time Context	Time, Day of the week
Driver Context	Age_band, Sex of driver, Education level, Vehicle_driver_relation, Driving _experience, owner of vehicle
Vehicle Context	Type of vehicle, service_year_of_vehicle, Defect of Vehicle
Road Context	Area_accident_occured, Lanes_of_medians, Road_allignment, Types of Junctions, Road_Surface_type, Road_surface_conditions
Environmental Context	Light_Conditions, Weather Conditions
Accident Movement Context	Type of Collusion, Number of vehicles involved, Number of causalities, Vehicle_movement
Casualty Context	Casuality_Class, Sex_of_Casuality, Age_band_Casualty, Casuaity_Serverity, Work_of_Casuality, Fitness_of_Casuality
Pedestrian Context	Pedstrain_Movement
Accident Cause Context	Cause_of_Accident
Severity Context	Accident_Servirity

Table 7 : Context to attribute relations

Each of the contexts may become a focus of Domain Experts for example government officials may be interested in Road Context to make policies to reduce accidents, Social Advisors might be interested in creating flyers on need to avoid driving in a set of Environmental Contexts.

Step 3: Binary Context/ Attribute Matrices

The next step of bridging Context and Data is to transform attributes into binary Yes/No level. Below we focus on Environmental Context. Each value of the attribute either happens Yes or does not No. Expert knowledge then give a situation term to these value combinations.

Features and values need to be mapped into binary level through a transformation process. The below table is a result of this transformation effort.

Binary Context / Attributes Matrices table

'	Light Context	'	'	'	Weather Context	'	Cause context	Severity context
Situation Description	Daylight	Darkness no li	Rainy	Windy	Snow	Fog & Misty	Cause of accident	Accident Severity
Daylight and windy	1	0	0	1	0	0	1	0
Darkness and Snow	0	1	0	0	1	0	0	1
Darkness and Rainy	0	1	1	0	0	0	1	1
Daylight and Rainy	1	0	0	1	0	0	1	0
Daylight and Snow	0	1	0	0	1	0	1	0
Daylight and Fog/Misty	1	0	0	0	0	1	1	1
Darkness and windy	0	1	0	1	0	0	1	0
Darkness and Fog/Misty	0	1	0	0	0	1	1	1
Darkness	0	1	0	0	0	0	1	1
Daylight	1	0	0	0	0	0	0	1

Table 8 : Binary matrices of the data

Binary transformation on table 6 is for the Indian Traffic dataset. Above looking subset of Light and Weather conditions Context created a binary matrix linking these condition as they relate to Accident cause and severity.

Data Context Relationship Maps

The physical world of cars, people and roads interacting need to be shown in diagrams for better communications. This mind or thought visualisation has been borrowed from other areas of study like psychology. Below is an Indian Traffic Accident Severity data lattice.

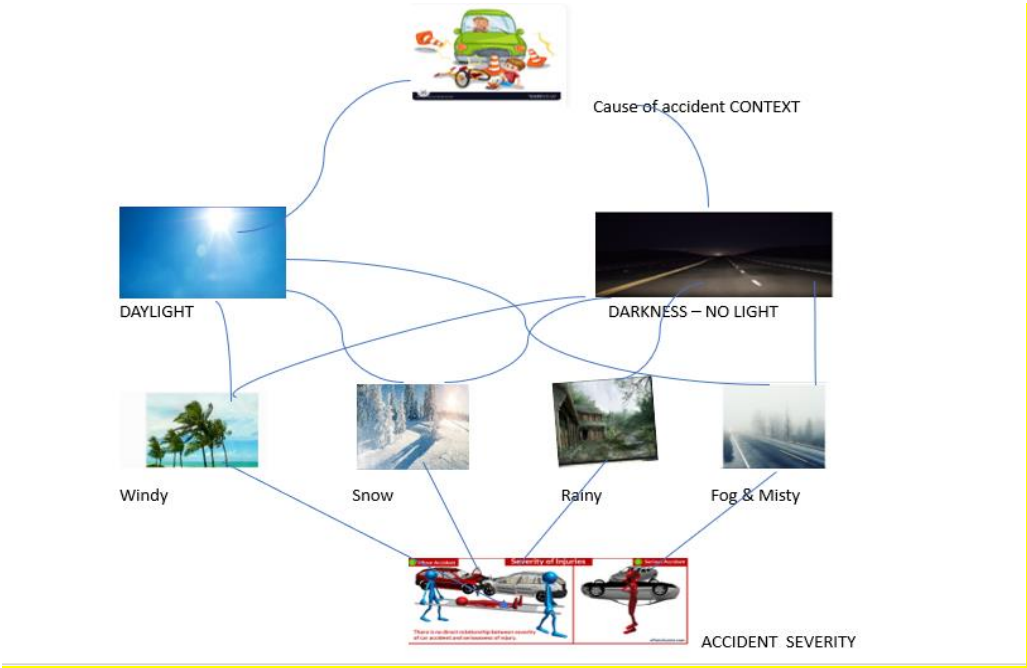


Figure 32: Indian Accident Relationship Lattice

The fig. 32 can answer investigation on Causes of an accident being darkness and foggy & misty weather conditions. A second perspective may look at Accident Severity blaming it on a Rainy and Daylight.

DATA SOURCE

DATA OUTLINE

System uses accumulated data mining methods knowledge. Algorithm selection is a two staged process starting with class (i.e., Classification, Association, Reinforcement,

Regression, Clustering) selection ending with selection of suitable algorithm. Each machine algorithm class has numerous algorithms grouped by the way they process data. Selecting the algorithm is out of the scope of this thesis.

EXPERT KNOWLEDGE

(Ditcharoen, et al., 2018) in a study of accident severity was of the view that similar research from different countries would always produce varying results, they went on to identify four severity factors as vehicle conditions, human, environmental and road infrastructure.

Over the years several techniques and many algorithms have been developed to improve knowledge extraction. Though this growth has led to specialised solutions, the large array of options has also resulted in algorithm selection challenges during model construction. Many considerations are undertaken including designing and purpose of algorithms. This project should improve this phase of model building, by investigating all key factors. The research aims to resolve the challenge of selecting the appropriate techniques to be used on a given data mining problem. The thinking is around teaching software to process existing knowledge and make some decisions.

It could be more informative to analyse a dataset with high a level of situational information. Context also enriches resulting insights. India as context has 71% of vehicles fifty-two million being MTW Motorised two Wheelers for example scooters, motorcycles, and mopeds. Looking at causes and influencing factors of TRI Road Traffic Injuries (Gururaj, 2008) cited both vehicle type and area population density.

5.2.2.3 Algorithm Class Selection for Context-Aware Data mining solution

Stage 4 Process Data Modelling adaption

CMAS Context Framework Method and Algorithm selection phase is through a script for this thesis, but same tool can be developed in other languages.

Method Selection

CMAS encourages an integrated approach to data mining, research and discussions with subject experts must be done before embarking into data analysis. Looking at the India Traffic dataset, decisions on the design of the analysis model must consider the situation in India. Knowledge of other issues that affect traffic related activities. Domain knowledge on every attribute needs to be considered.

The thesis is saying the environment affects behaviour so we either incorporate the effects or acknowledge them. Traditional data mining ignores effects of the environment; this is reflected in insight being far from real-world characteristics and difficult to interpret. Data mining frameworks can be designed to produce insights that can be actionable with minimum effort.

Decision on appropriate Data Mining technique to apply on the Traffic data had to depend on their characteristics. The diagram below attempts to highlight method characteristics.

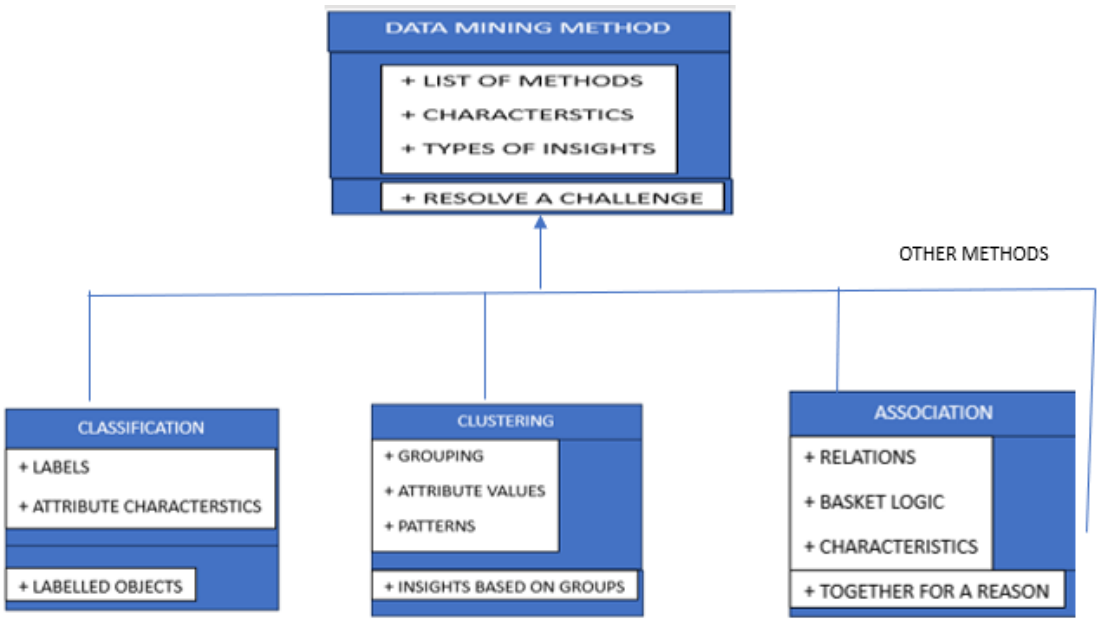


Figure 33 : Relations of Data Mining Methods

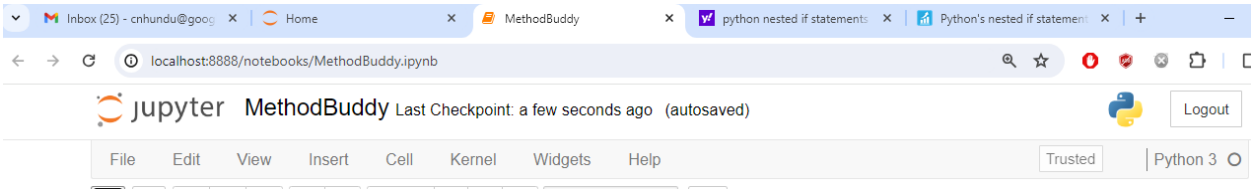
The Relations of data mining methods fig. 33 shows only Classification, Clustering and Association methods. The method selected for this use case is Clustering as detailed later.

Data analysis is resource intensive so the solution should be designed to apply effort on a focused scope. There are many challenges faced in Data Mining including overload of available methods and Algorithms sparsity and overwhelming volumes of insights. The proposal increases focus thereby reducing some of the challenges.

The CADM framework uses Question Answering System (QAS) technique, (Gupta & Gupta, 2012) an expert system that gets questions, comprehends them, looks up expert opinion from an Ontology of DM methods, and has a user interface to feedback the findings.

Method Selection script outline

Working setting for solution development and model running.



Work is done on Python 3 using the Jupiter Notebook development environment. The script is called Method Buddy as an assistant to select appropriate Data Mining methods.

Input

```
In [1]: ## Data Mining Method Buddy

In [56]: ### Suggestions are made by the DATA METHOD BUDDY depending on data entered into the screen ###

In [57]: ##  Buddy accepts options from keyboard. Interaction with the Analyst

In [51]: Dlabelling = str(input("Does the data contain information to label elements YES NO:"))
         Does the data contain information to label elements YES NO:NO

In [14]: Rnumeric = str(input("Are the results from analysis in numeric form YES/ NO:"))
         Are the results from analysis in numeric form YES/ NO:NO

In [16]: DSegmant = str(input("Does insight come from splitting data based on some features: YES/ No: "))
```

Logic of responses to key questions posed to the Method Buddy.

Driving Logic	Classification	Regression	Association	Clustering	Other
Label rows	YES	YES	NO	NO	NO
Numeric results	YES	YES			
Categorize objects	YES	NO		YES	
Segment objects			NO	YES	NO
Occurrence			YES		NO

Figure 34: Traffic Data Mining Method Selection logic.

Concept logic building starts by high level characterisation of basic data mining methods. The approach can then be expanded to handle many conditions and method classes. Variables created to accept options as provided by the Analyst based on basic questions about the challenge and local dataset. Note the questions asked are what a Data Analyst would ask themselves is manually modelling the data.

Display Response

```
In [52]: print( "labelling:", Dlabelling)
         labelling: NO

In [15]: print( "Numeric Results: ", Rnumeric)
         Numeric Results:  NO

In [54]: print( "Segmentation : ", DSegmant)

         Segmentation :  YES

In [19]: print( "Categorization: ", Tcategory)
         Categorization:  NO
```

These responses will be used later in making decisions. Above we provide a way to see what the Analyst has answered to key questions. The responses also work to transfer the Analyst understanding of data to the machine Method Selection Buddy.

Data mining solution may be a hybrid or ensemble of algorithms, with some data preparation activities like normalisation. Any opportunity for tuning may be considered ranging from AdaBoost, Gradient Boosting, Random Forest, and extra tree. The research highly valued data mining skills and domain knowledge contribution to the success of data analysis for example in the medical domain. Context-awareness improves numeric values to relate to the subject of analyses.

Data mining issues and available enormous amounts of data is making analysis more complex. Decisions on data mining methods to apply are also becoming difficult. CMAS attempts to address these issues through teaching software to apply historical expertise knowledge. Investigations need to be wide accepting that various methods can be used at various stages of a single analysis project. The thesis is not exhaustive, starting academic discussions in this direction will be an acceptable level of success.

Results

```
In [58]: ## options used in this case not labelling. They are interested in segmantation of data

In [59]: if Dlabelling == ("NO"):
             if DSegmant == ("YES"):print("Data Mining Method suggest your use Clustering")

         Data Mining Method suggest your use Clustering
```

Above the Method Buddy suggested using the Clustering method to mine insights given the challenge seems not to emphasise labelling subjects between some classes and is suggesting putting them in some segments.

Data mining methodologies aim to structure components for clear steps that optimise results. Design of a data mining framework starts with investigation on classification of Data Mining opportunity. At the highest level, depending on the presence of data labels

the opportunity can either be classified as supervised, semi-supervised or unsupervised. Supervised datasets have labels, but expert effort is required to create this type of data. Semi-supervised methods are two steps with labelling of small samples of dataset and use of these labels on the whole data. Data without labels, known as Unsupervised is the bulk of real-life datasets, hence increased effort on research on processing this data structure.

Decision Making Logic

```
In [46]: if Dlabelling == ("YES"):
             if Rnumeric == ("YES"):
                 if TCategory == ("YES"):print("Data Mining Method suggest your use Classification")

In [47]: if Dlabelling == ("YES"):
             if Rnumeric == ("YES"):
                 if TCategory == ("NO"):print("Data Mining Method suggest your use Regression")

In [48]: if Dlabelling == ("NO"):
             if Dsegmant == ("NO"):
                 if Dcooccour == ("YES"):print("Data Mining Method suggest your use Association")

In [49]: if Dlabelling == ("NO"):
             if Dsegmant == ("NO"):
                 if Dcooccour == ("NO"):print("Not able to prescibe Data Mining Method")
```

Each choice is a combination of decisions or options selected by the analyst. These options cover characteristics of the data and an overview of direction of investigations undertaken. Expert knowledge on abilities of Data Mining Method is a key consideration in construction of options and they are used in the Method Buddy decision making process.

Note the option with result “Not able to prescribe Data Mining Method” point to ensembles or situations where challenge need to be split before applying data mining methods. Like any automated system the human decisions are always required but at reduced levels. It may be of interest as to ways in which Artificial Intelligence will be used in future to replace the Data Mining Method Buddy.

The above effort should produce suggestions that are far better that human intuition. The accumulation of knowledge and speed of retrieval should massively improve this decision.

5.2.2.4 Clustering mining method

Following the CMAS framework, the selected DM needs to be applied to the sourced data. The thesis has left selection of algorithms outside its scope. This was done since other researchers have fully covered algorithm ranking and selection.

Clustering Implementation

Sample Dataset

Analysts and end users need to understand the actual data in the dataset. Below is a sample of the data.

	Sex_of_driver	Educational_level	Type_of_vehicle	Area_accident_occured	Road_surface_conditions	Light_conditions	Weathe
0	Male	Above high school	Automobile	Residential areas	Dry	Daylight	
1	Male	Junior high school	Public (> 45 seats)	Office areas	Dry	Daylight	
2	Male	Junior high school	Lorry (41?100Q)	Recreational areas	Dry	Daylight	
3	Male	Junior high school	Public (> 45 seats)	Office areas	Dry	Darkness - lights lit	
4	Male	Junior high school	NaN	Industrial areas	Dry	Darkness - lights lit	

Figure 35:Traffic data with text-based values.

Traffic data shown in fig 35 show categories for most features.

Datasets have values as per their use, for analysis some transformations may be necessary. The solution may create a Context-Aware Data mining framework that is simple and effective. Data mining has about five phases from data preparation, data wrangling, algorithm selection, data processing to interpretation of results. One of the key

challenges is identifying the most effective point to apply context concepts. Contextual factors can be applied to one or more stages of the Data Mining process.

	Sex_of_driver	Educational_level	Type_of_vehicle	Area_accident_occured	Road_surface_conditions	Light_conditions	Weathe
0	0	0	0	0	0	0	
1	0	1	1	1	0	0	
2	0	1	2	2	0	0	
3	0	1	1	1	0	1	
4	0	1	-1	3	0	1	
5	0	-1	-1	-1	0	0	
6	0	1	0	0	0	0	
7	0	1	0	0	0	0	
8	0	1	2	3	0	0	
9	0	1	0	0	0	0	

Figure 36: Traffic data transformed into numeric for analysis.

The data is transformed into binary as explained in earlier sections. Most drivers in the above sample are female as depicted by the value 0 on the column Sex_of_Driver.

Number of clusters

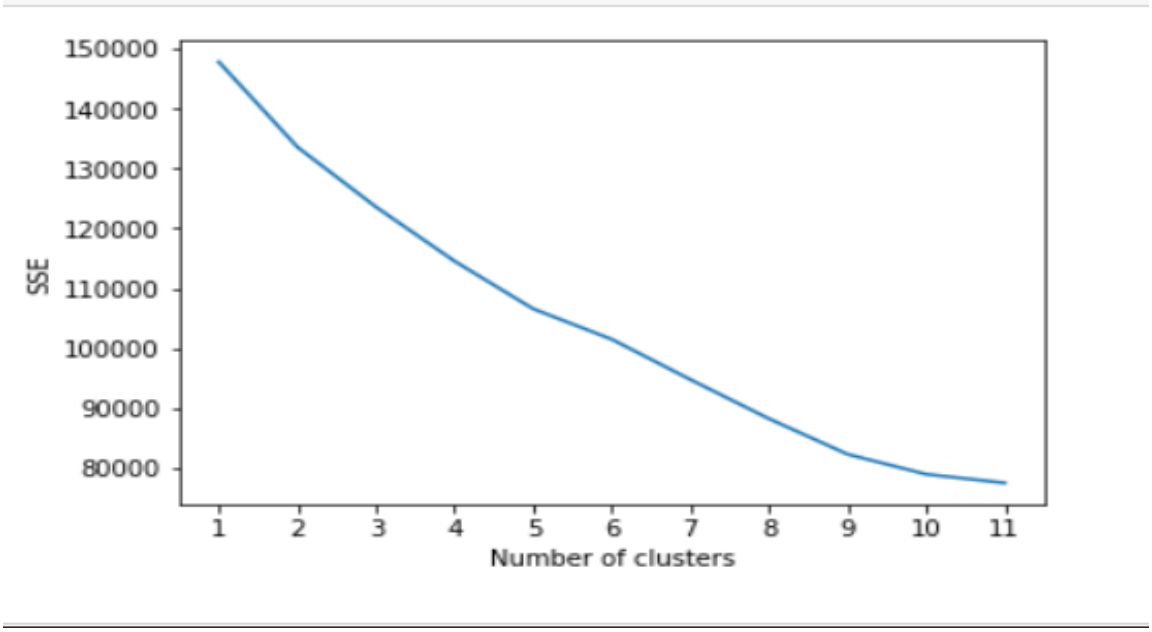


Figure 37: Number of clusters analysis.

Developers need to determine the optimum number of clusters to produce from the dataset. In this case the bent is around 5 clusters shown on fig 37. The clustering process is an iterative trying diverse number of clusters until one finds the best fit.

Analysed data can be linked back to source to explain groups using the original terms. One can combine text to describe clusters.

	Sex_of_driver	Educational_level	Type_of_vehicle	Area_accident_occured	Road_surface_conditions	Light_conditions	Weathe
0	Male	Above high school	Automobile	Residential areas	Dry	Daylight	
1	Male	Junior high school	Public (> 45 seats)	Office areas	Dry	Daylight	
2	Male	Junior high school	Lorry (41?100Q)	Recreational areas	Dry	Daylight	
3	Male	Junior high school	Public (> 45 seats)	Office areas	Dry	Darkness - lights lit	
4	Male	Junior high school	NaN	Industrial areas	Dry	Darkness - lights lit	

Figure 38:With a cluster's column added on can look back on the text values in the dataset.

The plot below shows Cluster which can be interpreted or made into insights.

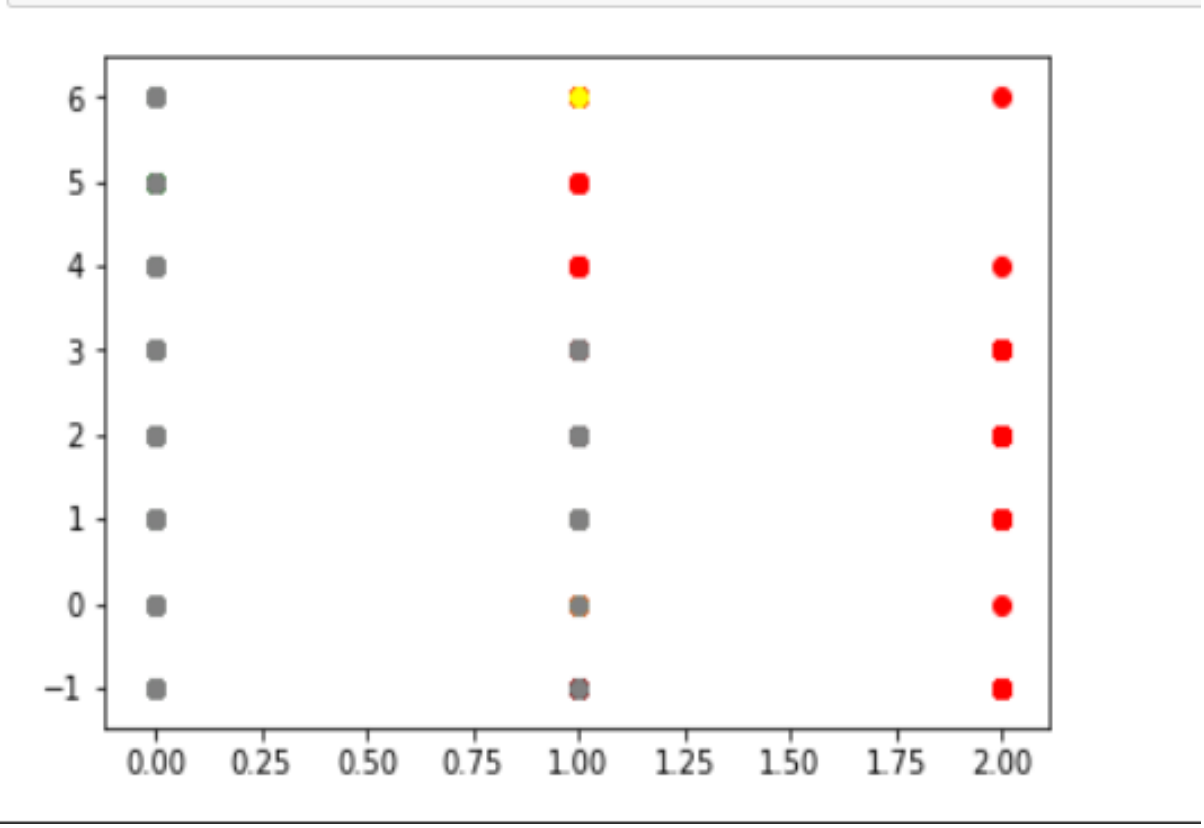
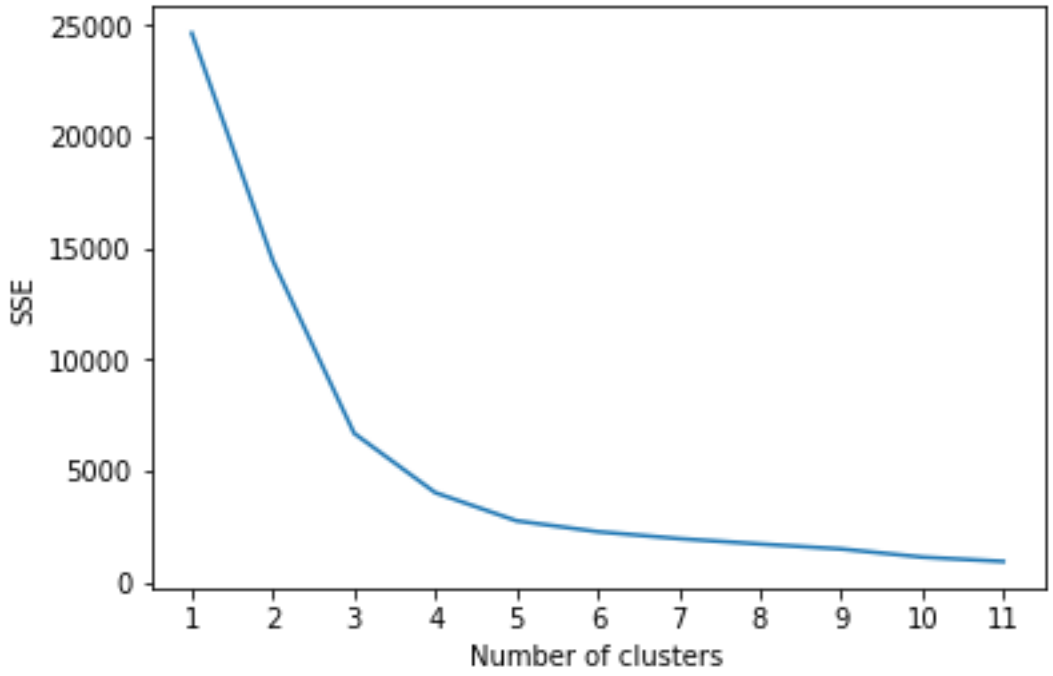


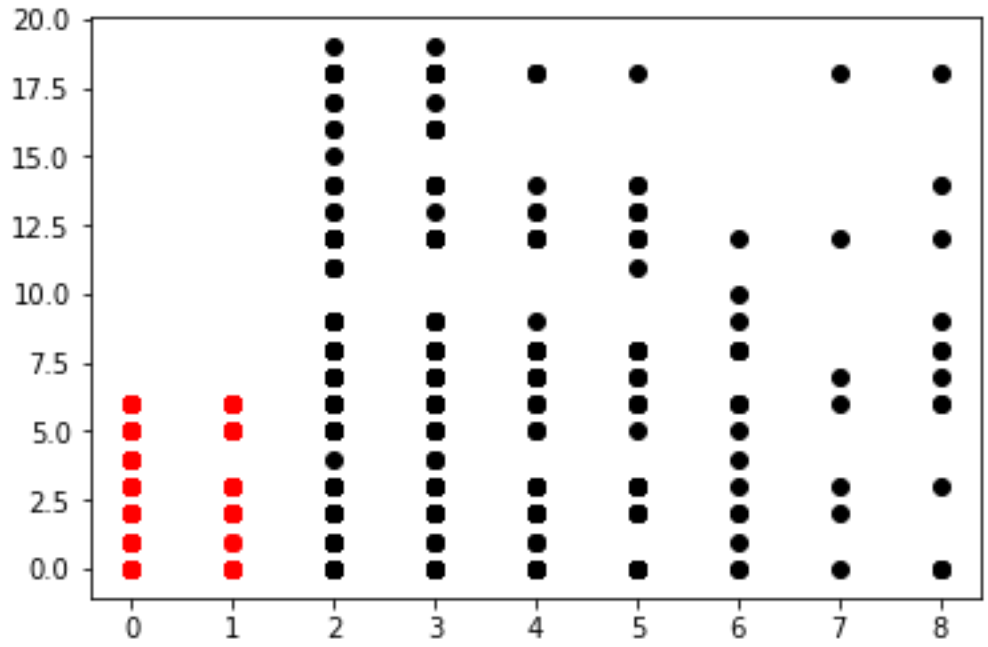
Figure 39: India Traffic Cluster plot

Clusters shown in fig 39 are in assorted colours.

Analysis of Type of Vehicle, Pedestrians movement, cause of accident and Accident Severity using 5 Clusters. This has given the below grouping, change of attributes or number of clusters will change the shape on the graph.

Checking contribution of Pedestrian Movement to Causes of Accident. Analysis done using three clusters based on the elbow graph below.






```
Silhouette score (n=5): 0.5755602237594535

Davies_Bouldin Index (n=5): 0.5824447643543826
```

5.2.2.5 Context-Aware Data Mining Model Development (Prototype)

A very low-level software assistant developed around selection of point to apply context and data mining methods. Note current effort on AutoML (Automatic Machine Learning) concentrates on selection of correct data mining algorithms. This research just discusses the selection of data mining method not algorithms. As mentioned above translation of problem into software understandable form will be a key pillar of effort of this research. The translated input should take a form like that which can be processed by the human brain.

The actual data mining model design is stakeholder and developer centred, there is needs to promote key aspects around data like context-awareness. Otherwise, the steps are too mechanical, reducing the likelihood of varying insight from a given dataset. The context consideration gives the entire process a new perspective and brings insight nearer to the situation at hand. Proposed framework for data mining will be looked at around a brief outline, development history, logic, advantages, and disadvantages of approaches. As interest in Data Mining has gathered momentum tools created exponentially leading to selection challenges. Well defined frameworks will increase uptake from other professions. Data Scientists and analysts are professionals in the main domain, there is needs to encourage others like bankers and clinicians to partake in DM.

CMAS final phases on 'Context-aware model implementation and insight interpretation' are lean and produce environment linked results. This is because the whole cycle involved checking relationships of local data and external factors.

CASE STUDY THREE

5.2.3 Air and Water pollution Dataset (Regression Analysis)

5.2.3.1 Air and Water pollution situation background

The perspective of Context varies depending on data and information provided. The context features in Context-Aware Data Mining can also be a subset of the situation being considered depending on project objectives. It can even be expanded to saying looking at data from a defined perspective. Scoping and situating a data mining challenge simplifies the process and improves the quality of insights. Detailing the situations can enrich the sense of resulting insight.

Traditional Data mining attempts to get insights from a given dataset. Contribution of subject experts' knowledge, changing real life situations and environment should be considered during data mining solution implementation. Thesis is motivated by the possibility of improving quality of results because of incorporating influencing context. Use

case could be analysis of diabetes dataset based on expert disease factor definition. Another application could be where states of subject are defined e.g. Stop in a Vehicle Navigation dataset as vehicle being on a location for longer than a given period.

If readers are interested in pollution and its effect on people's life, data sought should have attributes linking both interests. Quality of data mining activity's output including relevance to audience depends on the methodology followed. Analysts study the issue at hand and datasets characteristics. Research has of late moved towards context-awareness, algorithm performance comparisons and any data pre-preparations strategies. Covering records of context pollution worldwide and death related to it allowed a more comprehensive view.

The features can be temporal or spatial. Human expert knowledge can be involved in the model design (Context-Aware) or left out as in traditional data mining concepts.

The Methodology should have a way to measure improved qualities and extra effort enforced by the framework. An evaluation of insights from the three domains analysed using CMAS may provide a clear picture on the possible contribution of the novel approach. The latest solutions based on context have been researched on, having contextual features driving Recommender Process (Viktoratos, Tsadiras, & Bassiliades, 2018) used Association rules focused on a single user's check-in history. They in the framework combined Recommendation logic and Association rules algorithms from a context perspective.

Once created the solution will speed up model creation, improve quality and encourage other professionals to data mine. Experts' knowledge use in data mining increases the results relevance and effectiveness of the model. Creation of high-level frameworks that can be tailor made for future data analysis projects. Opportunity of sharing situation data, user understanding and data mining expertise in an interactive informative framework.

Looking at data from cities, air pollution is measured using emissions (Castells-Quntana, Dienesh, & Krause, 2021), though high population density turns to lower emissions per Capita.

Water pollution as a negative adjustment of water characteristics, (Khatun, 2017) it is affected by population density and industrialization.

The research is motivated by the possibility of having accumulated knowledge on data mining methods by using computer software to make key model design decisions. The use of knowledge repositories to inform data mining solution design is likely to improve resulting model and analyst performance. There is a possibility to semi-automate some solution design steps.

5.2.3.2 Contextual Information integration decisions

This is further complicated by the fact that Classification algorithms may be measured using different metrics from Association algorithms. Create and illustrate a data mining framework that is designed from accumulated methods knowledge. Illustrate the framework through real world examples to highlight logic and limitations. Algorithm selection is done by measuring performance using agreed metrics.

Data Summary

The case study analyse World Pollution challenges based on seven features covering measures and levels of death linked to the pollution.

Feature	Description
Share of death air pollutic	Percentage of country wide death that is linked to Air Pollution
Share of death Water poll	Percentage of country wide death that is linked to Water Pollution
pm10 concentration	Annual mean concentration of particulate matter with diameter of 10um or less (ug/m3
pm25 concentration	Annual mean concentration of particulate matter with diameter of 2.5 um or less (ug/m3
no2 concentration	Annual mean concentration of nitrogene dioxide (ug/m3
pm10 tempcov	Annual temporal coverage for pm10 on a 100 base (full year = 100; 1=1% of the year
pm25 tempcov	Annual temporal coverage for pm2.5 on a 100 base (full year = 100; 1=1% of the year

Table 9: Air and Water Pollution Features

Table 9 informs on measures and description of feature to be analyzed when investigating the pollution situation in the world. Data mining task for the Air and Water is very situation specific rendering most design activities to remain outside automation. Data was preprocessed with the main subject in mind. Data scientists still need to prepare data from many sources, clean it and make other environment related decisions like considering privacy issues. Breakdown of data mining tasks should be noticeably clear on the contribution of the human and the automated.

Data sources

:ountry_name	who_region	iso3	year	pm10_concentration	pm25_concentration	no2_concentration	pm10_tempcov	pm25_ten
France	3132	3132	3132	2590	1205	2678	2590	
Georgia	4	4	4	4	3	0	3	
Germany	3266	3266	3266	2707	1488	3111	2680	
Ghana	4	4	4	4	0	0	4	
Greece	164	164	164	143	57	115	139	
Guatemala	5	5	5	4	3	4	0	
Honduras	4	4	4	4	1	0	0	
Hungary	193	193	193	185	68	164	182	
Iceland	36	36	36	33	25	35	31	
India	1301	1301	1298	1264	220	270	147	
Indonesia	28	28	28	24	28	22	23	

Figure 40: Pollution data sample.

Water and air pollution are recorded all over the (fig. 40) world but not at the same level.

Air pollution data was collected as an average of measurements over several years. It was noted through counting that some countries have many measure taking tasks.

	Entity	Code	Year	Mortality_rate_water
0	Afghanistan	AFG	2019	16.64
1	Algeria	DZA	2019	4.05
2	Angola	AGO	2019	48.85
3	Antigua and Barbuda	ATG	2019	2.47
4	Argentina	ARG	2019	11.45

Data by country of total death related to water pollution. It is a snap view pollution activity which will be worked together with other average figures. It will be great to share these assumptions to allow interpretation of data to have accurate meaning.

Entity	Year	Share_deaths_pollution
Morocco	2004.5	11.940793
Mozambique	2004.5	12.430966
Myanmar	2004.5	19.607624
Namibia	2004.5	10.769775
Nauru	2004.5	3.925831

Looking at an average death from air pollution, note the year will need to be dropped as it is no longer carrying any meaning.

Data Statistical Summary

	Share_deaths_pollution	Year	Mortality_rate_water	pm10_concentration	pm25_concentration	no2_concentration	pm10_
count	110.000000	106.0	106.000000	110.000000	110.000000	110.000000	11
mean	9.681825	2019.0	9.082264	47.935181	23.942054	16.726205	5
std	5.208602	0.0	12.635827	52.610222	20.365069	19.938705	3
min	1.281300	2019.0	0.420000	0.000000	0.000000	0.000000	
25%	5.702150	2019.0	2.345000	18.104650	11.480675	0.000000	
50%	9.145450	2019.0	3.955000	29.351250	18.614950	15.310600	7
75%	13.554150	2019.0	8.872500	53.887425	30.041900	21.891475	8
max	23.202500	2019.0	71.730000	273.931400	119.774000	125.234300	10

Figure 41:Statistics for the Air and Water Pollution

Statistical analysis informs on the data distribution like Fig 41. This also influences the actual data mining process. Data with air, water pollution death and the air pollution concentrations. The statistics show values ranging from 23 to 273 therefore any meaningful check of variable relationships will need some form of normalizations.

Data details

The initial data had lots of data covering many years, but a decision was made to get averages. Time of occurrence was dropped for the Pollution Data as it was not informative. When analysis insights expressed using both data features and Context changes, they tend to be more informative. Knowledge like subject S *pollution* increase (behaviour I *cause death*) as environmental factor E (*Air Pollution*) decreases (behaviour d). The two-dimension S and E are more informative taken together than any of them on its own. As said before very, few activities happen in real life without the effect of the surroundings. In other experiments one could say we got these results assuming certain factors being constant. Human expert knowledge should be involved in the model design.

Define the model end to end for the reader. Measure contribution of human expertise to an analysis journey

Case Study clarifies the specific area application of Context to problem understanding. Key concepts of Context-Aware data mining are that at every design point is the situation being considered. Software is designed to assist in the decision-making process during model design.

Data and Context perspectives

Context-aware concepts application and method selection concepts are evaluated in three use cases. The proposed CMAS framework, its application to PIMA diabetes, Indian Accident Severity and World pollution datasets.

Data analysis is data at distinct levels of expertise from end user, data science students to experts' Data Scientists. The thesis effort is justified though review merits of the proposed novel Context-aware framework against the classical non-context-aware approaches. Varying angles of looking at the perceived benefits of the novel approach. A cost benefit analysis will also be performed, where cost is the added activities from Context-Aware approach like extended problem analysis and context definition steps.

As per DSR methodology implementation is used to evaluate the proposed direction of expanding academic knowledge. Narrowing down of data given focus. Imagine if the project is looking at hundreds of attributes and millions of rows of data. Context-aware would prune for a given situation and allow for deeper mining of the selected data subsections.

Thesis promotes the use of context aware principals to increase insight's relevance to real-world understanding. It also explores the possibility of software assisting in the selection of Data Mining methods, a key part of model development. Evaluation process aims to measure the level to which the proposed framework meets the mentioned objectives.

Clear communication of the framework to invite discussions on topics ranging from applications of the framework to suggestions on modifications of the proposed approach.

Evaluations will be designed as a two-way sharing of views and knowledge between researcher and subjects' experts.

The Context-Aware Data Mining Framework was demonstrated by applying it to a few uses case. Analysis of the implementations result should conclude on the key benefits of the thesis that is increase insight relevance to real world issues on a dataset, focus, improved performance of data mining tools.

The standard algorithm evaluation metrics like Hausdorf distance, classification accuracy, mean square error are best used when comparing algorithms in isolation of other efforts on the lifecycle. Therefore, the research focused on measuring the extent to which the proposed framework improves the analysis journey not just single stage or activity.

.

5.2.3.3 Context-Aware concepts application points

Expert knowledge

The dataset sits in the medical domain. Reading papers medical practitioners have agreed facts with limited variations. Any data analysis premised in this domain must be guided by this knowledge. Our analysis is based on medically agreed facts on which most of these will be explored. A link with issues discussed in the research will be explored to illustrate the proposed solution.

Develop or define Context.

- Define context C by looking at the D dataset, E expert knowledge and situation change parameters.
- Defines focus of analysis F, more of scope and direction of this research or analysis
- Transform features into context aligned form.
- Outline processing steps. Use of the class feature in understanding or getting varying insights.

- Outline each stage as it relates to the proposal. Also detail any challenges and probable future solutions.

Data Scientists make lots of decisions in the design and implement any data mining solution. Design of a Context-Aware solution starts with key decisions to transform problem description, data understanding and key environment knowledge into a Context-Aware solution. Key to this decision is integration of available context at the most appropriate stages in the data mining solution. In a simplified form this may be a single stage, but in real-life it is several stages.

Data mining technology has facilitated self-learning systems. Consideration has also been taken to make self-learning applications that accumulate performance and characteristics that may be compared to future context. This decision is complicated given variations within processing windows over the lifetime of the dataset. Task of transforming context information into a usable form adds into the key activities of the Context-Aware paradigm.

```
x=Worldpollutionclean[['Mortality_rate_water','pm10_concentration','pm25_concentration']]

y=Worldpollutionclean['Share_deaths_pollution']
```

```
print(model.coef_)
print(model.intercept_)
```

```
[0.06317286 0.02582579 0.11618712]
5.155909285302036 5
```

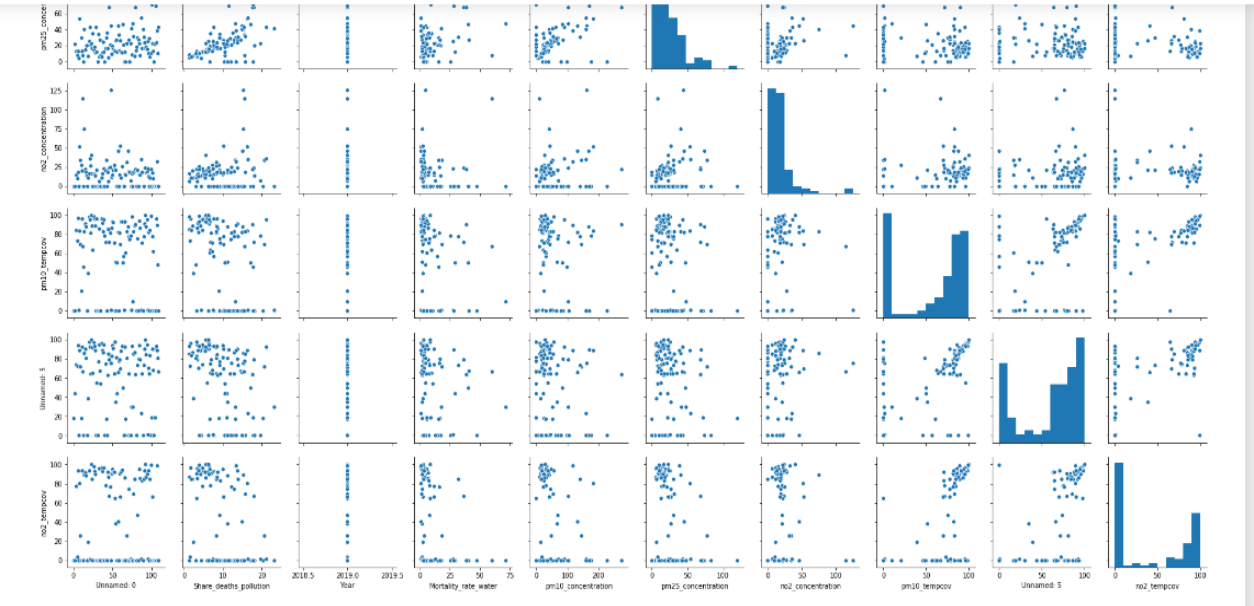


Figure 42: Pollution correlations.

```
plt.hist(y-predictions)
```

```
(array([ 2.,  9., 20., 18., 24., 14., 10.,  4.,  1.,  4.]),
 array([-7.95073846, -6.08020468, -4.20967091, -2.33913713, -0.46860336,
         1.40193042,  3.27246419,  5.14299797,  7.01353174,  8.88406552,
        10.75459929]),
 <a list of 10 Patch objects>)
```

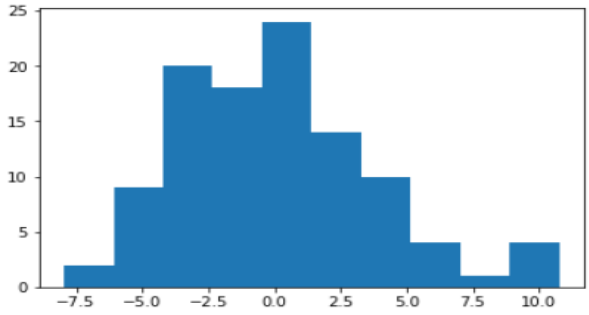


Figure 43: Data fitness for Regression.

5.2.3.4 Algorithm Class Selection for Context-Aware Data mining solution

Advancement in technology has led to Intelligent Discovery Assistants (IDAs), (Serban, Vanschorn, Kietz, & Bernstein, 2013) built around self-learning systems that can get knowledge from huge databases of a given field, in our case DM methodologies.

Modern system developers need to create solution that are not resource hungry. Resources are always limited therefore need to be managed well from a system design perspective. Need for data pruning is increasing the application trends. Data analysis is now not a reserve of high-capacity computers, enabling smart devices application to everyday life. As data mining solutions are being applied to day-to-day activities, there is pushing for processing on portal machines.

Analyst

```
: if Dlabelling == ("YES"):
    if Rnumeric == ("YES"):
        if Tcategory == ("NO"):print("Data Mining Method suggest your use Regression")

Data Mining Method suggest your use Regression
```

Analyst response to Method but questions about the situation at hand got the above results from expert knowledge.

5.2.3.5 Regression Analysis rule mining method

The most logical illustration of the proposed approach is to compare it with the way data mining traditionally performed. There is a need for proof of concept but more importantly attempts to justify the extra effort being introduced by the proposal. The analysis, evaluation and discussion in chapter 5 should clarify expected model improvements.

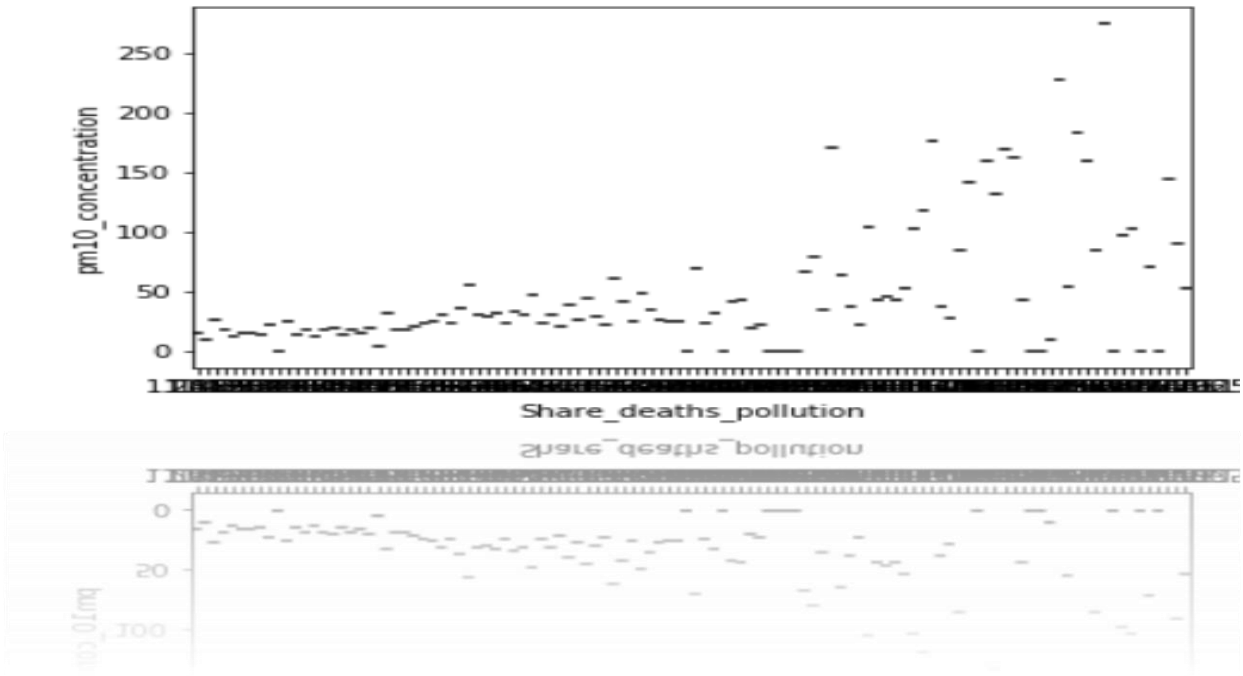


Figure 44: Plot of PM10 air pollution and death rate.

Death linked to air pollution trends to increase as pollution raises fig 44.

Perform (CDM), Classical Data Mining on a sample dataset. This is data analysis without taking into consideration any environmental information. Because of research time limitations the algorithm selected above will be used. CDM does all stages using the basic dataset, data preparation and application of data mining varying from CADM, Context-Aware Data Mining which considers the environmental factors.

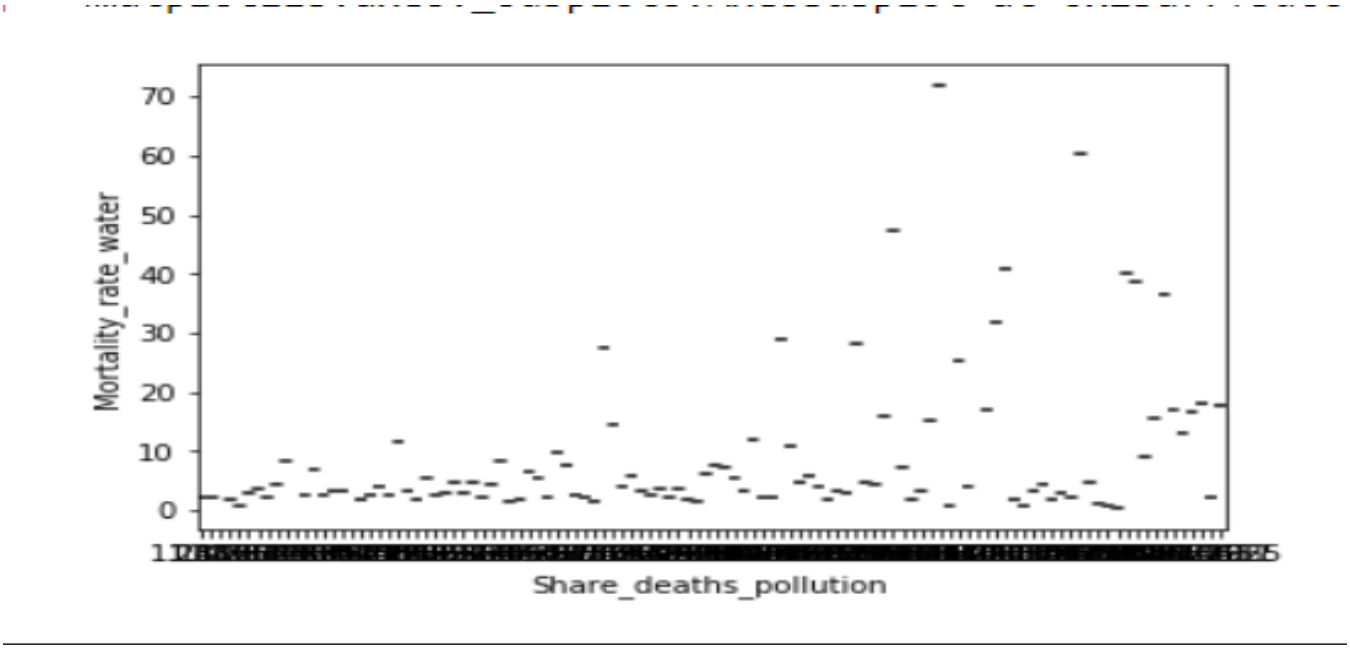


Figure 45: Relationship of the two death rates.

The death rate linked to water pollution is not influenced by air related death. In Context-aware analysis further work can drop the other feature.

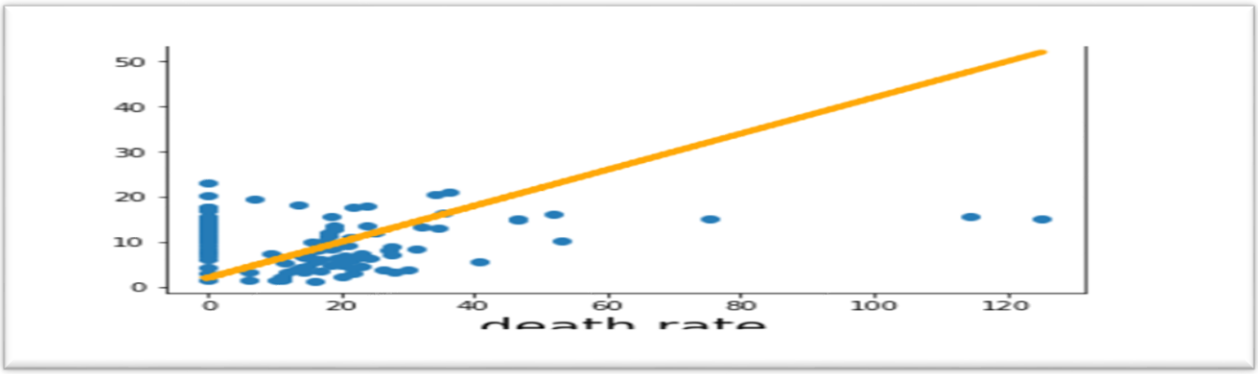


Figure 46: Regression line Air and Water

It will be used for prediction using the above regression relationship. In real life it is essential work through data and make decisions based on knowledge coming from data mining.

5.2.3.6 Context-Aware Data Mining Model Development (Prototype)

For the selected dataset algorithms performance comparison will be performed and any challenges discussed. The research will illustrate the proposal by using the created script on a dataset. Algorithm performance will also be performed for the selected dataset.

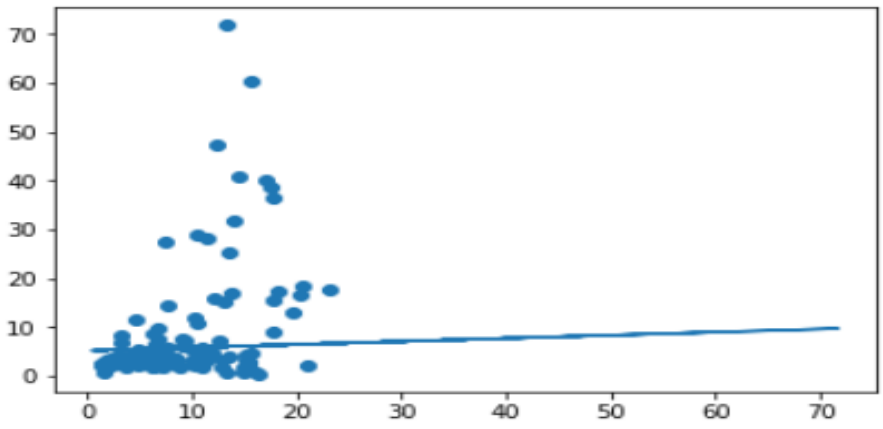
The top ranked algorithm will integrate into the proposed model. Research wishes to be a proof of concept rather than a comprehensive detail of measuring unlimited volume of algorithms and their classes.

Data mining task is very situation specific rendering most design activities to remain outside automation. How are they handled by a Data Scientist and any help that researchers can offer to these tasks. Data scientists still need to prepare data from many sources, clean it and make other environment related decisions like considering privacy issues. Breakdown of data mining tasks should be truly clear on the contribution of the human and the automated.

In the next section we explore the effort made so far to share code and design of established algorithms. The algorithms should be pluggable into any created models. Project focuses on selection of techniques and algorithms. We explore the background of most components and logic behind their design. This knowledge can transform into decision making factors.

```
#plt.scatter(x,predictions)
plt.figure(2)
plt.plot( x.iloc[:,0].values,0.06317286* x.iloc[:,0].values+5.1559)
plt.scatter(y, x.iloc[:,0].values)

plt.figure(3)
plt.scatter(y, x.iloc[:,1].values)
plt.figure(4)
plt.scatter(y, x.iloc[:,2].values)
```



Principle of matching opportunity to methods is the logical next step. Starting for scoping the work correctly, selection of the right processing algorithm is the biggest determinant of the accuracy of the results. Complexity of the output has a bearing on interpretation by the end-user.

```
airp2group22.count()
```

country_name	who_region	iso3	year	pm10_concentration	pm25_concentration	no2_concentration
France	3132	3132	3132	2590	1205	2678
Georgia	4	4	4	4	3	0
Germany	3266	3266	3266	2707	1488	3111
Ghana	4	4	4	4	0	0
Greece	164	164	164	143	57	115
Guatemala	5	5	5	4	3	4
Honduras	4	4	4	4	1	0
Hungary	193	193	193	185	68	164
Iceland	36	36	36	33	25	35
India	1301	1301	1298	1264	220	270
Indonesia	28	28	28	24	28	22
Iran (Islamic Republic of)	132	132	132	76	80	36
Iraq	16	16	16	8	4	13

Figure 47: Volumes of pollution measurement by country.

Developed countries record pollution (Fig.47) more than underdeveloped countries. Not sure if this affects the data quality.

```
x=Worldpollutionclean[['Mortality_rate_water','pm10_concentration','pm25_concentration']]
```

```
y=Worldpollutionclean['Share_deaths_pollution']
```

Data mining is the main domain, which with time has led to development of various methodologies including Context-Aware approach. Following section will looking firstly at background literature on the main domain Data Mining, then comprehensive analysis of context-aware concepts. Context or environmental features formulation and integration approaches borrow designs from a wide range of established fields.

```
from sklearn.linear_model import LinearRegression
```

```
from sklearn import metrics
```

Implementation of the full cycle of data mining will be performed as illustration on how to apply the suggestions. These key steps are only successful if the project manager has lots of environment knowledge, making key decisions more applicable.

Approach to data mining algorithm performance should consider both classes of dataset e.g., nominal, and numeric. It should also consider methods of measure rank on a key metric or an average of selected metrics. In an analysis involving forty algorithms and fourteen metrics, (Nasor & Ali, 2019) found Logical Model Tree (LMT) and Random Forest (RF) performed well on nominal datasets. Also, M5P and Linear Regression (LIR) performed well on numeric datasets.

Algorithm performance metrics are wide including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Receiver Operating Characteristic (ROC) and accuracy. (Gayati, Nickolas, Reddy, & Chitra, 2009) used a 10-fold cross validation test to measure J48, CART, Random Forest, BFtress and (NBC).

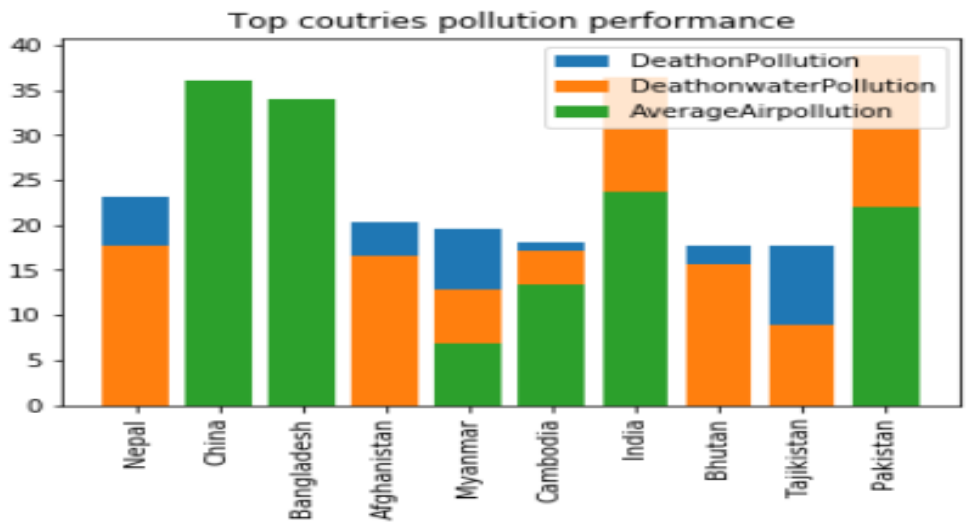


Figure 48: World Pollution view

Looking in the above Fig. 48 World Pollution view one can see elevated levels of death from water pollution for Nepal, Afghanistan, and Bhutan which happens to be an underdeveloped world. Countries are characterised by poor infrastructure. Their rate of death in air pollution is also not low.

The air pollution rate is high for China, Bangladesh, and India, which are countries with high population and highlight developed industries. Industries being a source of air pollution.

Analysis on the context of these countries with expert knowledge on causes and management of pollution would give an even better picture of the situation. The investigation attempted to bring a wide range of factors to light, but it is not exhaustive.

Chapter 5 Summary

The chapter applies the Method Selection Assistant and Context-Aware Data mining concepts to three life situations. The context of the dataset was looked at for PIMA Indians diabetics, India Road Accidents and City Pollution datasets. Each presented its own unique characteristics to explore ability of the concepts in a practical setting.

CHAPTER 6 Concept Evaluation, Analysis and Discussion

Chapter 6 Introduction

This chapter evaluates the proposed CMAS framework by comparing it with similar efforts made by other researchers. It also evaluates the CMAS solution against thesis aims and objectives. Analysis is also done on challenges and achievements during the implementation of CMAS in three real-world situations. The Discussion section also includes limitations and directions for future research efforts.

6.0 Summary of Evaluation, Analysis, Discussion of Results

Case Study PIMA Indian Diabetes

Dataset

Analysis of publicly available Kaggle with 768 observations and eight diagnostic measures. Context is built by categorizing diabetes measures based on WHO and PIMA Indian specific background knowledge. Studies on diabetes have given the amount of health complications it causes including kidney, blindness, and blindness. Knowledge on specific to diabetes and PIMA Indian women.

Data Mining Method selection

CMAS starts with understanding the situations presented, designing models with environment in mind and visualizing insights. The Data Mining method is selected based on the possibility of getting meaningful insights on co-occurrence of health indicators, so the Association method is likely to be a class of algorithms to apply. The logic in the Buddy dropped other classes that work for supervised and segmentation-based logic.

By using medical categorization, a range of health indicators were named in medical thematic that are later used when advising patients. Age could be 39 middle age, 80 Elderly

Analysis Results

Analysis of diabetes one must understand BMI Body Mass Index that is Patient Height and weight as health status indictor. BMI is a wellness indicator that links height to weight and is used to provide sensitive situation advice. When looking Obese/Overweight in WHO scale 23 to 24.9 and for Indian Women expert Scales 24.1 to 30. Advice from practitioners should be context based or else it will be wrong and endanger patients' well-being. A woman with BMI 29 kg/m2 will be classified differently by the two scales of is the advice.

Result minimum confidence being 0.75 Severely Obese, HighSkinThickness) that is patients with High Skin Thickness are also 75% likely to be Severely Obese based on WHO scaling or categorization. Looking at Conviction of values infinity would confirm that relationship between High insulin or Medium Insulin is not by coincidence linked to a health situation. Key CMAS decision is on the stage to apply the context knowledge during the Data Mining cycle.

Significance of the CMAS approach to medical domain

Data mining projects are usually performed by a team that include subject experts, their purpose is to supply contextual information. This dynamics of searching for understanding environment informing a dataset enriches the insight quality. The Recommender approach proposed here sits well with the modern technology of Generative AI where search will bring back rich knowledge. The same can also be said of knowledge repositories available in the open resources' community. The is a need for foundation of future where data mining model development is assisted by technology and existing expertise.

In real life datasets do not happen in a vacuum. Patterns in the data have varying levels of influence from their surroundings. When applying CMAS to medical domain one needs to understand disease based on expert knowledge. The expertise applies the knowledge to the aspects of the datasets. This will lead to insights explained in terms of the medical

domain. Data analysis is always done to specific audience stakeholders like medical professionals.

Case Study Indian Accident Severity

Accident and Indian Context

The Indian traffic is a complex jungle demanding actional insights that are environment sensitive. Public Kaggle Dataset is 32 columns given in 12 316 rows. Box analysis of Causalities had a result of 2 pointing to people travelling with partners.

With 32 attributes, and groupings context considerations posed an analysis path. The Indian country has unique facts around traffic and people's culture. Other countries are known to have drinking and driving problems. Context defined on the groupings and background traffic knowledge to direct.

The context may be combinations time may be in months that are then grouped together to form seasons. There are behaviors linked to seasons. Data analysis faces many challenges including contrast stakeholder requirements and data being of high dimensionality. Models are built based on understanding thematic of Accident Severity. This drives both the data transformation and link of insights to aspects that are important to domain experts. Stakeholders like car dealers may be interested in making cars that are safer than currently available ones.

Data mining Selection

The situation is matched with knowledge to confirm that it has labelling, numeric format but not interest in Categorization these being best achieved by algorithms in the Clustering class. Selection of Clustering Data Mining Method was done by considering the characteristics of the dataset and expert based knowledge of the capabilities of the method. Referencing Data Mining subject experts clustering was selected as it takes care of groupings, Patterns, and segmentations. The Semi-automated Assistant Buddy will match the situation to data mining methods capabilities. The speed of this semi-

automated selection saves analysts time compared to the trial and errors approach. One could then apply WEKA, RapidMiner to select an algorithm within the class or method.

Analysis Performance

CMAS tries to reduce the overwhelming selection of Data Mining Methods and produce situation-focused insights. Data preparation calculations like correlations need to be context considerations before being applied, for example 0.040687 driver age and light conditions may have no real-life relationship. Some attributes might have a real-life relationship like that between Accident Severity and Road surface conditions with value 0.00937. CMAS does propose to relate to domain activities as we go through as analysis stages.

Cluster quality is measured by distance between clusters and between members of the same cluster. Analysis needs to bring the whole picture and go deeper into the context of a specific view. Relationship of Pedestrians Movement against Cause of Accidents may answer worries like to what extend does Pedestrian Movement cause accidents in India as per given dataset. In this case working with 5 clusters SS 0.424 ND db index 0.788.

Discussion on the analysis of results

Contextual factors are applied at various stages of CMAS data analysis. The strength of CMAS data analysis approaches is understanding the data as it relates to the situation. The resulting insights should be easy and accurately actionable.

Insights that are blind of Context on weather, drivers or road infrastructure are likely to be difficult to convert into actionable advice. Use of context allows translation of figures into domain linked statements that may have clarity to stakeholders. Traditional / Classic data mining focuses on a given features and attributes resulting in patterns that need huge human effort to become actionable.

Models are built based on understanding thematic of Accident Severity. This drives both the data transformation and link of insights to aspects that are important to domain experts. Stakeholders like car dealers may be interested in making cars that are safer than currently available ones.

One could focus on attributes that interact with specific stakeholders. The group consists of about 10 groups like Time, Driver, and vehicle and a real-life perspective to the data. Context considered in this case like Drivers is made-up of Age band, Sex of driver, education band, Driving experience, owner drivers, vehicle driver relation. Driver is a principal factor on accidents such that related patterns may inform on actionable policies formulation.

Clusters quality is measured using Silhouette score and Davies Bouldin Index with model adjustments to improve it. CMAS works with known context of the subject, in this case it is known that 71% percent of vehicles in India two wheelers. Interest may be to find whole these types of vehicles affect accident severity.

Case Study World Air and Water Pollution

The problem is that insights from traditional data mining tend not to easily be linked to practical situations. In the Air and Water pollution use case the context that interest of pollution specialists and population affected by pollution. The other problem is traditional is traditional trial and error selection of Data Mining method apply.

Data Mining Method Selection

The Method Selection Buddy recommended Regression Method for the World Air and water Pollution problem. The Method Selection Assistant Buddy made the recommendation using situation responses from the Analyst and reference to method abilities as provided by Data Mining experts. Interactive questions asked by the Buddy can easily be answered by novice analyst with basic situation information.

Quality of Insights

This knowledge is on methods capabilities and data types it handles well. Characteristics of data attributes and need for domain understandable insights sometimes determine methods to apply.

Selection Buddy assisted in selecting a Data mining method that matches Air and water pollution situations. Buddy saved resources by avoiding the hit and miss approach used in traditional data mining. The hit and miss are time consuming and likely to lead to failure. Buddy selected a single method in case it had selected multiple one has to decide on logic to be followed by an ensemble of methods or rank them.

The selection Bubby increases the speed of creating an enriched model. Semi-Automated method selection on accumulated data mining and pollution domains. CMAS scoped the issue at hand to levels of pollution and its effects. The limited scoping encourages deeper aligned understanding of the subject.

Regression method is known to predict future given historical relationships. The result confirms speed and application of expert knowledge in a fly. Expertise knowledge is required on Air and Water pollution to inform the analysis, in real life one may be a member of a team that has many stakeholders. The Buddy's importance is bringing together lots of knowledge in a quick decision-making process.

Regression method is a result from Assistant Buddy by matching ability nature of Air and Water pollution being not categorized, its numeric and could also have some labelling.

Model Performance discussion

The prediction model attempted to allow one to foresee death rate in each country if provided Air or Water pollution levels. Algorithm data mining method Regression is measured in performance on the Air and water pollution dataset. For this thesis, the focus is on improving the whole analysis effort which result to relate to view of pollution professionals. Data mining method has performance measures improved by overall analysis approach, CMAS. Insight on death rates relationship with levels of pollution is relevant to subject professional. The analysis is highly linked to the pollution situation and the model is optimized through data preparation.

Quality of Regression solution may be measured in Mean Square Error (MSE) or Root Mean Square Error (RMSE). Expert knowledge informs analysts in this case to understand Air and Water pollution are measured. The results like RSQR and MSE measure how good a model is in predictions.

The overall effort of CMAS approach is to attempt to produce insights that can be easily understood by pollution experts or other stakeholders. Air and Water Pollution is a vast domain of vast dimensions of interest. Expert meaning and importance of pollution measures like PM 10 Concentration or PM25 tempcov assist in understanding resulting insights.

Based on proposed approach software was used to recommend possible data mining methods likely to produce usable insights around Air and Water pollution. The solution refers to data and method capabilities as accepted by Data Scientist over the years.

Context-Aware considerations

Dataset development

Dataset one 40 100 rows and 9 columns

Dataset two 195 rows and 20 columns

Dataset three 68 540 rows and 4 rows

The first challenge is to develop a context-aware input, in this case data that can inform pollution experts. One interest of these experts is to reduce pollution side effects which includes death rates.

The combined dataset presented features that are relevant to domain experts. CMAS was applied to Air and Water pollution context to produce domain linked insights.

Given that any one dataset it failed to tell a meaningful story, the need for full context pushed analysts to search for more datasets. Air and Water pollution dataset is a combination of three original public datasets to build a local and context perspective. Data Analyst need to think around what would be of interest to end users of the insights for example Pollution Experts. Data transformation to have common keys like country and average values forced for dropping features like Time.

Domain interest points to need to address by insights to inform practical decisions. Pollution is influenced by population density, level of industrialization, water treatment facilities but analyst considers these to be out of scope for this analysis phase. The results or insights linked well with health concerns around the pollution issue. A known negative

effect of pollution is increase in death rates. Analysis to enable prediction of death rates is helpful in focusing Pollution Scientist effort to improve quality of life.

The whole analysis process integration of three datasets to enable the model to produce insights that directly relates to the practical situation of levels of pollution and related death rates. Application of the approach looking at Air and Water pollution in the context of causing death. The selected method Regression developed an optimized model that produced insights focused on the subject. The terms embedded into the insights is from the Pollution domain.

Data mining resource optimization through pruned focus of effort on appropriate perspectives of the challenges. Water and Air pollution analysis only focused on recording of levels of pollution and related death rates. Results of a few linked attribute ..... relationship

Pollution could be affected by population density, infrastructure, industrialization are possible additional situation or context to be considered. Accepting resource limitations makes CMAS encourage right levels of scoping. Define well what is being covered and what will remain outside the current analysis process.

6.1 Datasets

To illustrate and evaluate the novel framework, end-to-end analysis was performed on three different datasets using the proposed approach. Three use cases on Indian PIMA diabetes, Indian Traffic and World Pollution datasets were used in the case studies. The datasets were explored, resulting in the following summaries. Understanding the challenge at hand and the logic of solution design is clearly informed by the characteristics of the data.

6.1.1 Indian PIMA Diabetes Dataset

The case study Indian PIMA Diabetes is a publicly available dataset that consists of statistics from 768 women of PIMA Indian origin. The data have eight attributes listed below in Table 9. Data analysis was done using two perspectives. The WHO and PIMA Indian specialist views defines Context of the analysis. The question is whether the use of two viewpoints improves the resulting analysis. Data are a mixture of wellness indicators and inherited factors that medical studies view as affecting the likelihood of having diabetes.

Indian PIMA features table

Feature	**Medical definition**	**Value Range**
Number of pregnancies	There might be a of volume of pregnancies and having diabetes	0 to 8
Plasma glucose concentration	Plasma glucose concentration tested more a two-hour oral glucose tolerance test Between 70 and 130 mg/dl milligrams per deciliter normal range	70 to 200 mg/dl

Triceps skinfold thickness (mm)	Amount of fat reserves in the body	18.7 (+/- 8.5) mm
Age (years)	Age of participants given most diseases kick in with age	0 to 90
Diabetes pedigree function	Family line affect the likelihood of developing diabetes in life	1 to 0
2 – hour serum insulin (mm U/m)	High level signal insulin resistance low point to diabetes or pancreatitis	2 to 20 mcU/mL
Body mass index	That is weight in kg divided by height in m2	18.5 to 24.9
Blood pressure	Normal being systolic less than 120 and diastolic 80 mm/Hg	120/80
Diabetes Status	Whether person is diabetic or not	0 or 1

Table 10:PIMA Features Summary

The below table 10 shows categorisation of values of the BMI attribute makes analysis more aligned to physical. Every human field or domain have their own terms to communicate situations. These are medical terms that are used mostly to describe health problems related to weight.

Categories	**WHO international ranges kg/m2**	**Indian Women kg/m2**
Normal/ healthy weight	18 to 22	18.5 to 21
At Risk weight	23	21.1 to 24
Obese/overweight	23 to 24.9	24.1 to 30
Severely Obese/Obese range	>25	>30

Table 11:Context of BMI categories

The CMAS framework promotes consideration of context during analysis of data, as mentioned before, and this can happen at any stage of the development cycle. In this case context was added to the data preparation stage based on established expert knowledge. The effect of these contexts can be seen throughout the stages of data analysis. At the point of interpreting outcomes one can easily link knowledge to the insights.

Looking at a woman of BMI 22 kg/m2 the WHO will say that they have a healthy weight, but if the subject is of Indian PIMA she has a weight that is at risk. Medical practitioners relaying on our analysis are going to give a more situation sensitive advice which is more appropriate.

6.1.2 Indian Accidents Severity Dataset

Road accidents and related injuries are of great concern for humankind. Data mining and Context-aware concepts can play a role to give insights on this issue. The data set (Kaggle, 2024) has 12316 rows and thirty-two columns on Indian road accident records. Data preparation, thirty of the columns are in text categories that had to be mapped for data analysis. For example, Sex_of_Driver column had female and male mapped to 0 and 1, respectively.

Data can be analysed from causes of accidents, grouping can be weather, road features, junction design, and Impacted object like vehicle. Features around an aspect are many to give a clearer picture for example Time and Day of the week can be more informative when used together. Another example Driver Information like (Age, driving experience and education level)

Various stakeholders are interested in different contexts of data for decision-making. Columns can be grouped by context for example Time Context, Driver Context, Road Context or Vehicle Context. Road contexts consist of (Area_accident_occured, Lanes_of_medians_Road_alignment, Types_of_junction etc)

Results and analysis can be tailored to Stakeholders requirements, allowing for focused analysis and better application of resources. Government officials may target reducing accident severity by understanding the Road Context influences. Therefore, context approaches can be used to resolve the high-dimensionality challenge commonly faced in data analysis processes.

Table of Feature / Attributes for the Indian Traffic Data

Feature name	**Data type**
Time	object
Day_of_week	object
Age_band_of_driver	object
Sex_of_driver	object
Educational_level	object
Vehicle_driver_relation	object
Driving_experience	object
Type_of_vehicle	object
Owner_of_vehicle	object
Service_year_of_vehicle	object
Defect_of_vehicle	object

Area_accident_occured	object
Lanes_or_Medians	object
Road_allignment	object
Types_of_Junction	object
Road_surface_type	object
Road_surface_conditions	object
Light_conditions	object
Weather_conditions	object
Type_of_collision	object
Number_of_vehicles_involved	int64
Number_of_casualties	int64
Vehicle_movement	object
Casualty_class	object
Sex_of_casualty	object
Age_band_of_casualty	object
Casualty_severity	object
Work_of_casuality	object
Fitness_of_casuality	object
Pedestrian_movement	object
Cause_of_accident	object
Accident_severity	object

Table 12: Traffic feature list and data types

Work through data from python ‘describe’ command.

6.1.2 World Pollution dataset

Data analysis needs to be guided by reality on the ground and expertise knowledge. Reading on pollution experts point to three contexts affecting and informing the situations. It was necessary to fetch for datasets covering death from Air Pollution, death rate from Water Pollution and air pollution rate data. The resulting analysis was enriched by these broader contexts leading to more informative insights.

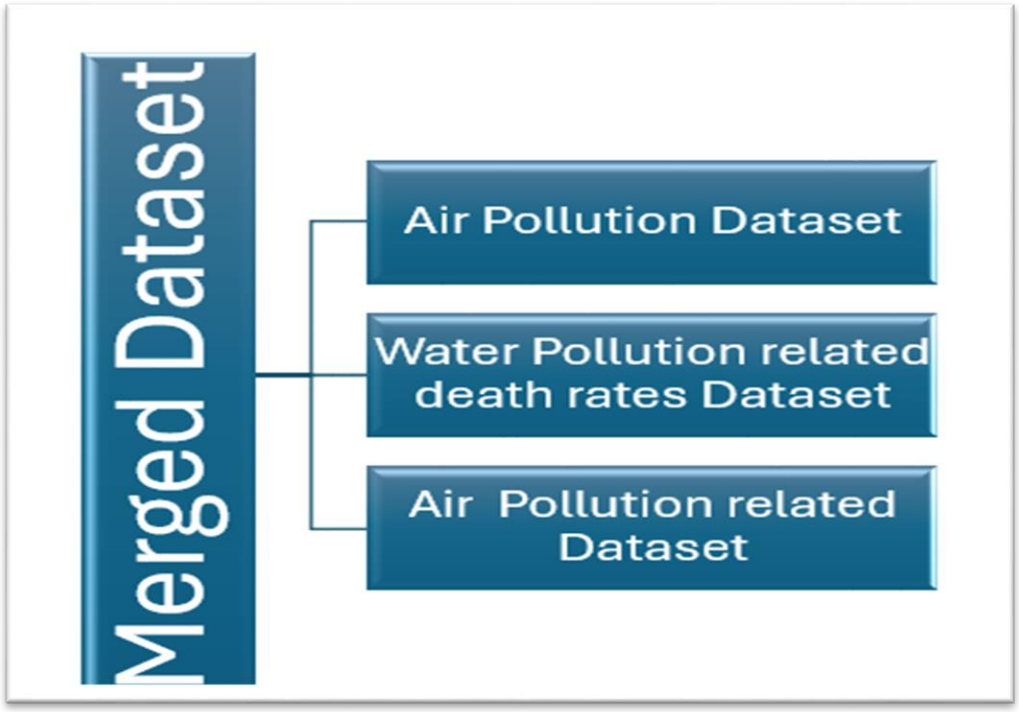


Figure 49: World Pollution Dataset Building

Fig. 49 shows the datasets used to create the data frame used in the World Pollution use case. Understanding problem at hand and Data Preparation tone and direct all remaining steps in the framework. Built based on country names, the merged dataset has 110 rows and nine columns, created from averages in three input datasets. Air Pollution Dataset (World Health Organisation, 2024) had 40 100 rows with twenty columns, Water Pollution Death 195 rows and four columns and finally Air Pollution related death 6840 rows on four columns.

Feature	Description
Share of death air pollutic	Percentage of country wide death that is linked to Air Pollution
Share of death Water poll	Percentage of country wide death that is linked to Water Pollution
pm10 concentration	Annual mean concentration of particulate matter with diameter of 10um or less (ug/m3
pm25 concentration	Annual mean concentration of particulate matter with diameter of 2.5 um or less (ug/m3
no2 concentration	Annual mean concentration of nitrogene dioxide (ug/m3
pm10 tempcov	Annual temporal coverage for pm10 on a 100 base (full year = 100; 1=1% of the year
pm25 tempcov	Annual temporal coverage for pm2.5 on a 100 base (full year = 100; 1=1% of the year

Table 13: Air and Water Pollution Features

The pre-processed data set is based on country data, and most attributes had a row count of 110. On death rates Share of death related to air pollution had a row count of 110. Rate of mortality linked to water has a count 106. This means four countries provided air data but no water data.

The context-aware approach encourages understanding of a domain throughout the analysis life cycle. Around data exploration it can inform on outlier and rows that can be removed from analysis. Data provided in percentage like death attributes are straightforward to make sense of. The air pollution data needs expert support to understand. All concentration columns (pm10, pm25 and no2) had a minimum of 0.0. The averages of pm10 concentration 47.93, pm25 concentration 23.94 and no2 concentration 16.73 may indicate disturbing pollution statistics. The available data also shows maximum values of pm10 concentration 273.93, pm25 concentration 119.77 and no2 concentration 125.23.

6.2 Evaluation

6.2.1 Performance Metrics

6.2.1.1 PIMA Indian Diabetes Use Case (Association Rule)

The PIMA Indian Diabetes analysis was performed using the Association Rule data mining method. Association method produces many rules whose information and occurrence need to evaluate. A rule says the items are likely to occur together. In medical, diabetic situation in particular common occurrence may tell conditions that practitioner need to be wary of. The evaluation check if items are occurring together by coincident and other measures. Metrics like Support, Confidence and conviction were used to understand rules found on this dataset.

Support

An Association rule is a basic if then logic, with the antecedent (*if*) being an itemset in the dataset. The consequent (*then*) being an item found together with the antecedent. In the medical case, a health condition or set of health conditions is being checked as to other conditions it occurs with.

A diabetic patient with 'HighSkinThickness' antecedent is likely to be 'Middleage' consequent.

'HighSkinThickness' → 'Middleage'

Support is a measure of a given dataset's IF/THEN relationship. Investigations are based on how and why two health conditions are found in a patient together.

$$\text{Support } (X \rightarrow Y) = \text{Frequency } (, Y) / n$$

(Akram, 2024)

X being first condition in this case ‘HighSkinThickness,’ Y being another condition ‘Middleage’ and n being a population size (total number of patients)

Importance and understanding of values from Association metrics need to do with expertise guidance, a low Support value on a large dataset may still warrant further investigation. Other metrics interpretation distortions have been found with situations of (High Support and Low Confidence or High Confidence and Low Support). The closer Support value is to value 1 the more relevant and important the rule might be. Support is a pointer for further investigation.

The building of items requires specification of computation metrics in this example min_support, giving control to the level of relevance of outcomes from the function. A range of metrics can now be extracted from the Association Rules. The thesis will just give a few as focus is more on Context integration than general model building.

	SUPPORT	ITEMSETS
0	0.5000	(HIGHINSULIN)
1	0.7500	(HIGHSKINTHICKNESS)
2	0.5000	(MEDIUMINSULIN)
3	0.5625	(MIDDLEAGE)
4	0.5625	(NORMALBLOODPRESSURE)

Support is the number of patients in which a specific range of health attributes occurred. ‘Highskinthickness’ has Support of 0.75 points for many patients with diabetes who have this skin state. As we progress a health attribute becomes a set of attributes occurring together, itemset. So, members of an itemset start from one to many, also called set length.

Confidence

Association rules should be measured using the Confidence metric, the reported relationship is TRUE. In other words, it is a measure of chance for an item being found based on the rule.

Confidence = Support (X and Y) / Support(X)

X being the first condition in this case ‘HighSkinThickness’ and Y being another condition ‘Middleage.’ The equation checks that the regularity item X ‘HighSkinThickness’ and Y 'Middleage' occur together based on the number of X 'HighSkinThickness' occurring.

A high Confidence encourages making decisions based on the co-occurrence. In this medical situation advise maybe to work to prevent combinations of condition to minimize frequency of diabetes or reduce complications.

Due to result interpretation challenges the analyst needs to decide on a practical minimum value known as support threshold. Any processing after this step will prune those that are not within the limit. Working below uses a threshold of 0.5.

```
frequent_itemsets1=apriori (df, min_support=0.5, use_colnames=True)

print(frequent_itemsets1)
```

==== Basic output from the Association workout ===

```
0.7500 (SeverelyObeseBMI, HighSkinThickness)
```

We could say with 75% minimum confidence that for patients with diabetes.

Those with HighSkinThickness are also 75% likely to be SeverelyObese based.

on WHO scaling (WHO Context)

There are many terms that are key to implementation of apriori analysis. Filter products that are below a set frequency are known as non-frequency itemsets. On further processing the system removes itemsets with non-frequent itemsets.

Conviction

Association rules are investigating dependence of items on each other. X is likely to be there if Y is there. Conviction measures levels of dependence or lack of independence. Also confirming if rule was found by chance. When high, Conviction points to Y being reliant on X.

Conviction = Probability (X) Probability (! Y) / Probability (X, Y)

Sign “!” stands for the logical NOT.

	ANTECEDENTS	...	CONVICTION
0	(HighInsulin)	...	inf
1	(HighSkinThickness)	...	1.125000
2	(Post Prandial) ...		1.083333
3	(HighSkinThickness)	...	inf
4	(SeverelyObeseBMI)	...	1.000000
5	(MediumInsulin)	...	inf
6	(Middleage)	...	1.687500
7	(PostPrandial)	...	1.137500
8	(Middleage)	...	inf

Figure 50: Conviction Report for Indian Diabetes

Conviction checks whether a relationship between two health measure occur together by chance or real. Above fig. 50 HighInsulin and MediumInsulin is infinity meaning the relationship is not found by coincident. Conviction of one like that of SeverelyObeseBMI shows that attribute is independent.

6.2.1.2 Indian Accident Severity Use Case (Clustering)

The Indian Accident Severity data has thirty attributes detailing valuable information on each accident. Attributes can be put into six categories based on viewpoints. Each attribute has named values for example Attribute Area_accident_occured has values Residential, Office and industrial areas.

The model aims to group accident data points by similarity. Clusters can then be integrated in physical world terms. Model performance may be improved by trimming data based on context. There may be extra patterns when analysing accidents as the relate to four attributes linked to say Weather Conditions. Evaluation of performance of the model on separate contexts will inform on the overall performance of the CMAS context-aware approach.

Decent quality cluster are said to have high similarity. That is, you have a high similarity of points in the same cluster, inter-cluster distance. There should also have low intra similarity that is distance between elements in neighbouring cluster is huge. Quality of clusters was measured using Silhouette score and Davies-Bouldin Index

Silhouette score.

Silhouette score or Silhouette coefficient has value range from -1 to 1.

$$\text{Silhoutte} = (n - i)/\text{Max}(i,n)$$

n is mean nearest cluster distance, distance between each sample and nearest cluster that the sample is not part of.

i is mean intra-cluster distance, is distance been samples in the same cluster.

Three score values 1, 0 and negative have special meanings. A silhouette scores closer to value 1 is good as it signifies that the groups are further from the others. Low score like zero shows points in the cluster decision boundary, near neighbouring clusters. When Silhouette is zero, it points to error in classification of the data point.

```
print(f"Selhouette score (n=5): {silhouette_score(scaled_df,labels1)}")
```

```
Selhouette score (n=5): 0.18251953048825922
```

Silhouette score of 0.182

Is a measure of the severity of key measures of Clustering results, these are inter-cluster distance, elements within a cluster should not have high distance from each other. The other is intra-cluster distance, distance between elements in different clusters. The intra-group distance needs to be high for a good clustering solution. A silhouette is not too close to one pointing to clusters that are not well apart.

Davies-Bouldin Index

The Davies-Bouldin index metric has advantages over other measures. A low index value is preferred is a result of low intra-clustering and high inter-cluster distance. Below is a basic equation to calculate the Davies-Bouldin index.

Davies-Bouldin Index (*internal evaluation technique*) can be calculated as follows:

$$Davies - Bouldin\ Index = \frac{1}{c}\sum_{i=1}^{c} \max_{j \neq i} \frac{\sigma_i + \sigma_j}{d(c_i, c_j)}$$

(Alla, 2024)

c number of clusters

ci centroid of cluster i

d (ci ,cj) distance between two Centroids

qi average of distances of all clusters in i and c

Davies Bouldin Index for the Indian Accident models.

```
print(f"Davies_Bouldin Index (n=5): {davies_bouldin_score(scaled_df,labels1)}")
```

```
Davies_Bouldin Index (n=5): 1.9271706409066982
```

Davies Bouldin index is 1.927.

Traffic data was split into clusters that were measured for quality. This was a way to find the performance of the clustering method. Davies-Bouldin Index find Similarity average for those that are most like the cluster.

6.2.1.3 World Air and Water Pollution Use Case

The World Air and Water Pollution dataset has two dependent variables, it also has six independent variables … The Regression model aims to enable prediction of a dependent variable value given values of the independent variables. In the context of pollution, the scientist is interested in reducing the death rates linked pollution factors.

Performance was measured using the following.

- R-Squared (R2) performance metric
- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)

RSQR is the R-Squared (R2)

The R2 is given as a percent to which the regression model developed can accurately predict, say, the rate of death from air pollution given a reading of PM10 concentration.

The R2, measure of goodness fit, informs on the effectiveness of the regression model. It also evaluates extend to which independent variable influence dependent ones in a clear term.

$$\text{R2 Squared} = 1 - \frac{SSr}{SSm}$$

SSr = Squared sum error of regression line

SSm = Squared sum error of mean line

(regression-metrics, 2024)

The R2 is the square root, where the SSR is the sum of the difference between the actual and predicted values. Predicted values are in the plotted line, while actual values are data points.

SST total squares measure dependent variable like the Death Rates in the Regression model

```
RSQR = r2_score(y,predictions)
```

```
print(RSQR)
```

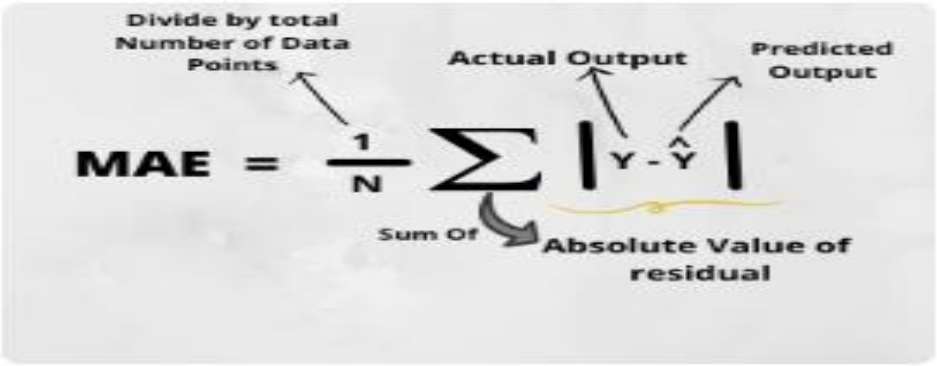
```
0.4861296777947278
```

The root square value of 0.486 points to 48 % of the prediction for the value of ‘Share of death from pollution’ can be explained from values of the three factors Mortality due to water pollution, the concentration of PM10 in the air and the concentration of PM25 in the air.

Mean Absolute Error (MAE)

As a performance metrics MAE measures model predictions against the actual values in the dataset. MAE produces the absolute discrepancies of the sets of values.

MAE



Advantages of MAE

Accepted working for the MAE value.

n number of points in the dataset.

y observed value of point i

ŷ predicted values in the regression line

Air and water pollution Mean Absolute Error (MAE) values.

Air and Water pollution MAE

```
MAE = metrics.mean_absolute_error(y, predictions)
```

```
print(MAE)
```

```
2.909512571217287
```

Mean Absolute Error (MAE) is known not to be over affect by outliers. Outliers being attributes with values far from the population averages. It shows the average difference in predictions and actual values. In this case the model is predicting values of Shared of Death from Pollution.

Mean Squared Error (MSE)

Results are squared prediction error average to improve interpretation. The MSE is rate of wellness of a prediction model.

$$MSE = \frac{1}{n}\sum(y - \hat{y})^2$$

Comment (The square of difference between actual and predicted)

MSE is calculated using ‘n’ as total datapoints.

Sum of difference y and ŷ squared divided by datapoint count.

y observed value of point i

ŷ predicted values in the regression line

(Agrawal, 2024)

Air and Water Pollution Regression model MSE

```
MSE = metrics.mean_squared_error(y, predictions)
```

```
print(MSE)
```

```
13.54419004996983
```

MSE, good and why trimming helps.

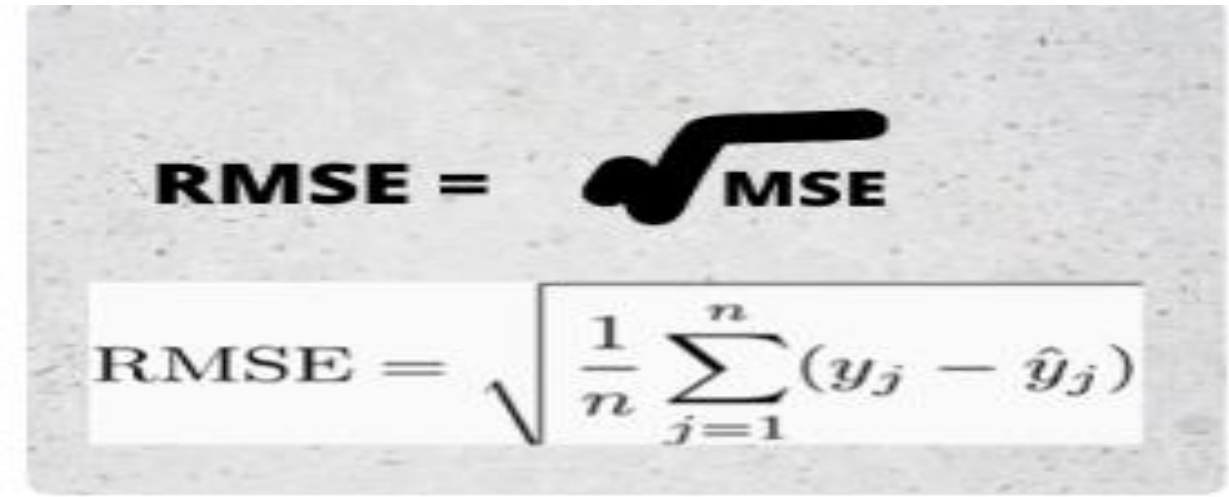
Mean squared Error (MSE), is the amount error of statistical Regression model on Air and Water Pollution dataset. The model enables predictions, therefore there this measure difference between predicted values and observed. The value 13.54 shows that the Regression model has less error it produced more precise predictions.

Root Mean Squared Error (RMSE)

The Root Mean Squared error (RMSE) handles well continuous numeric input values.

The RMSE quantify average distance between predicted values from the model and actual values from the Air and Water dataset. A low value of RMSE means the model can fit the Air and Water dataset. In other terms it represents the square root variances of residuals.

Giving values of the same units the RMSE gives an approximate error of the prediction model.



Square root of the MSE

y observed value of point i

ŷ predicted values in the regression line

:

Root Mean Square Error for the Air and Water Pollution below.

```
]: RMSE = np.sqrt(metrics.mean_squared_error(y, predictions))

]: print(RMSE)
   3.6802432052745955
```

This being the average error made by the model in predicting the share of Death from pollution given the three predicted attributes. It is low (3.68) showing that our context developed model is performing well. There might be other factors, but effort invested in making analysis domain focused surely improves model performance.

Table of Case Study working towards objectives

Objectives	**Case Study 1 PIMA diabetes**	**Case Study 2 India Traffic**	**Case Study 3 World Pollution**
Create Context-aware Framework	Context build on variance of classification of biometrics from World Health Organisation and those that are specific to the PIMA Indian people.	Context of Indian Traffic linked to the high volumes of bicycles and tricycles. Context of lighting, infrastructure etc was investigated.	Three datasets sought to complete the Context of pollution. Pollution can be air or water based. Results of pollution include increase levels of Death
Semi-Automated Data Mining Method selection	Novice Analyst feed characteristics of the situation at hand to recommend a few possible appropriate data mining methods. In this case Association Analysis	Clustering was selected by the Method Buddy. Ability of Clustering to group attributes based on values and characteristics. The amount of Context that need to be processed for a full understanding of insights	Nature of the data being numeric and attribute relations being informative in prediction pushed the Method Buddy to promote Regression Data Mining Methods
Evaluate Resources utilization	Scope limited to variance when looking at the same data from WHO	Context defined and trimmed according to intended use of insights	Data was transformed to averages as target was to find relationships. Some of

	perspective and a given population.		the datasets were collected over many years.
Application of domain knowledge	Readings and knowledge were sort from medical papers	India traffic and accidents severity knowledge was explored to premise that data analysis	General link of population death rate and the pollution in environment. Country was used to demarked areas of study.
Alignment of analysis and domain thematic	Medical and more Specifically Diabetes terms from understanding of issues to interpretation insights	Context illustrated in a concept lattice. Need to borrow approaches from other knowledge-based fields like philosophy	Understanding of pollution metrics sought before undertaking data analysis task
Easy of Insight Interpretation	Medical and PIMA terms are in the data analysis results making it easy to interpret.	With Context insight can easy be applied to situation by stakeholders like Government departments	The Regression equation can have features of interest can be substituted to tell a real-life story

Table 14 : Case Study Construction logic

6.2.2 Thesis Hypothesis, proposed framework, and case study

Hypothesis one

H1) Implementing Data Mining with considerations of contextual factors improve the quality of the resulting insights. Quality taken as to include relevance to the physical world, focus, ease of interpretation and level of insight applicability to solution development. Therefore, predict that Data Mining performed in a Context-aware approach will result in high quality insights.

Alternate Hypothesis

H1a) Implementing Data Mining with consideration of contextual factors negatively affects the quality of resulting insights.

Null Hypothesis

H1n) Implementing Data Mining with consideration of contextual factors does not affect the quality of resulting insights.

Case study 1 PIMA Indian Diabetes – Consideration of the expert medical context of world-wide perspective and one specifically for the Indian population result in more appropriate and accurate insights.

Case study 2 India Accident Severity situation – Splitting focus by context allows insights to be distributed per given stakeholders. Analysis for government officers can be for Road Context etc.

Case study 3 World Pollution – Analyst created a better understanding of pollution by having a wide Context from the data perspective.

Hypothesis two

H2) Data Mining could be aided by an application that selects appropriate Data Mining methods for a given situation. Therefore, predict that a Data Analyst using software to select a DM method will find the process simpler and produce a model that reflects years of domain experts' knowledge.

Alternate Hypothesis

H2a) Software method selection negatively affects Data Scientist making of appropriate model design decisions.

Null Hypothesis

H2n) Software method selection does not affect Data Analyst ability to make appropriate model design decisions.

Method Selection Buddy: - The scripts used in all three use-case referred to known expert knowledge on capabilities of Data Mining method to select the most appropriate one to apply. The decision is made by matching situation to methods abilities.

6.3 Results

The CMAS framework was used to analyse the medical diabetes dataset, activities in India's roads, and the pollution challenges facing the world. Below are results from this contextual linked analysis. As detailed before, Context-Aware data mining adds an effort to consider the effects of situation on patterns being investigated.

6.3.1 PIMA Indian Diabetes Use Case

0.7500 (SeverelyObeseBMI, HighSkinThickness)

We could say with 75% minimum confidence that for patients with diabetes.

Those with HighSkinThickness are also 75% likely to be SeverelyObese based.

on WHO scaling (WHO Context)

With understanding that the data were looked at on two contexts WHO and local experts' perspectives, the above outcome will be more informative with this background. Other than the source dataset, general expert knowledge was necessary to create a meaningful analysis model. The thesis investigates the benefits of considering context when performing data mining. Data exist in a real-world environment; therefore, it is influenced by this. Analysis without fully understanding background is likely to be bias in the wrong direction.

Frequent Itemsets were developed using a dataset of diabetes data from the Indian population of PIMA. Set minimum support to 0.5 for the analysis to focus on interesting diabetes health indicators linked to well-being. The itemsets are combinations such as

'HighSkinThickness' and 'SeverelyObeseBMI', which are analysed to direct further research effort.

The model generates rules, that is, pairs of itemsets that have relationships to be investigated. The rules on this instance had confidence metric above 0.6

Below is graph of relationships between Confidence and Lift for the PIMA Indians diabetic disease in the population.

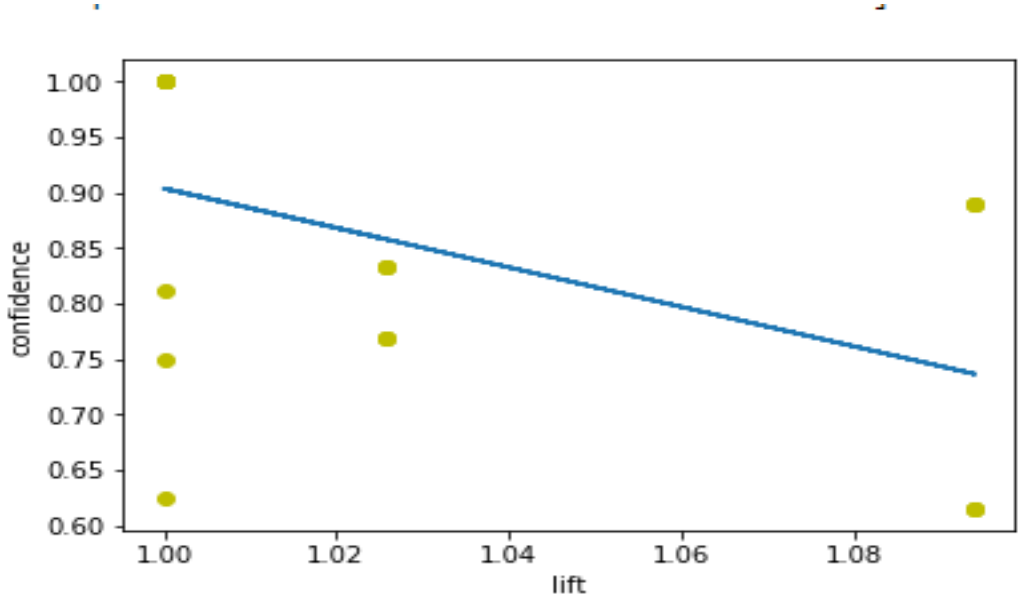


Figure 51: Itemsets Confidence and Lift graph.

Fig 51 indicates that Confidence decreases as lift is increased for this model and dataset.

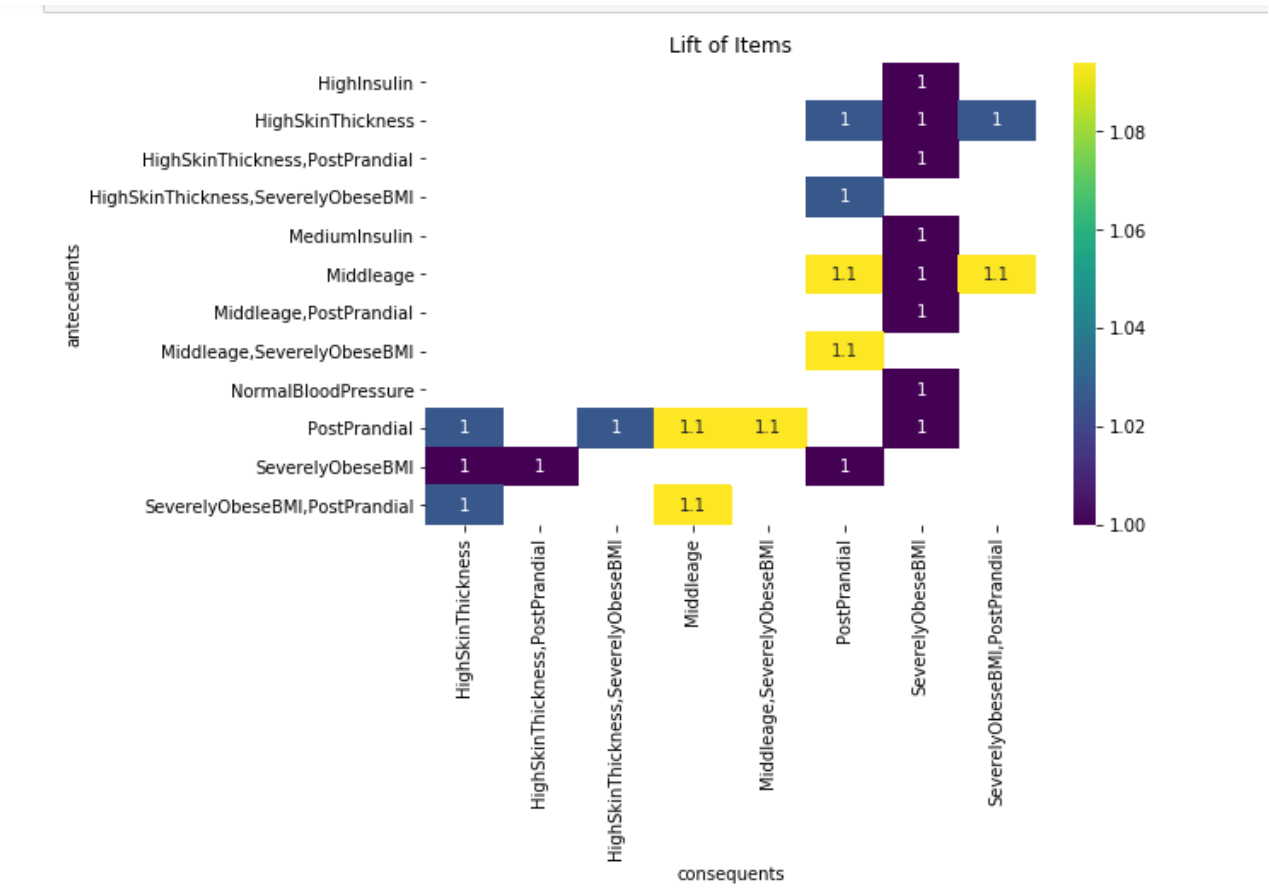


Figure 52: Heatmap for itemsets and Lift metric

dfas[dfas['antecedents'] == 'SeverelyObeseBMI']

```
dfas[dfas['antecedents'] == 'SeverelyObeseBMI']
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
4	SeverelyObeseBMI	HighSkinThickness	1.0	0.7500	0.7500	0.7500	1.0	0.0	1.0
10	SeverelyObeseBMI	PostPrandial	1.0	0.8125	0.8125	0.8125	1.0	0.0	1.0
16	SeverelyObeseBMI	HighSkinThickness,PostPrandial	1.0	0.6250	0.6250	0.6250	1.0	0.0	1.0

Figure 53: SeverelyObese focus

BMI Body Mass Index is a health measure of great interest as it looks at a person's height and weight. The model further analyses rules were Antecedents, left side of the Rules being 'Severely Obese.'

Sensitivity Analysis of Association Rules

Analysis of Severely Obese related rules the below checks changes on average Support, Confidence, Lift, Leverage and Convictions as the minimum confidence is changed.

Min Confidence	Support	Confidence	Lift	Leverage	Conviction
0.6	0.6875	0.6875	1	0	1
0.8	0.8125	0.8125	1	0	1
0.2	0.6015625	0.6015625	1	0	1

Table 15: Confidence Sensitivity Severely Obese

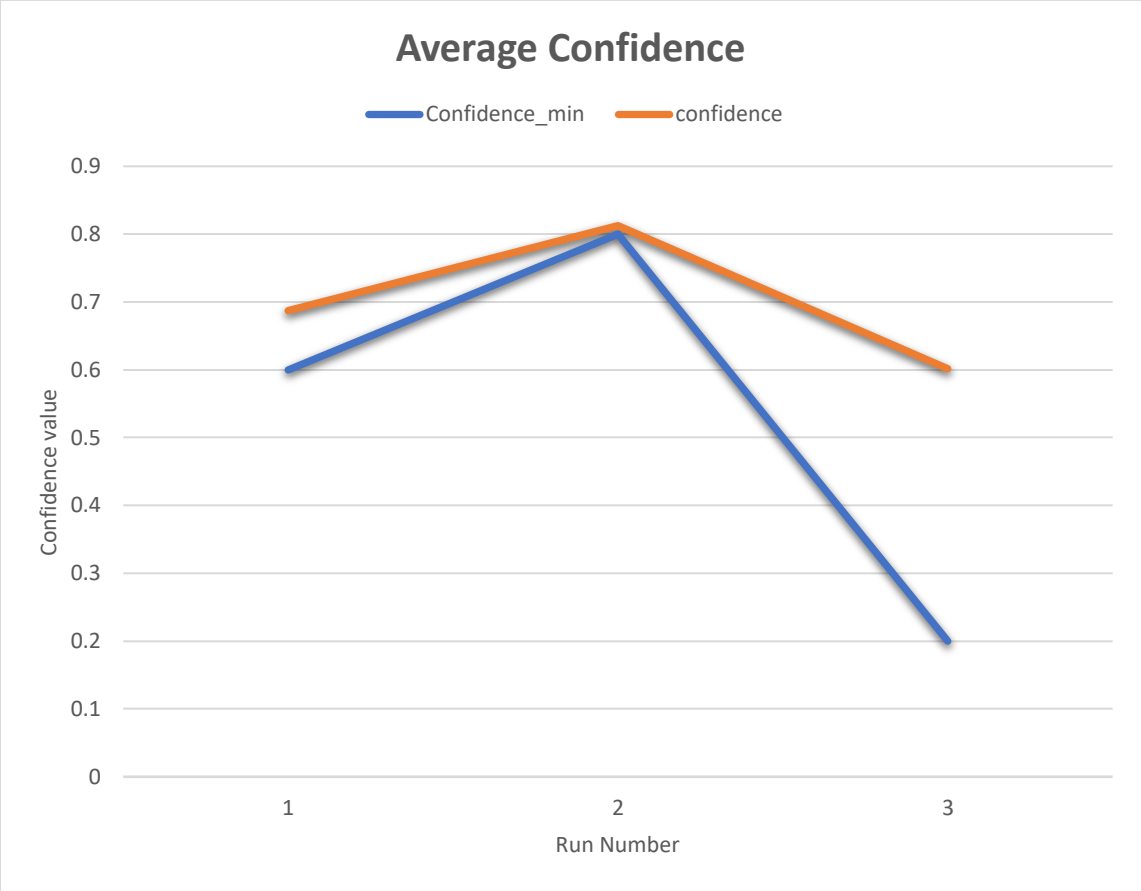


Figure 54: Change of other metric with min confidence

Minimum Confidence Sensitivity shows Confidence being around 0.6 for minimum of 0.2 and 0.6. as shown on Fig. 54.

Varying Conviction and other metrics

min conviction	antecedent support	consequent support	support	confidence	Lift	leverage	conviction

0.5	1	0.601563	0.601563	0.601563	1	0	1
0.7	1	0.601563	0.601563	0.601563	1	0	1
0.3	1	0.601563	0.601563	0.601563	1	0	1

Table 16: Table of Conviction Sensitivity

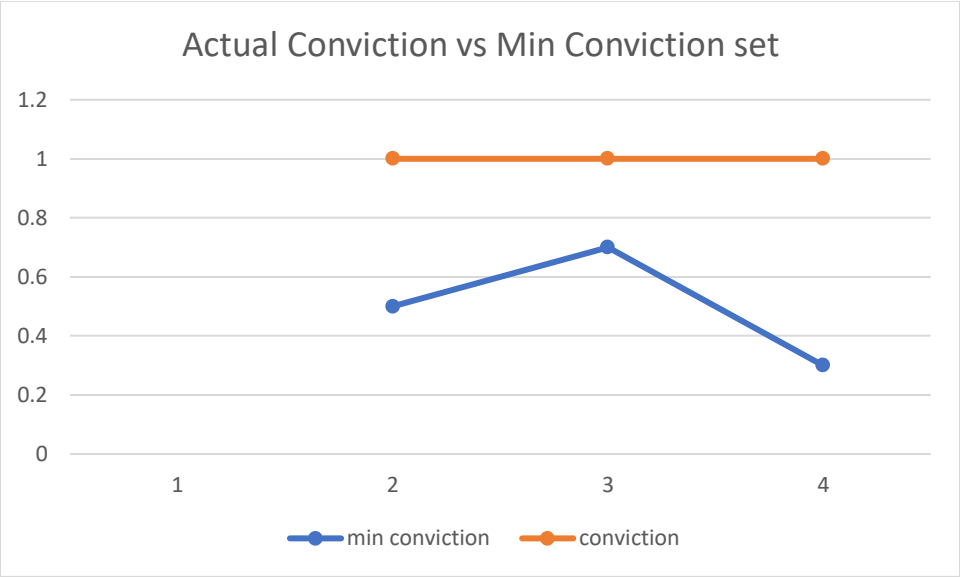


Figure 55 : Obese Conviction Analysis Graph

The fig 55 Conviction, Support and other metrics seem not to change with changes in Minimum Conviction.

6.3.2 Indian Accident Severity use case.

The Indian Accident Severity dataset represents a situation of many interacting factors. Application of Context-Aware data mining concepts enables investigations that can be broad, looking at whole picture in one, on the other hand deeper context specific view. In this section we measure performance of a Clustering Model.

The following covers variance analysis of performance measures Silhouette Score and Davies_Bouldin Index on the same data and all other parameters. Clusters were also analysed for varying context, and sensitivity analysis of the Silhouette Score when number of clusters is changed.

Summary Clustering Performance

Firstly, used sets of attributes to get clusters of attribute values reporting on number of clusters, silhouette score and Davies_Bouldin index.

Indian Accidents Cluster Model Performance table

Data Attribute	Number of Cluster	Silhouette score	Davies_Bouldin Index
Vehicle type, Pedestrain, Movement, Cause of Accident, Accident Severity	**5**	0.4247916969211908	0.7878770218408153
Pedestrian Movement \and cause of accident	**3**	0.5755602237594535	0.5824447643543826
Road surface conditions, Type of Collision, Accident severity	**3**	-0.012391981682854323	70.52902877902302
Vehicle Movement, Pedestrian Movement and Accident severity	**4**	0.019224335863483568	39.029979851113026

Figure 56: Cluster Model Performance Data

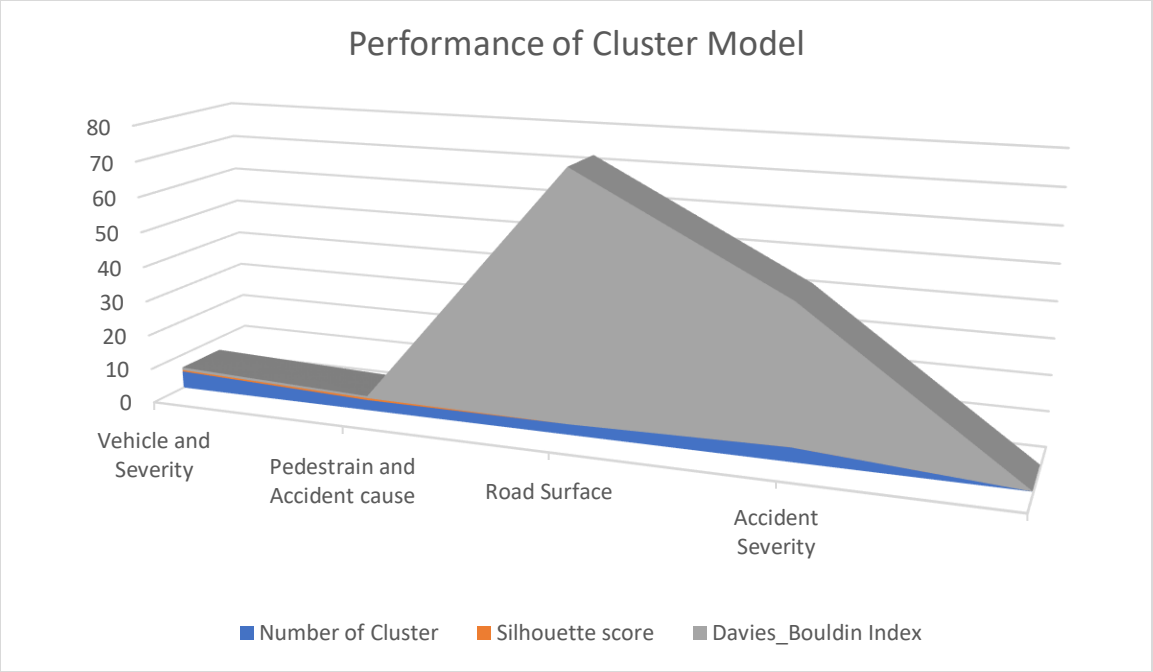


Figure 57: Performance Graph for Indian Data

Performance of the Cluster Model shows fig 57 a high Davies_Bouldin index for clusters linked to Road Surfaces Conditions. The numbers of clusters used in the model appears to have limited influence on model performance.

Silhouette score cluster volume relationship!

Next, we investigate relationship of number of clusters to the performance measure Silhouette score.

Silhouette Score vs number of clusters graph

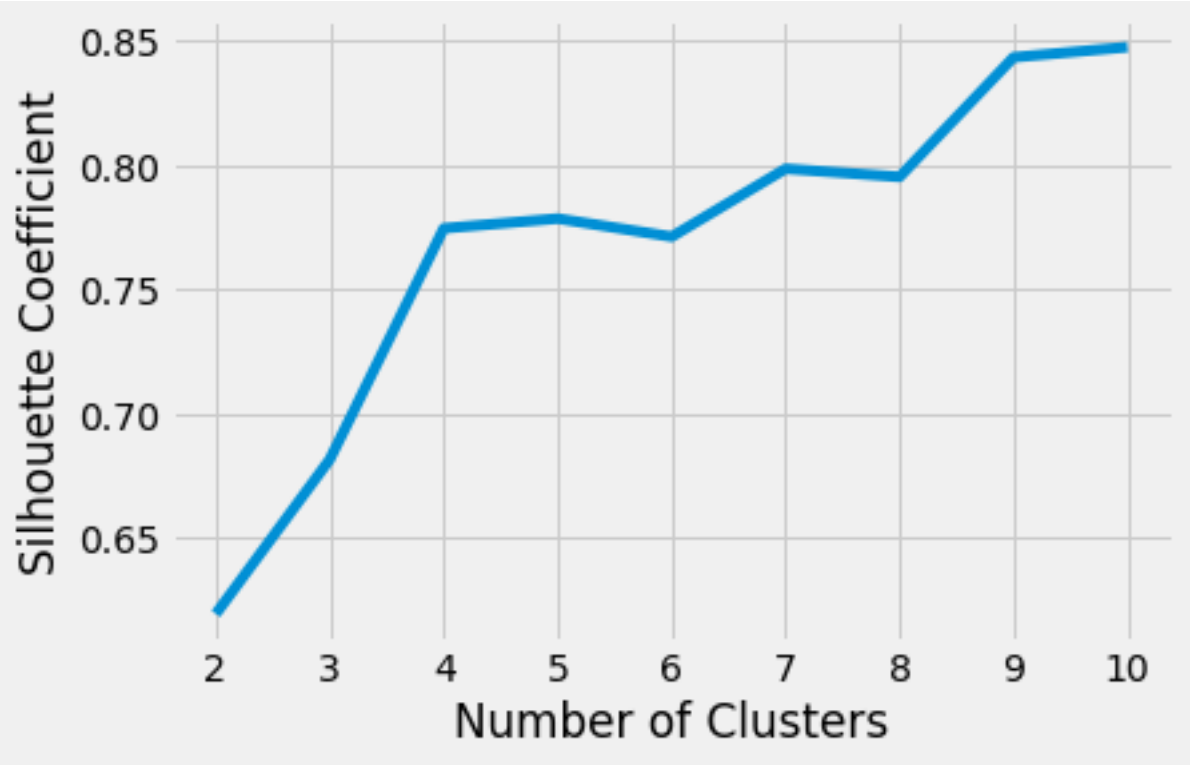


Figure 58: Silhouette clusters volume relationship.

The Silhouette score for the Indian traffic cluster model increases with the number of clusters considered.

Accident Severity, Pedestrian Movement and Vehicle Type

Stakeholders interest in reducing Accident Severity, they will limit investigations to the context of four attribute Types of Vehicles, Pedestrian Movement, and cause of accident. Analysis of Type of Vehicle, Pedestrian Movement, Cause of Accident, and Accident Severity using 5 Clusters. This has given the following grouping; change of attributes or number of clusters will change the shape on the graph.

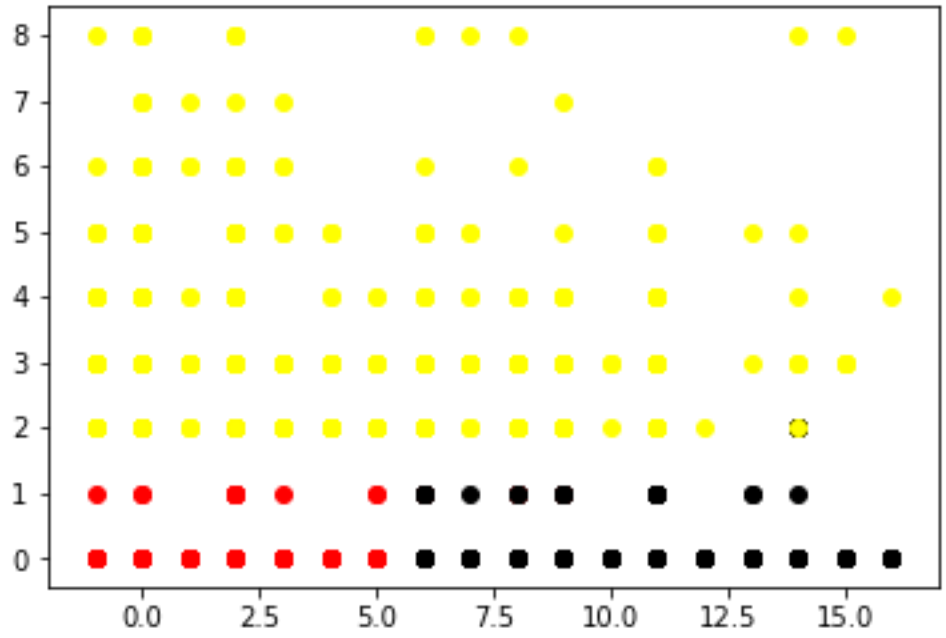


Figure 59: Clusters of Accident Severity given other four measures.

```
Silhouette score (n=5): 0.4247916969211908

Davies_Bouldin Index (n=5): 0.7878770218408153
```

Pedestrians Movement and Causes of Accident Analysis

When accidents occur, there are causes and people affected. It is of great interest to know the contribution of Pedestrians to causing accidents. Changes may be made to separate pedestrians from traffic. In large cities there are Pedestrian zones where traffic is prohibited.

Checking the contribution of pedestrian movement to the causes of accidents. Analysis was performed using three clusters based on the Elbow graph below.

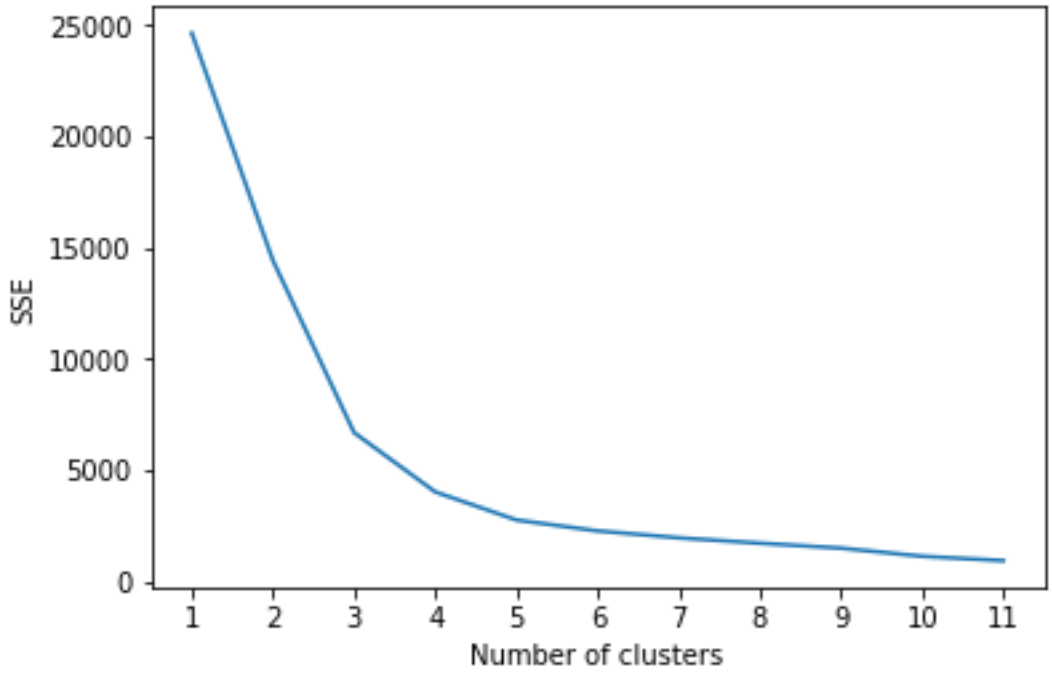


Figure 60: Pedestrian Accident Cause analysis clusters volumes

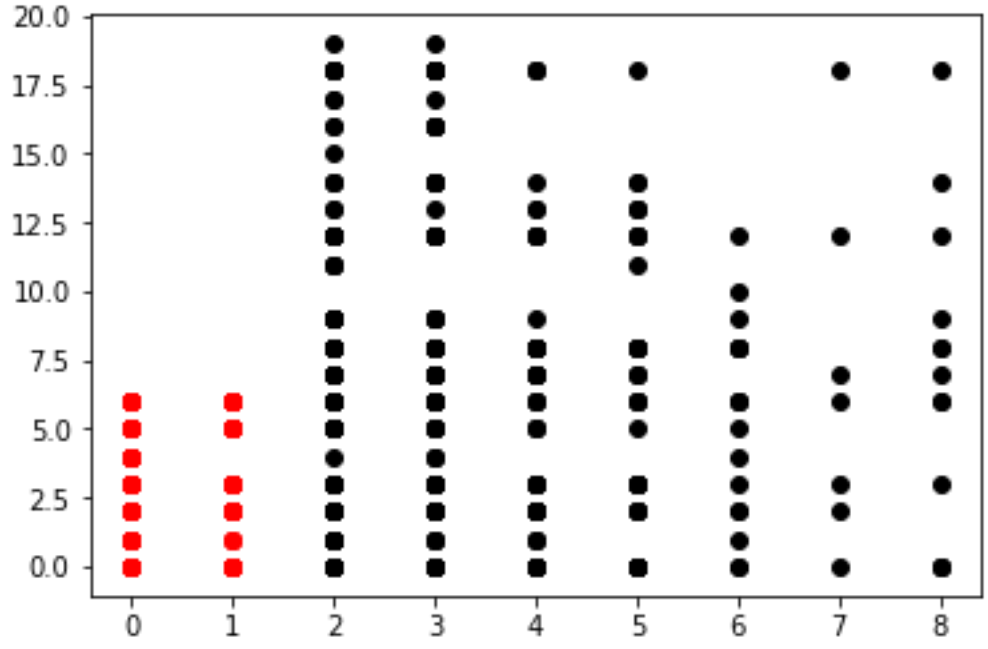


Figure 61 : Pedestrian and Accidents clusters.

```
Silhouette score (n=5): 0.5755602237594535
Davies_Bouldin Index (n=5): 0.5824447643543826
```

Accident Severity relationship with Road Surface Conditions and Types of Collision

There is so much that governments and local authorities can do to improve the road surface conditions to minimise some types of collisions. The results from this analysis can inform these decisions. Using three clusters on groupings values of road surface, type of collision, and accident severity. One may want to make decisions based on attribute relationships.

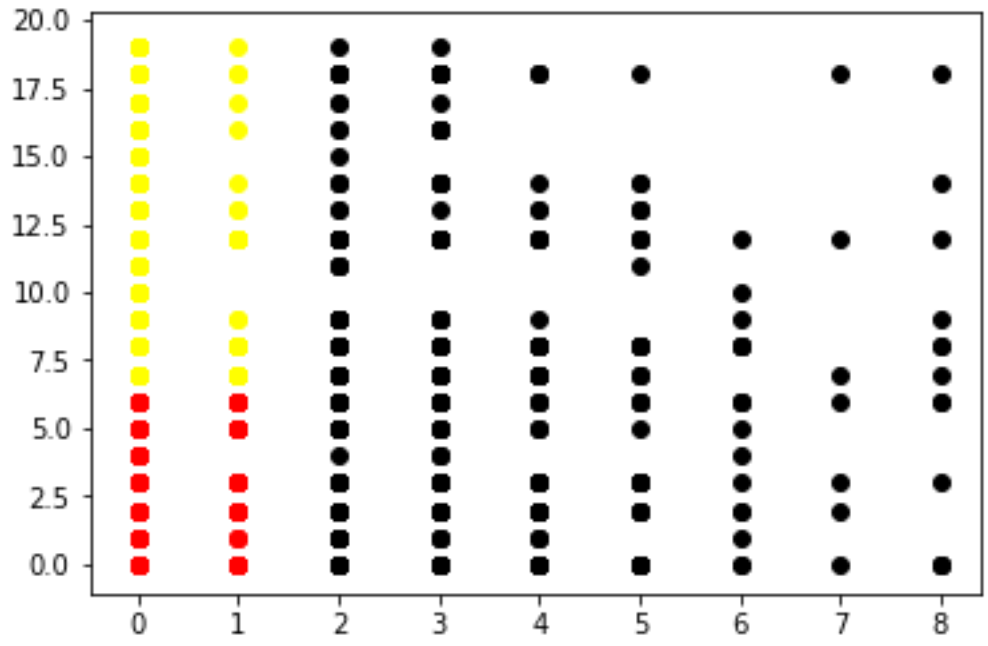


Figure 62: Clusters of Accident Severity attributes

```
Davies_Bouldin Index (n=5): 70.52902877902302

Silhouette score (n=5): -0.012391981682854323
```

Accident Severity caused by Vehicle Movement and Pedestrian Movement

Cluster model analysis of vehicle Movement, Pedestrian Movement and Accident Severity using four clusters as per the bent below.

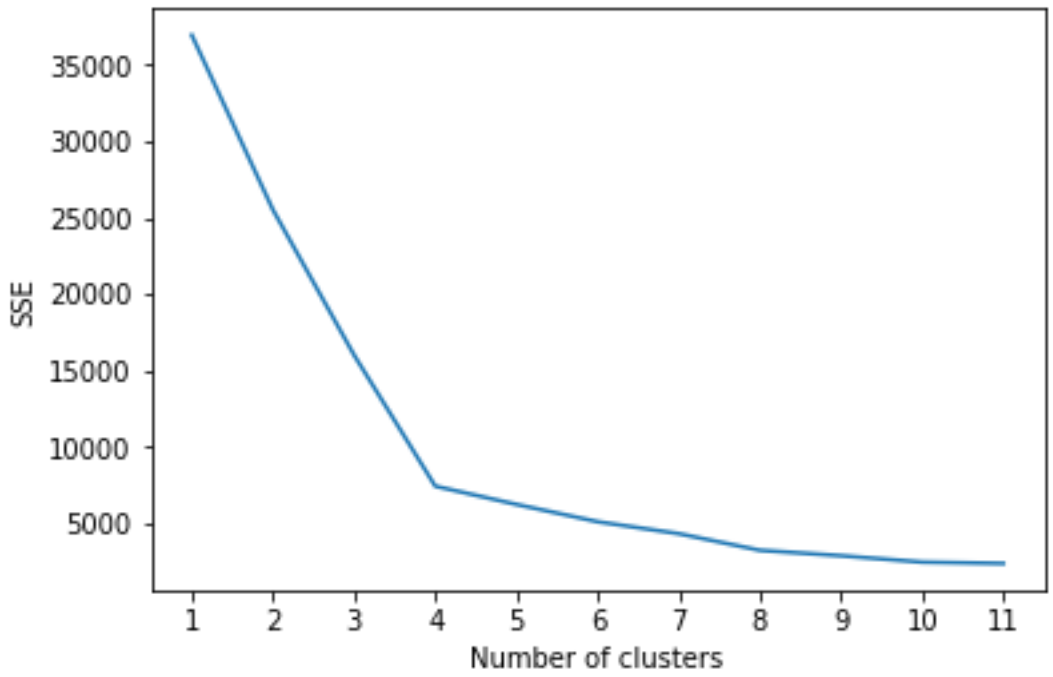


Figure 63: Finding the right number of clusters for a model.

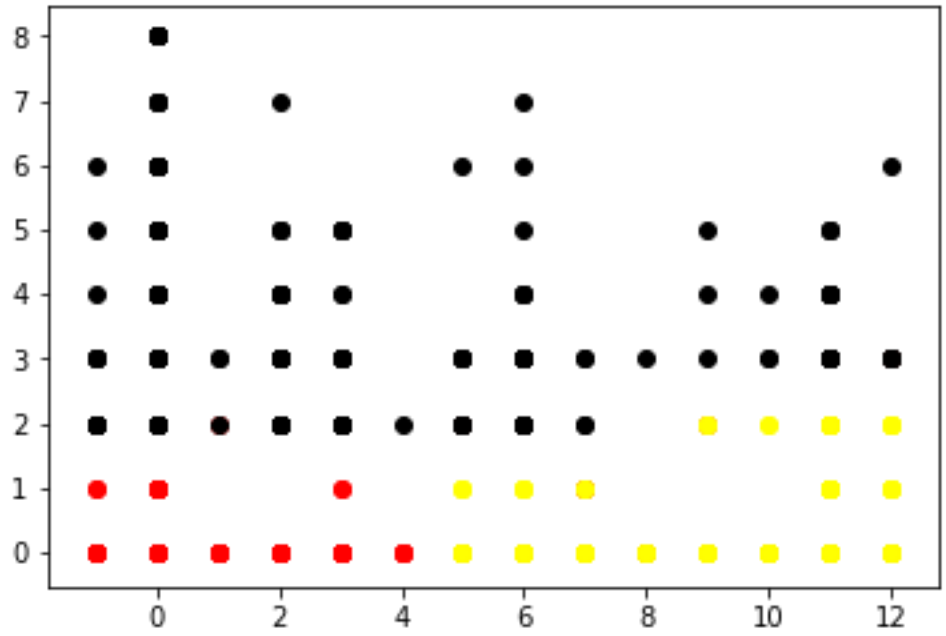


Figure 64: Accident Severity linked to movements.

```
Silhouette score (n=5): 0.019224335863483568

Davies_Bouldin Index (n=5): 39.029979851113026
```

With four clusters the plot represented the above fig 64. Each cluster is represented by a colour.

Gaussian Mixture

Below applied different Cluster analysis model based on Gaussin Approach.

	Light_conditions	Weather_conditions	Type_of_collision
0	Daylight	Normal	Collision with roadside-parked vehicles
1	Daylight	Normal	Vehicle with vehicle collision
2	Daylight	Normal	Collision with roadside objects
3	Darkness - lights lit	Normal	Vehicle with vehicle collision
4	Darkness - lights lit	Normal	Vehicle with vehicle collision

Table 17:Sample of the dataset

```
Gaussian Mixture(components=3)
```

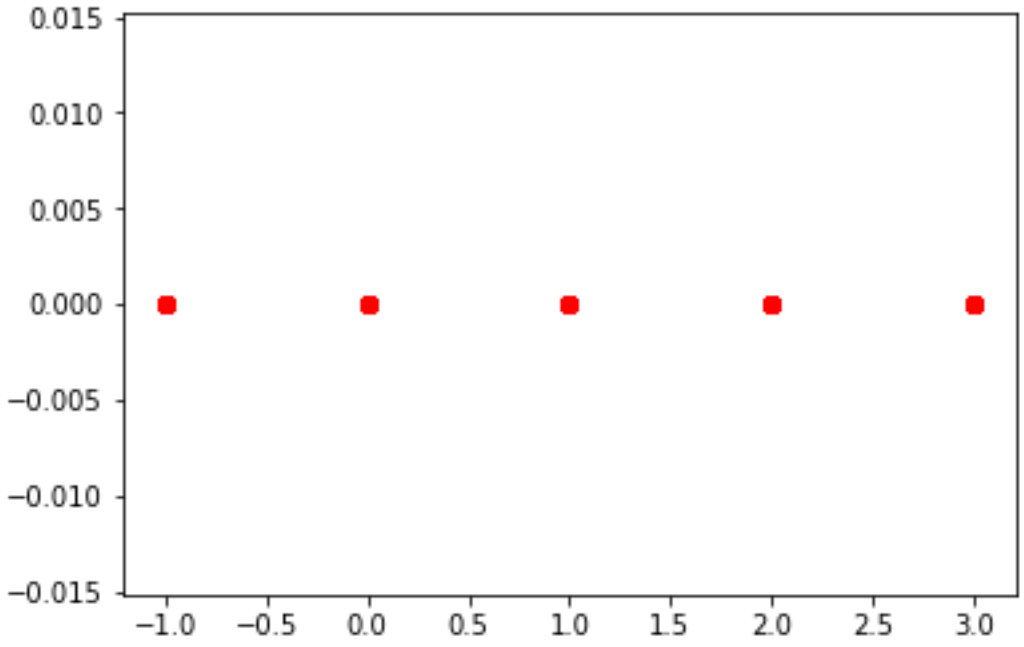


Figure 65:Gaussian Mixture Cluster plot.

The above cluster plot if different from the other ones, this could be due to the data subset or principles behind that approach.

6.3.3 World water and Air Pollution use case.

Outliers

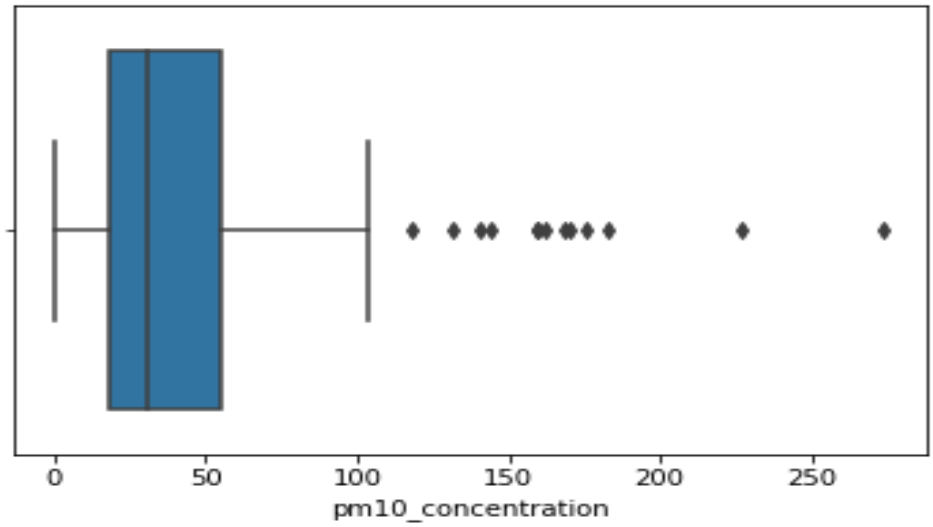


Figure 66 : PM10 concentration box plot

Fig 66 shows that for pm10 air concertation, those above 100 are outliers which may be left out of analysis as they may negatively impact outcomes.

There might be interest on how levels of PM10 concentrations relate to Share of death due to pollution worldwide.

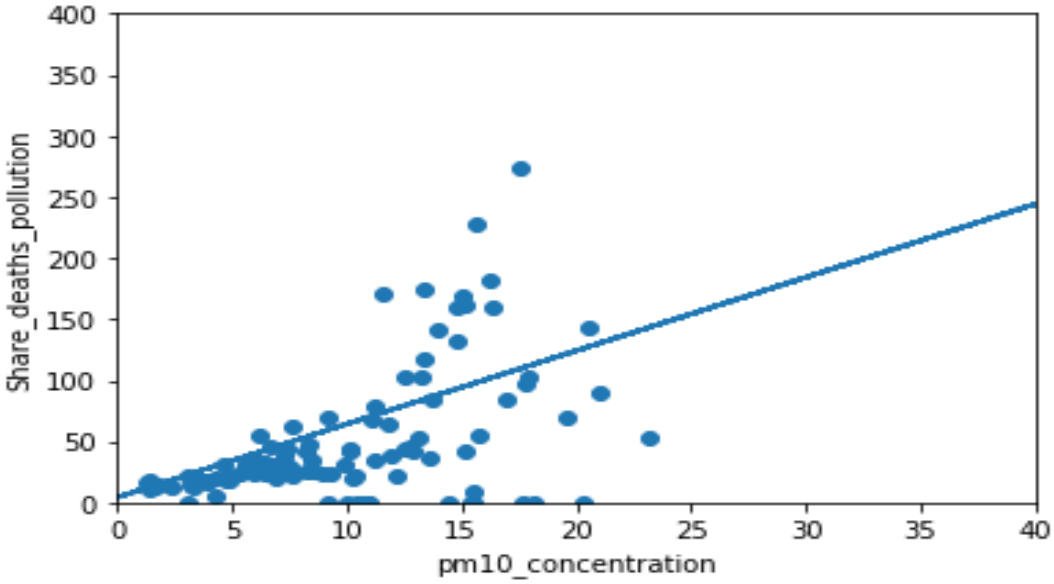


Figure 67: Regression plot of death rate and level of air pollution.

Fig 67 above confirms that most rates are below 50 with air pm10 concentration below 10. The distribution become sparse with air concentrations above 10. The Prediction line can be used to predict rate of death give an air pm10 concertation value.

Changing approach by making the death rate an independent variable and pollution a dependent one returns interesting information.

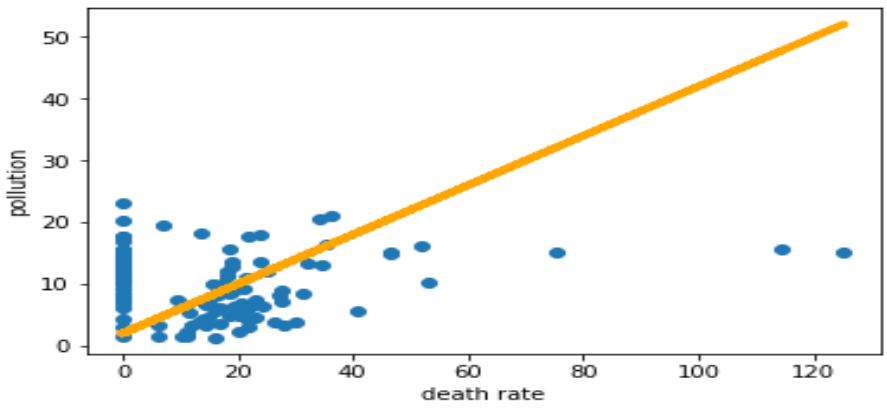


Figure 68: Regression with death rate as independent Variable

Fig 68 says many countries have death related to pollution being near zero. Most death rate are near twenty which corresponds to pollution rates below 10. Note death rate point80, 125 and 120 could be treated as outliers. Considerations may be made to exclude them from the analysis.

Least Squares Lines

Measuring strength of a regression analysis by analysing the Least Square measure over some metrics.

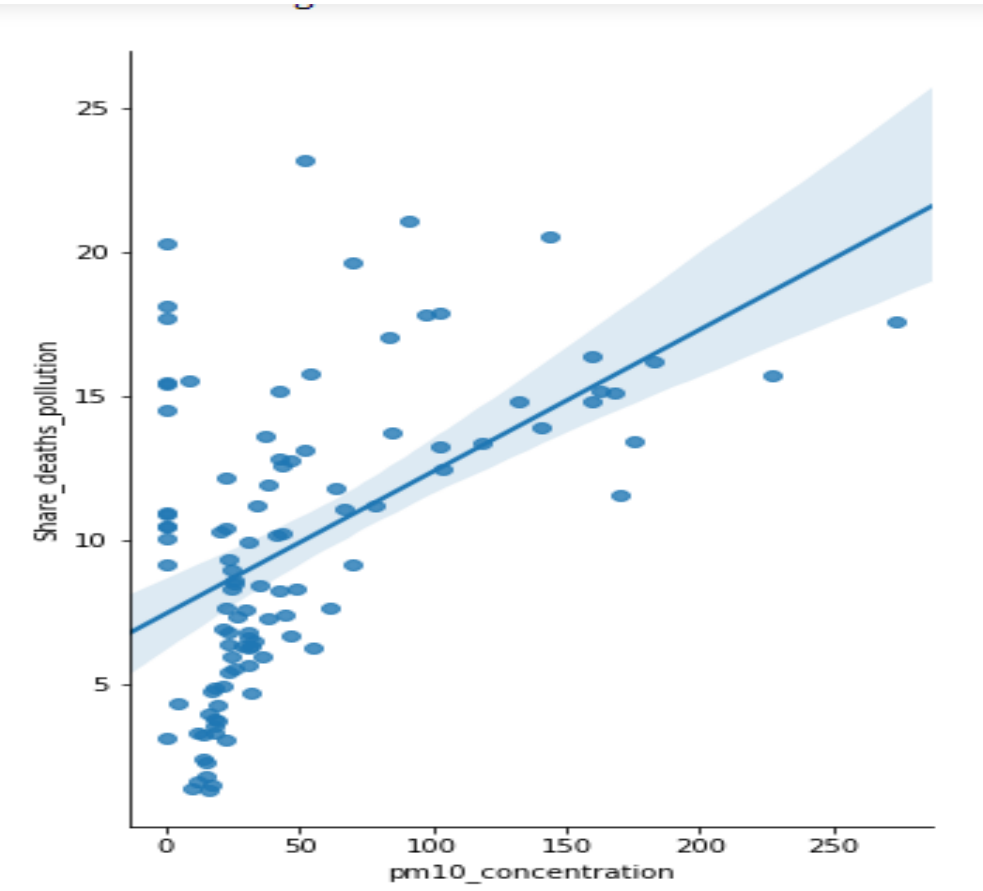


Figure 69: Least Square Lines Graph

The spread of the Least Squares after concentration above 100 pm10 informs model developers on data ranges to concentrate on.

Heatmaps

The map visualizes interesting relationships in a dataset.

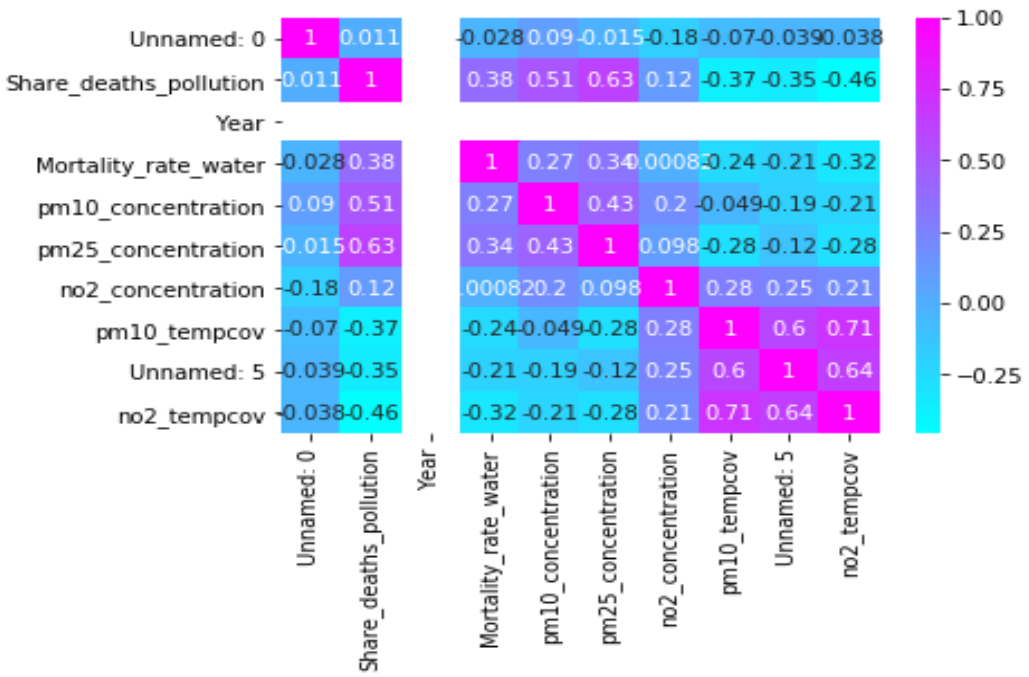


Figure 70: Heatmap for the pollution and death rates worldwide.

The coefficient of 0.63 between PM25 air concentration and share of Death is of interest as it relates a single pollutant to death rate allowing for scientist to workout intervention strategies.

Another interesting relationship is the 0.38 coefficient of Share of death by Air pollution to Mortality Rate linked to water pollution, wonder if areas of high-water pollution are also areas of high air pollution. There might be other factors influencing like levels of medical infrastructure.

6.4 Analysis

Past research has indirectly measured the performance of context-aware frameworks by has to be done. Adoption of the framework is towards improving the whole analysis journey. The following paragraphs look at the current study versus past approaches in line with a holistic view. Later, we move to benefits like relevance, focus, lean scoping and saving of processing resources.

The CMAS framework follows the six CRISP-DM steps Business Understanding, Data Understanding, Data preparation, Modelling, Evaluation and Deployment with modifications of adding context were appropriate. Understanding of contribution of CMAS must be done at four levels, analysis of the framework outline, method selection aspect and effects on the performance of a selected algorithm.

Analysis of CMAS framework performance is complex. This evaluation needs to be of the whole approach, not just the data mining algorithm. The relevance of insight to the section of users is key; so is the reduction of dimensions. The approach might also reduce the amount of effort needed for processing data and translating results into insight. The algorithm performance evaluation metrics may be applied to the Association Rules, Clustering or Regression as shown by the three use cases. This is of benefit if focus is on the algorithm being given pre-processed lean input. Otherwise, they do not serve much purpose for the proposal since the novel CMAS approach is meant to improve the whole analysis life cycle.

Instead, the research will analyse the key objectives of insight domain relevance, ease of model development and inherent trimming from perspective focus of context-aware approach.

6.4.1 PIMA Indian Diabetes CMAS application

The context-aware approach CMAS is likely to be aligned to domain semantics and view of the world. Looking at the PIMA analysis expert knowledge and terms were used in the

design of the model. As a result of this approach the insights will be easily understood by medical stakeholders. Therefore, the insights are easy to interpret for the relevant stakeholders, medical partitioners.

Diabetes is a complex disease, from health indicators on the patient to complications resulting from the disease. For the PIMA Indian population studies have confirmed the characteristics of this population. The application of CMAS allows for these unique aspects to be factored into the analysis. Data mining performed on the nature or structure of the PIMA dataset was modified in response to the context approach.

Specific contexts for medical, traffic, and pollution were looked at in the use cases. The medical data analysis was performed supported by experts in medical science supported by experts in factors related to diabetes Melius and a specific population. Analysis was more informative since it was built on foundation of expert knowledge. In terms of words, a dataset surrounded by an influencing environment. Investment in considering situational factors is likely to produce relevant information on the subject and a focused analysis effort.

(Schnor, 2024), looked at PIMA data from a dietitian perspective using Data Mining Algorithms focused on disease onset to predict those likely to develop it.

CMAS was applied to PIMA analysis stages Business Understanding, Data Understanding, and Data preparation. Business understanding was done through reading medical references which detail the population characteristics relating to diabetes and those of the world according to the World Health Organisation (WHO).

6.4.2 India Traffic CMAS application

CMAS was applied to the Indian Traffic dataset from the point of looking at complex situation by splitting and focusing on contexts involved. Clustering data mining was

applied to the whole dataset and subsets like Drivers Context. In this case the key indicators Cause of Accidents and Accident Severity was analysed in relation to Contexts

Context-aware should increase analysis dynamics and allow analyst to make focused decisions. For example, in the Indian Traffic use case one can streamline analysis to a few contexts depending on an area of interest. A vehicle dealer might be interested in introducing safer vehicles to negate accidents experienced by tricycle and bicycle users. The streamlining may lead to analysis of reduced amounts of data. Resource saving will be at many stages of the analysis cycle up to insight creations stage.

India Traffic data has been analysed by researchers like (Autnafa & Kour, 2017) who tested tools like (R, WEKA, RAPID MINER, KNIME) resolve complexity of this context and explore efficiency on resources. Data was mined using Clustering method, CMAS improvements to the process could be on streamlining focus to one context at a time. The clarity and reduced volume of attributes is likely to produce results that are high quality and linked to the contexts.

The Indian Accident Severity dataset is complex but the application of CMAS managed to split it into subsets that are context specific. The split helps in scoping and trimming the whole analysis into manageable stakeholder specific chunks. This change simplifies that process and makes outcomes easy to be related to domains.

6.4.3 World Air and Water Pollution CMAS application

The third user case World pollution benefited from CMAS determination of key factors affecting a subject. The insight was more informative especially for researchers interested in pollution and its effects to humankind survival.

Context-aware, data mining, method/tools, scoping, data domain, context concept, resource optimization, insight domain alignment, Application at multiple stages of data mining.

Hybrid or ensembles suit mixed context and data from multi sources (Represa, Fern'andeza-Sarri'a, Porta, & Palomer-Vazquez, 2019) proposed a framework including Clustering, Time Series and Classification among many tools to resolve to understand Air Pollution data.

Views of environment interacting with the dataset increase the level of reality in both processing and understanding data. These factors could make the context-aware solution vastly different but also more relevant to the given situation. Research effort needs to increase in understanding features and the environment in which the dataset finds itself. Efforts of understanding environmental improve both significance of insights and ease of application to real life concerns.

6.4.4 CMAS vs other frameworks

Proposed novel data mining approach CMAS follows and adds activities to the six stages developed for the CRISP-DM methodology. Context-Aware approach does focus the analysis and promote some pruning effort during most stages in the life cycle of these data projects. The research aims to improve the quality of insight from data mining by incorporating facts of environment surrounding the subject dataset.

Scenarios Platform Collaboration Data Mining Context-Aware Data Mining (SP-CCADM) added context to analysis. Performance was measured by comparing accuracy of a range of classifications algorithms Deep Learning (DL), decision tree (DT), Gradient Boosted Tree (GBT) and k-mean neural network (k-NN). CMAS has checked performance for Association, Clustering and Regression algorithms for three real life contexts.

CASP-DM structured along the six stages of CRISP-DM added models to promote re-use like (Versatile, Reframe and Revised models) and context. Context was broken into changes, plots, descriptions, goals, and success criteria. Communications of improvements was done through analysis of a real-life situation. The thesis looks at PIMA

diabetes, Indian accident severity and World pollution situations. Analysis integrates expert knowledge and Context data to create a situation specific data understanding. Step by step reflection of the Content-Aware data processing does highlight the success of the novel framework. Both CASP-DM and CMAS show an improved focus on a situation at hand with domain perspective.

CADM focuses on improving prediction accuracy through context and local data hybridization. Noted a 76.61 % increase in accuracy due to increase in availability of context data. CMAS used the World Pollution dataset to predict based on a Regression algorithm. The wide breadth of the source data may improve accuracy of the prediction.

Comparison of frameworks considers many factors and measures. Measures may not be clear needing references and definition to remain within a confined space and focus. Proposed CMAS approach does allow efforts to be focused on purposefully targeted scope, care need to be taken not to limit model view on the challenge. The process should not leave out perspectives that significantly affect the data. An initial need assessment should be undertaken before incorporating context into a data process effort. Perceived benefits of CMAS on a given data mining situation should justify the added activities.

6.5 Discussion

.

The thesis developed a framework (CMAS), proposing a prominent level of considerations of key factors that might affect a dataset. These factors need not be included in the dataset but be close enough to explain analysis outcomes.

The research method is DSR with case studies being used as a tool of thought demonstration. Main problems in systems design are simplifying and conceptualising ways to accurately mixing environment factors to data analysis. Classical Data mining tends to ignore changing real-life surrounding events (context-awareness), thereby not giving results that are easy to interpret. The local data (main dataset of the issue at hand) is usually affected by changing environmental factors whose effects need to be factored into any solution-building decision-making process.

In this section we related the study to questions raised at the beginning of thesis.

6.5.1 Question 1 Context and analysis results

Does Context-Aware Mining (CMAS) achieve increased domain relevance, improve quality, high accuracy, domain themes, and right relationships of outcomes to the real world?

The proposed CMAS approach views a data set as an element of a wide ecosystem of interactions. The interactions are defined in human professional or specialist terms. This way of looking at analysis activity encourages a localised tone and increase chances of high-quality insights. The three case studies reflect an improved understanding of medical, traffic and pollution situations. Insights within a defined social sphere are likely to apply terms and accurate assumptions on fields.

On the PIMA diabetes the CMAS investigated background expertise and applied it from step one Situation Understanding right to last step deployment. On the other hand, Indian Traffic data were split into context by applying CMAS to investigate the relationship of situation to human understanding of the physical world in question. The World Air and Water pollution data was not found on a single data source, but data had to be integrated to return a full perspective on subject pollution.

Framework starts with integrating the local dataset and context features, both being generated from the initial problem analysis. Research should contribute to improve Data Mining understanding and uptake by a wide range of professionals. Data mining in most settings involves data analysts and end users. Communication to bridge the knowledge gap between specialist, user, and existing data mining technology could be covered by the CMAS approach. It should start a discussion on whether the art of data mining is standardisable through the development of helper functions. Exploring and encouraging a focused, more real-life approach should improve the data analysis journey and the output quality.

CMAS achieves improvements on both the data mining process and insights quality. The design of the solution and the increase in the relevance of the outcomes is achieved through a holistic review of the complete approach, from understanding the challenge at hand to interpretation of any outcomes. A clear simplified solution enables proper use of

limited resources, be it processing capacity, human result interpretation ability, or processing ability associated with data mining methods and algorithms. Examples of methods with some limitations include decision tree and association rules which produce complex sets of outcomes unless the input is pruned. Context-Aware approach does focus the analysis and promote some pruning effort during most stages in the life cycle of these data projects.

6.5.2 Question two Context and Analysis scoping

Does Context-Aware approach optimize use of data mining resources. Does the approach achieve appropriate task scoping to balance results and data processing effort?

CMAS encourages the analyst to consider environmental factors linked to a data set. In these considerations boundaries are form and decisions on subject to investigate are made. The original PIMA original dataset included the population with or without diabetes, but early consideration dropped the non-diabetic ones. On the Indian Traffic data, clear definition of contexts opened opportunity to analyse them separately. One could analyse the effect of the road condition context on the cause of accident and the severity of the accident. To some extend CMAS does promote clear project scoping.

The data mining process has many challenges, including massive datasets, a large volume of dimensions, and data sparsity. Processing effort and insights that ignore contextual factors. The aim of this thesis is to reduce challenges that occur during insight development. Imminent difficulties arise from data mining logic and inputs that are too abstract instead of aligned with the surroundings and domain features.

The thesis uses a medical dataset around a single disease, diabetes based on PIMA Indian population. The second use case is based on the Indian Traffic Dataset. The third real-life use case is based on air and water pollution in the world. Improvements on

improvements (model quality, insights relevance, effort optimization) leveraged by Context-Aware concept and semi-automated method selection approach.

Context-aware, data mining, method/tools, scoping, data domain, context concept, resource optimisation, insight domain alignment, Application at multiple stages of data mining. CMAS achieved varying improvements per application cycle. The improvement includes improved insights relationship with real world; improved domain relevance, including tone easy relation to domain concepts. On the other hand, it also created clearly defined scoping of data mining projects, improved accuracy of insights and improved application to solution of insights. Improvements are subjective, but building analysis with domain and context considerations is likely to be anchored on a vital trimmed scope.

Data mining efforts have other limiting considerations of context, location when presenting sensitive output, and effects of privacy laws on both source datasets and output. Are they perspectives that are out of reach for data analysis due to political, ethical, or any other limitations. Looking at medical data, at what point does protection of privacy affect the relevance or quality of data analysis?

6.5.3 Question three Semi-Automated method selection

Can a semi-Automated Data Mining Method Selection script benefit novice Analyst and key stakeholders.

Most research effort has been done to select the most appropriate data mining algorithm for a dataset at hand. It also investigates the possibility of improving model quality by using stored tested knowledge to select the most appropriate data mining method. This research focuses on a level above this, that is the class of tools to be applied rather than the actual tool. Semi-Automated method selection effort. How this sits with other automations, suggestions systems. Maybe expand on AUTal which focuses on algorithms. Are there any lessons learnt? Also, any challenges and uptake rate %. Data

mining tool Selection Assistant is an aid the seeks to harvest on existing knowledge and software's improved abilities.

Knowledge on abilities Data Mining methods is pivotal to selection of appropriate methods. Popular methods being Deep learning, Association mining, Clustering, Regression, Classification, and ensembles. CMAS adds context to activities like Feature Selection, Feature Transformation, and data preparation analysis. Selection Assistant proposal attempts to increase insight relevance to the real-world influence, prune or focus effort and simplify the analysis.

There are design considerations when creating a system that selects the most appropriate data mining method for a given Data Mining challenge Depth of decisions and design considerations that need to be made during applications of the framework can vary with complexity of the situation.

In all the three use cases CMAS was used to integrate the context and semi-automation through real life test cases. The novelty of this approach is around incorporating Context on all six CRISP-MD stages. The role of human expertise, real life definitions, environment that changes forcing data to be looked at in varying ways and situations defined by combining features to create named states (Context) add tone to this approach. When well taught this approach can simplify data mining field.

It would be great to be able to produce analysis with context in mind and comparing this with that without context. Research should be able to measure the effect of the environment on the data mining results. Can results of a data mining assignment have environmental changes factored in or out?

6.6 Chapter summary

Chapter puts together the whole thesis looking at whether aims and objectives have been met. Results from application of the proposed CMAS approach to three use cases are analysed.

CHAPTER 7 Conclusion

7.1 Conclusion

This thesis developed Context-Aware Method and Algorithm Selecting (CMAS) framework to improve model design and outcomes based on the settings in which dataset is found. The novel CMAS proves to push analyst to investigate environment in which dataset are found. Models were created with domain terms and aligned perspective of the world. The result from the evaluation shows improved Data Mining process and quality of insights. Application of the approach for example, promoted efforts to understand the Indian Accident Severity setting including other factors to enrich insights from a data mining effort. Data analysis examines complex situations, dynamic approaches, and resource saving strategies. CMAS incorporates context at any stages of CRISP-MD like for Indian Accident Severity grouping of attributes and matching them to stakeholder interest to get enriched, stakeholder linked insights.

Thesis was illustrated the approach being applicable to medical, Traffic and pollution domains. A focused scoping and extended preprocessing of context data resulted in accurate and correctly subject linked insights. This thesis explored ways to make data analysis results more situation sensitive by incorporating contextual factors. Study has found Context-Aware approach turns to make resulting insights easier to interpret. Domain thematic, expert knowledge was found to increase the link between insights and the real-world challenge at hand.

Results indicated that Context-Awareness produce, CMAS framework defines situation and adjust all investigations based on this definition. Most Data Mining activities incorporate context factors. This research found that unguided analysis of data is resource wasteful. It established that Data Mining need to trim analysis to a narrow specific space that is easily linked to problem at hand. Context-Aware approach accept the fact that, all fields like medical, finance, or industry have ways and terms to address transactions and key activities in day to day. CMAS built based on domain terms, focus and ranking of issues. These terms are used for communication, so CMAS produce insights that can be understood by stakeholders. CMAS map scope to context freeing Analyst to decide on the appropriate level and breadth of investigations to be undertaken. The scoping aids in simplification of the process in line with requirements.

Seme-Automated Data Mining Method, Method Assistant Bubby developed in Python illustrated proof of concept by improving robustness in model design. Context-Aware expert knowledge based Semi-Automated Data Mining method selection improved model quality. We found that harnessing past expert knowledge simplifies model building. The resulting model is domain and situation sensitive. This thesis demonstrated through the Method Assist Buddy demonstrated that software can be used to recommend an appropriate Data Mining Method for a given dataset. The solution managed to make recommendation be based on methods capable of and situation under analysis. Thesis confirms that the knowledge processing ability of software technology may be used to resolve this selection problem.

A focused scoping and extended preprocessing of context data resulted in more complete data analysis. CMAS performance for each case study was reported based on the Data Mining technique applied like Association, Clustering, and Regression. Investigations pointed to insights easy to interpret as the Context-Aware approach design solution in terms of domain being analysed. Insights are also easy to translate into practical solution as the directly address a given situation.

Data mining projects are usually encountered with many constraints ranging from algorithm design, computing processing power down to human result interpretation limitations. A focused context-aware solution is most likely to be simple and easier to interpret. The results of the experiment confirm the increase in relevance and accuracy of the analysis output from analysis of complex and dynamic datasets.

Data mining effort have other limitations considerations of context, location when presenting sensitive output, effects of privacy laws on both source datasets and output. There are perspectives that are out of reach for data analysis due to political, ethical or any other limitations. Looking at medical data, at what point does protection of privacy affect the relevance or quality of data analysis.

CMAS and Method Selectors improved model design in the three domains used in the use cases. The improvement came from both expert knowledge and context considerations. Performance measures used were the standard one for any applied Data Mining Method. The logic is that nature of the problem at hand and dataset aid in the

selection of suitable data mining methods. The inclusion of Context makes the resultant insight easy to relate to the real world.

‘Context Data’ step of the CMAS framework was found to promote incorporation of specific situation factors to Data Mining. Data Integrator stage turns to marry data to environment thereby ensuring that the analysis is less abstract but more situation specific. This thesis demonstrated that experts view situations differently depending on their domain. Perspective of a situation could vary for Customers and Business Managers for the same dataset. CMAS promoted an increase analysis and transformation during solution design, which improves results and saves on the interpretation effort.

Thesis demonstrated three approaches to sourcing contextual factors accept expert knowledge, subset of local dataset or additional data sought before analysis. Thesis resolved the abstractness of data mining results by incorporating contextual factors.

7.2 Limitations

Context-Aware research presented a few limitations: -

- Scope can easily be too large. Data Mining involves stages; therefore, a novel approach may need to be inclusive of all stages.
- Metrics to measure performance of framework. Performance like easy of interpretation, domain relevance, linkage to physical world have not been traditionally measured. This poses a challenge in getting the most appropriate measuring units.
- The idea development had to borrow a path from other fields. The other fields with developed process do not fit well with Data Mining. Data Mining is a process that has no rigid structure.
- The research managed to develop a concept; it did not create a domain specific solution. That it illustrates the application of concepts but does not develop principles to be applied to multitude of settings.

7.3 Future directions

Research can in future apply concepts to the datasets and compare output with analysis that leaves out Context concepts. Future directions of research may include testing use of the Assist Buddy to select the DM method and then applying Automated Machine learning principles to match algorithms to datasets.

The research confirms the diversity of the environment in which the data are collected. The focus of a data analysis loosely called the local data is an extremely limited data space of the environmental data defining the situation, context data. The research focused on high-level environmental considerations. Future research could formulate the Context Data to be compared with parallel changing Local Data.

Research may need to be done with participants trained on the concept and those that a not forced to used CMAS, the compare outcomes.

7.4 Chapter summary

Chapters looks at Context-aware data mining principles, around their contribution and challenges that may be faced during implementation on use cases. It also reports back the application of the Method Selection Assistant software. The chapter concludes with the objectives, objectives, and contribution of knowledge of the thesis. An outline of the whole thesis from what the thesis aims to achieve to how this might be done. An outline of limitations was also provided.

A. References

Abdel-Basset, M., Mohamed, M., Smarandache, F., & Chang, V. (2018). Neutrosophic Association Rule Mining Algorithm for Big Data Analysis. *MDPI Symmetry, 10*(106).

Agrawal, R. (2024, 04 22). *know-the-best-evaluation-metrics-for-your-regression-model*. Retrieved from https://wanalyticsvidhya.com: https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model/

Akram, B. (2024, 4 20). *Understanding association rule mining*. Retrieved from Educative: https://www.educative.io/blog/understanding-association-rule-mining

Alla, S. (2024, 4 21). *Ml-evaluation-metrics-part-2*. Retrieved from Paperspace: https://blog.paperspace.com/ml-evaluation-metrics-part-2/

Al-Shargabi, A. A., & Siewe, F. (2020). A lightweight Association Rules Based Predication Algorithm (LWRCCAR) for Context-Aware System in IOT Ubiqitous, fog and Edge Computing Environment. *Proceedings of the Future Technologies (FTC) 2020. 2*, pp. 22-30. Advances in Intelligent Systems and Computing. Retrieved 02 25, 2023, from https://link.springer.com/chapter/10.1007/978-3-030-63089-8_2

Amir, F. u., Minhas, A., Asif, A., & Arif, M. (2016). CAFE-Map: Context Aware Feature Mapping for mining high dimensional biomedical data. *Comput. Biol. Medicine, 79*, 68--79. doi:10.1016/j.compbiomed.2016.10.006

Autnafa, B., & Kour, G. (2017). Survey on Analysis and Prediction of Road Traffic Accident Severity levels using Data Mining Technques in Maharashtra, India. *International Journal of current Engineering and Technology, 7*(6).

Avram, A., Matei, O., Pintea, C. M., & Anton, C. (2020). Innovative Platform for Designing Hybrid Collaborative and Context-Aware. *CoRR, abs/2007.13705*. Retrieved from https://arxiv.org/abs/2007.13705

Avram, A., Matei, O., Pintea, C. M., Pop, P. C., & Anton, C. A. (2019). Context-Aware Data Mining vs Classical Data Mining: Case Study on Predicting Soil Moisture. In *14th International Conference on Soft Computing Models in Industrial and Environmental Applications {(SOCO} 2019) - Seville, Spain, May 13-15, 2019, Proceedings* (pp. 199--208). Seville: Springer. doi:10.1007/978-3-030-20055-8_19

Avram, A., Matei, O., Pintea, C.-}., Pop, P. C., & Anton, C. A. (2020). How Noisy and Missing Context Influences Predictions in a Practical. In *15th International Conference on Soft Computing Models in Industrial and Environmental Applications, {SOCO} 2020, Burgos, Spain, 16-18* (Vol. 1268, pp. 22--32). Burgos: Springer. doi:10.1007/978-3-030-57802-2_3

Avram, A., Matei, O., Pintea, C.-M., & Anton, C. (2020). Innovative Platform for Designing Hybrid Collaborative and context-Aware Data Mining Scenarios. *Mathematics, 5*(8), 684.

Bezdek, J. C. (1973, January 01). Pattern recognition with fuzzy objetive function algorithms . Logan, Utah, USA.

Bhadane, C., & Shah, K. (2020, 9 4). Context-Aware next location prediction using Data Mining and Metaheuristics. *Evolutionary Intelligence, 14*, 871-880.

Bhadane, C., & Shah, K. (2021). Context-aware next location prediction using data mining and metaheuristics. *Evol. Intell., 14*(2), 871-880. doi:10.1007/s12065-020-00469-7

Brocke, J., Hevner, A., & Maedch, A. (2020). Introduction to Design Science Research. In J. Brocke, A. Henver, & A. Macdeh, *Design Science Research. Cases* (pp. 1-13). Springer Link.

Cabri, G., & Nocetti, G. (2024). A Context-Aware Application to Monitor the Air Quality. *12th EAI International Conference, ICCASA 2023*, 175-185.

Castells-Quntana, D., Dienesh, E., & Krause, M. (2021). Air population in urban world: A global view on density, cities and emissions. *Ecological Economics, 189*.

Chen, L., & Xia, M. (2021, February 1). A Context-Aware Recommendation approach based feature selection. *Applied Intelligence, 51*(2), pp. 865-875. Retrieved march 3, 2023, from https://dl.acm.org/doi/10.1007/s10489-020-01835-9

Delen, D., Walker, G., & Kadam, A. (2005). Predicting Breast Cancer Survability: A comparison of three data mining method. *Artificial Intelligence in Medicine, 34*(2), 113-127.

Delen, D., Walker, G., & Kadam, A. (2005). Predicting Breast Cancer Survivabilty: a Comparison of Three Data Mining Methods. *Artificial Intelligence in Medicine, 34*(2), 113-127.

Dey, A. (2000). *Providing Architectual Support for Building Context-Aware Application.* Georgia: Georgia College of Computing .

Dhaneshwar, S. S., & Patil, M. R. (2016). Context Aware in Data Mining Applications. A Survey. *International Journal of Science and Research (IJSR), 5*(1).

Dhaneshwar, S. S., & Patil, M. R. (2016). Context Awareness in Data Mining Application: A Survey. *International Journal of Science and Research (IJSR), 5*(1).

Ditcharoen, A., Aphivongpanyd, N., Chhour, B., Maneerat, K., Traikunwaranon, T., & Ammarapala, V. (2018). Road Traffic Accident Severity factors: A Review Papere. *2018 5th International Conference on Business and Industrial Research (ICBIR) Bangkok, Thailand.* Bangkok: IEEE.

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era Big Data - evolution, challenges and research agenda. *International Journal of Information Management, 48*, 63-71.

Duan, Y., Ong, V. K., Xu, M., & Mathews, B. (2012). Supporting decision making process with "ideal" software agents - "what do executives want?". *Expert Systems with Applications, 39*, 5534-5547.

Duo, A., Robinson, M. D., & Soneson, C. (2018). A systematic performance evaluation of clustering methods for sindle-cell RNA-seq data. *F1000 Research, 7*, 1141.

Escobar, M. O., Espinosa, R. L., Espinosa, J. M., Monroy, J. J., & Solar, G. V. (2019). Applying Process Mining to Support Management of Predictive Analytics Data mining projects in Decision Making Center. (pp. 1527-1533). The 2019 6th International Conference on Systems and Informatic.

Fatima, F., Talib, R., Muhammad, K. H., & Awais, M. (2020, November 18). A Paradigm-Shifting From Domain-Driven Data Mining Frameworks to Process-Based Domain-Driven Data Mining-Actionable Knowledge Discovery Framework. *IEEE Access*, pp. 210763 - 210774.

Fogil, D., & Guida, G. (2013). Knowledge-centered design of decision support for emergency management. *Decision Support System, 55*(1), 336-347.

Fung, U., Jianxin, L., Xueguan, L., Milan, A., & Zhaoquan, G. (2023, June). Robust Image Clustering Via Context-aware Constrative graph Learning. *Pattern Recognition*.

Gayati, N., Nickolas, S., Reddy, A. V., & Chitra, R. (2009). Performance Analysis of Data Mining Algorithms for Software Quality Prediction. (pp. 393-395). ART Com 2009 International Conference on Advances in Recent Technologies in Communication and Computing.

Guo, Y., Wang, K., Liu, S., Guo, L., & Lu, H. (2015). A Context-Aware Data Processing Model in Power Communication Network. Shenzhen: IEEE.

Gupta, P., & Gupta, V. (2012, 9 25). A Survey of Text Question Answering Techniques. *International Journal of Computer Applications, 53*(4), 1-8.

Gururaj, G. (2008). Road Traffic deaths, injuries and disabilities in India. *The National Medical Journal of India, 21*(1), pp. 14-20.

Haghighi, P. D., Zaslavsky, A., Wang, S. K., Gaber, M. M., & Loke, S. (2009). Context-Aware Adaption Data Stream Mining. *Intelligent Data Analysis Knowledge Discovery from Data Streams, 13*(3), 423-434.

Hollmann, N., Muller, S., & Hutter, F. (2023). Large Language Models for Automated Data Science: Introducing CAFFE for Context-Aware Automated Feature Engineering. *Annual Global Survey*.

Hotz, N. (2023, 01 19). Data Science Process Alliance. What is CRISP-DM.

Huang, C., Wang, Q., Yang, D., & Xu, F. (2018). Topic mining of tourist attractions based on a seasonal context aware {LDA} model. *Intell. Data Anal., 22*(2), 383--405. doi:10.3233/IDA-173364

Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2018). DMME: Data Mining Methodology for Engineering application - a holistic extension to the CRISP-DM model. *12th CIRP Conference on Intelligent Computation in Manufacturing Engineering* (pp. 403-408). Gulf of Naples: Elsevier B V.

Jakhar, D., & Kaur, I. (2020, 01). Artificial itelligence, machine learning and deep learning: definitions and differences. *Clinical and Experimental Dermatology (CED), 45*(1), 131-132. Retrieved from https://doi.org/10.1111/ced.14029

Jenkins, A. M. (2000). *Research Methodology and MIS research.* Semstic Scholar.

Jia, Y., Zhang, J., & Huan, J. (2011). An effient graph-mining method for complicated and noisy data with Real-World applications. *Knowledge and Information Systems*, 423--447.

Kaggle . (2024, 4 19). *Kaggle.com*. Retrieved from www.kaggle.com: https://www.kaggle.com/datasets/rachit239/road-accident-data-2020-india/data

Khan, S., & Shaheen, M. (2021). From Data Mining to Wisdom Mining. *Journal Of Information Science*.

Khanbabaer, M., Alborzi, M., Sobhanl, F. M., & Rafar, R. (2019). Apply Clustering and Classification data mining techniques for competetive and knowledge improvement. *Knowledge Processes Management, 26*, 123-139.

Khatun, R. (2017). Water Pollution: Causes, consequence, prevention method and role of WBPHED with special reference from Murshidabed district. *International journal of Scientific and Research Publication, 7*(8).

Kim, Y., & Chung, M. (2016, 11 23). Unstructured Data Service Model Utilizing Context-Aware Big Data Analysis. *Advances in Computer Science and Ubiquitous Computing*, 926 - 931. Retrieved 03 1, 2023

Konar, A. (2000). *Artificial Intelligence and Soft Computing: behavioral and Cognitive modeling of the human brain.* Florida: CRC Press.

Lakshmi, K. R., Kumar, S. P., & Krishma, M. V. (2013). Performance Comparison of Data Mining technques for Prediction of Breast Cancer Disease Survivability. *Asian Journal of Computer Science and Information Technology, 3*(5), 81-83.

Lange, M. (2017). Time series data mining for context-aware event analysis. L{\"{u}}beck, Germany. Retrieved 12 18, 2021, from http://www.zhb.uni-luebeck.de/epubs/ediss1873.pdf

Lin, X., Coller, R., & O'Hare, G. M. (2017, October). A Survey of Clustering Techniques in WSNs and consideration of the challenges of appling such to 5G IoT scrnarios. *IEEE Internet of Thing Journal, 4*(5), 1229-1249. doi:https://doi.org/10.1109/JIOT.2017.2726014

Liu, X., Wang, G., & Bhuiyan, M. A. (2021). Personalised context-aware re-ranking in recommender system. *Connection Science*. doi:10.1080/09540091.2021.1997915

Lu, Q., & Sheng, B. (2013). A Weighted Association Rules Mining Algorithm with Fuzzy Quantitative Constraints. *2013 International Conference on Information Science and Cloud Computing Companion*, 481-487.

Majid, A., Chen, L., Chen, G., Mirza, H. T., Hussain, I., & Woodward, J. (2013). A context-aware personalized travel recommendation system based on geotagged social media data mining. *Int. J. Geogr. Inf. Sci., 27*(4), 662--684. doi:10.1080/13658816.2012.696649

Malik, N., Mahmud, U., & Javed, Y. (2007). Future Challenge In Context-Aware Computing. *IADIS International Conference*.

Mao, X., Mitra, S., & Swaminathan, V. (2017). Feature Selection for FM-Based Context-aware Recommendation System. *2017 IEE International Symposium on Multimedia (ISM).* Taichung: IEE. Retrieved 6 4, 2021

Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I., & de Mendonca, A. (2011). Data Mining Methods in the prediction of Dementia: A real-data comparison of the accuracy sensibility, speciciaficity of linear discriminant analysis, logistic regression, neural network, support vector, classification trees and random foressts. *BMC Research Note*(4).

Martinez-Plumed, F., Contresus-Ochando, L., Ferri, C., Flach, P., Hernandez-Orallo, J., Kull, M., . . . Ramirez-Quintana, M. J. (2019). CASP-DM : Context-Aware Standard Process for Data Mining. San Fransisco: Deep AL.

Martinez-Plumed, F., Ochando, L. C., Ferri, C., Flach, P. A., Hernandez-Orallo, J., Kull, M., . . . Ramirez-Quintana, M. J. (2017). {CASP-DM:} Context Aware Standard Process for Data Mining. *CoRR, abs/1709.09003*. Retrieved 10 15, 2020, from http://arxiv.org/abs/1709.09003

Matei, O., Rusu, T., Bozga, A., Sitar, P. P., & Anton, C. (2017). Context-Aware Data Mining: Embedding External Data Sources in a Machine Learning Process. In F. Javier, *Hybrid Artificial Intelligent Systems - 12th International Conference, {HAIS} 2017, La Rioja, Spain, June 21-23, 2017, Proceedings* (Vol. 10334, pp. 415--426). La Rioja, Spain: Springer. doi:{10.1007/978-3-319-59650-1_35

Mijnsbrugge, D. V., Ongenae, F., & Van Hoecke, S. (2023). Context-aware deep learning with dynamically assembled weight matrices. *Information Fusion, 100*.

Mohan, K. K., Mariamichael, A., & Kishore, K. (2011). Community Specific BMI cutoff point for South India females. *Journal Obesity*, 8.

Moyle, S., & Jorge, A. (2001). RAMSYS-A Methodology for supporting rapid remote collaborative data mining project. *PKDD01 Workshop: Itergrating Aspects of Data Mining*.

Nallaperuma, D., De Silva, D., Alahakoon, D., & Yu, X. (2017). A cognitive data stream mining technique for context-aware IoT systems. In *{IECON} 2017 - 43rd Annual Conference of the {IEEE} Industrial Electronics Society, Beijing, China, October 29 - November 1, 2017* (pp. 4777--4782). Beijing, China: IEEE. doi:10.1109/IECON.2017.8216824

Nasor, M., & Ali, S. (2019). Performance Analysis and Ranking of Data Mining Algorithms Across Multiple Datasets. IEEE 19th International Symposium on Signal Processing and Information Technology.

Onim, S. H., Thapliyal, H., & Rhodus, E. k. (2024). Utilizing Machine Learning for Context-Aware Digital Biomarker of Stress in Older Adults. *Information, 15*(5), 274. Retrieved from https://doi.org/10.3390/info15050274

Osei-Bryson, K.-M. (2012). A Context-aware data mining process model based framework for supporting evaluation of data mining results. *Expert Systems with Application, 39*(1), 1156-1164.

Pechenizkiy, M., Pauronea, S., & Tsymbal, A. (2008). Does Relevance Matter to Data Mining Research? In T. Y. Lin, Y. Xie, A. Wasilewska, & C. J. Liam, *Data Mining : Foundation and Practice, Studies in Computational Intelligence* (pp. 251-275). Berlin: Springer.

Pei, L., Vidyaratne, L., Rahman, M. M., & Iftekharuddin, K. M. (2020). *Context aware deep learning for brain tumor segmentation, subtype classification, and survival prediction using radiology images.* Scientfic Reports.

Popat, S. K., & Emmanueal, M. (2014). Review and Comparative Study of Clustering Techniques. *International Journal of Computer Science and Technologies, 5*(1).

Pourghasemi, H. R., Yousefi, S., Kornejady, A., & Cerda, A. (2017, December 31). Performance qssessment of individual and enseble data mining techique for gully erosion modeling. *Science of the Total Environment, 609*, 764-775. doi:https://doi.org/10.1016/j.scitotenv.2017.07.198

Ramezani, M., Moradi, P., & Tab, F. A. (2013). Improve Performance of Collaborative Filtering System Using Backwards feature Selection. Shiraz: IEEE.

Rauch, J. (2005). Logic of Association Rules. *Applied Intelligence, 22*, 9-28.

regression-metrics. (2024, 4 22). Retrieved from https://www.geeksforgeeks.org: https://www.geeksforgeeks.org/regression-metrics/

Remeseiro, B., & Bolon-Canedo, V. (2019, September). A Review of feature selection methods in Medical Applications. *Computers in Biology and Medicine, 112*. doi:https://doi.org/10.1016/j.compbiomed.2019.103375

Represa, S. N., Fern'andeza-Sarri'a, A., Porta, A., & Palomer-Vazquez, J. (2019, November 27). Data Mining Paradigm in the study of Air Quality. *Environmental Process, 7*, 1-21.

Saleen, B., & Masseglia, F. (2011). Discovering Frequent Behaviors : Time is an essential element of context. *Knowledge Information System* , 311-331.

Santoso, M. H. (2021, November). Application os Association Rule Method using Apriori Algorithm to find Sales Patterns Case study of Indomaret Tanjung Anon. *Brilliance Research of Artificial Inteligence, 1*(2), 1-13.

Sarker, I. H. (2021). Data Science and Analytics : An Overview from Data-Driven Smart Computing, Decision-Making and Application Perspective. *SN Comput Sci, 2*(5), 377. Retrieved 01 23, 2021

Sarker, I. H. (2021). Machine Learning : Algorithm, Real-World Applications and Research Directions. *SN Comput Sci, 2*(3), 160. Retrieved 01 23, 2021

Schnor, N. (2024, 05 07). *Pima-Indians-Diabetes-Dataset/blob/master/README.md*. Retrieved from Github: https://github.com/npradaschnor/Pima-Indians-Diabetes-Dataset/blob/master/README.md

Scholze, S., & Barata, J. (2017). Context Awareness for Flexible Manufactoring Systems Using Cyber Physical Approaches. *7Th Doctoral Conference on Computing, Electrical and Industrial Systems (doCEIS)*, 107-115.

Serban, F., Vanschorn, J., Kietz, J.-u., & Bernstein, A. (2013). A Survey of Intelligent Assistant for Data Analysis. *ACM Company Surveys, 45*(3), 1-35.

Shafique, U., & Qaiser, H. (2014). A Comperative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research, 12*(1), 217-222.

Sheikhpour, R., Sarram, M. A., Gharaghani, S., & Chahooki, M. A. (2017). A Survey on Semi-Supervised feature Selection Methods. *Pattern Recognition, 64*, 141-158.

Sherif, M., & Alesheikh, A. A. (2018). Context-Aware Movement Analysis : Implications, Taxonomy and Design Framework. *Wire : Data Mining and Knowledge Discovery, 8*(1).

Silva, C., Saraee, M., & Saraee, M. (2019). Data Science in public mental health: a new analytic framework. *IEEE Symposium on Computers and Communications*. Barcelona, Spain. Retrieved 01 01, 2021, from http://usir.salford.ac.uk/id/eprint/51688

Singh, S., Vajirkar, P., & Lee, Y. (2003). Context-Based Data Mining Using Ontologies. *22nd International Conference on Conceptual Modeling* (pp. 405--418). Chicago: Lecture Notes In Computer Science. Retrieved 01 05, 2022, from https://link.springer.com/chapter/10.1007/978-3-540-39648-2_32

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *In Proceedings of the Symposium on Computer Applications and Medical Care*, 261,265. Retrieved july 20, 2022, from https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database or http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

Sridevi, K. N., & Prakasha, S. (2021). Analysis Comparison of Various Clustering Classifier Algorithms. Madurai: IEEE.

Sridevi, K. N., & Prakasha, S. (2021). Comparative Study on Various Clustering Algorithm Review. (pp. 153-158). Madurai: IEEE.

Standl, E. (2021). *Global Statistics on Diabetes.* Sophia Antipolis, Brussels: European Society os Cardiology (ESC).

Suppa, P., & Zimeo, E. (2016). A Context-Aware Mashup Recommender Based on Social Networks Data Mining and User Activities. In *2016 {IEEE} International Conference on Smart Computing, {SMARTCOMP} 2016, St Louis, MO, USA, May 18-20, 2016* (pp. 1--6). St Louis: {IEEE} Computer Society. doi:10.1109/SMARTCOMP.2016.7501672

Suragala, A., Venkateswaralu, P., & China Raju, M. (2020). A comparative Study of Performance metric of Data Mining Algorithm on Medical Data. Hyderabad: Lecture Notes in Electrical Engineering.

Temenos, A., Temenos, N., Tzorrtzis, I. N., Rallis, I., Doulamic , A., & Doulamis, N. (2024, 4). C2A-DC: A context-aware adaptive data cube framework for environmental monitoring and climate change crisis management. *Remote Sensing Applications: Society and Environment, 34*.

Thabet, D., Ganouni, N., Ghannouchi, S. A., & Hajjami, H. (2019). Towards Context-Aware Business Process Cost Data Analysis Including the Control-Flow Perspective - {A} Process Mining-Based Approach. In A. Abraham, P. Siarry, & K. Ma, *Intelligent Systems Design and Applications - 19th International Conference on Intelligent Systems Design and Applications {(ISDA} 2019), Auburn,WA, USA, December 3-5, 2019* (Vol. 1181, pp. 193--204). Auburn: Springer. doi:10.1007/978-3-030-49342-4_19

Think Insights. (2023, Match 9). *CRISP-DM – A Framework For Data Mining And Analysis.* Retrieved from thinkinsights.net: https://thinkinsights.net/data/crisp-dm/

Treffinger, D. J., Selby, E. C., & Isaksen, S. G. (2008). Understanding individual problem solving style: A key to learning and applying creative problem solving. *Learing and Inividual Differences, 18*(4), 390-401.

Tuget, V. T., Binh, N. T., Quoc, N. K., & Khare, A. (2021). Content Based Medical Image Retrieval on Salient Regions Combined with Deep Learning. *Mobile Networks and Application*, 1300 -1310.

UK Essays. (2022, December 12). *Reseach Onion Explanation of Concept.* Retrieved from UKEssays.com: https://www.ukessays.com/essays/psychology/explanation-of-the-concept-of-research-onion-psychology-essay.php#:~:text=The%20research%20onion%20was%20developed%20by%20Saunders%20et,of%20the%20research%20process%20%28Saunders%20et%20al.%2C%202007%29.

Van Amasterdam, B., Jackson, M. J., & Stoyanov, D. (2021). Gesture Recognition In Robotic Surgerey : A Review. *IEEE Transaction on Biomedical Engineering, 68*(6), 2021,2035.

Van Houdt, G. (2020). Mining Behavioural Patterns from Event Data to Enable Context-Aware. In C. D. Ciccio (Ed.), *Proceedings of the {ICPM} Doctoral Consortium and Tool Demonstration* (Vol. 2703, pp. 9--10). Padua, Italy: CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-2703/paperDC5.pdf

Viktoratos, L., Tsadiras, A., & Bassiliades, N. (2018). Combining community=based knowledge with association rule mining to alleviate the cold start problem in context-aware recommender systems. *Expert Systems with Application. 101*, pp. 78-90. Elsevier.

Vinh, P. C., Anti, L. N., & Siricharoen, W. V. (2017). Context-Based Project Management. *Context-Aware System and Application*, 12-21.

Wachowicz, M., & Bogorny, V. (2009). A Framework for Context-Aware Trajectory. In L. Cao, P. S. Yu, C. Zhang, & H. Zhang, *Data Mining for Business Application* (pp. 237-257). New York: Springer.

Wang, H., & Liu, X. (2011). The research of Improved Association Rules mining Apriori algorithm. *2011 Eight International Conference on fuzzy systems and Knowledge Discovery (FSKD), 2*, 961-964.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard Process Model for Data Mining. In *Proceedings of fouth international conference on the practical application of knowledge discovery and data mining* (pp. 29-39). Pennsylvania: The Pennsylvania State University.

World Health Organisation. (2024). WHO Ambient Air Quality Database (updated 2024). Geneva: WHO.

Wrede, J. C., Visa, D. K., Alerge, U., & Aztriria, A. (2018). A Survey on the Evolution of the Notion of Context-Awaness. *Applied Artficial Intelligence*, 1-22.

Yau, J. y., Dickert, S., & Joy, M. (2010). A mobile context-aware fromewark for managing learning schedules - Data Analystis from Dairy Study. *Educational Technology & Society, 13*(3), 22-32. Retrieved 03 03, 24

Yuan, J., & Wu, Y. (2008). Context-Aware Clustering. *Computer Vision and Pattern Recognition, 2008, CVPR 2008, IEEE.*

Zschech, P., Horn, R., Janiesch, C., & Heinrich, K. (2020). Intelligent User Assistance for Automated Data Mining Method Selection. *Business Information System, 62*(3), 227-247.