



REPUBLIC OF BENIN
MINISTRY OF HIGHER EDUCATION
AND SCIENTIFIC RESEARCH OF BENIN

ABOMEY-CALAVI UNIVERSITY

**INSTITUTE OF TRAINING AND
RESEARCH IN COMPUTER SCIENCE**

BP 526 Cotonou Tel : +229 21 14 19 88
<http://www.ifri-uac.net> Courriel : contact@ifri.uac.bj



THESIS

for obtaining a

Master's degree in computer science

Option : Software Engineering

Presented by :

Linuse TIKPON

Interpreting complex machine learning models using association rules

Under supervision of :

Dr (MA) Vinasetan Ratheil **HOUNDJI**

Dr (MC) Arnaud **AHOUANDJINOU**

Examination comitee :

HOUNGUE Pélagie Dr (MC) IFRI President of the Jury

KONNON Abel Dr (MC) IFRI Examiner

HOUNDJI Ratheil Dr (MA) IFRI Reporter

Année Académique : **2023-2024**

Summary

Dedication	iii
Acknowledgements	iv
Résumé	v
Abstract	v
List of figures	vii
List of tables	viii
List of Algorithms	ix
Acronyms and abbreviations	x
Introduction	1
1 Literature review	5
Introduction	5
1.1 Introduction to the literature review	5
1.2 Complex machine learning models	5
1.3 Challenges in interpreting complex models	7
1.4 Model interpretation techniques	8
1.5 Limitations and challenges of current interpretation methods	12
1.6 Association rules [1, 2, 3]	14
2 Material and methods	19
Introduction	19
2.1 Data preparation	20
2.2 Development of the rule extraction algorithm.	28
2.3 Implementation and tools section	34
3 Results and discussion	40
Introduction	40
3.1 Presentation of results	40

3.2 Implementation details	50
Conclusion	51
3.3 Critical analysis of results	51
3.4 Comparison with other interpretation techniques	55
Conclusion	57
Bibliography	59
Webography	60
Contents	63

Dedication

I dedicate this thesis to Almighty God for His grace and guidance throughout my academic journey.

To my parents, for their unwavering support, love, and constant encouragement. You have always been a source of inspiration and motivation.

To my brothers and sisters, for their presence and moral support during all these years.

To my friends, who have accompanied me with joy and positivity, and who helped me stay strong during the challenging moments.

To my classmates, with whom I shared unforgettable moments, challenges, and successes. Your company made this academic experience even richer.

Acknowledgements

I would like to express my deepest gratitude to everyone who contributed to the successful completion of this thesis.

First and foremost, I thank Almighty God for His infinite grace, guidance, and protection throughout my academic journey. Through His divine providence, I have been able to overcome every obstacle and reach this milestone.

I extend my sincere appreciation to the INSTITUTE OF TRAINING AND RESEARCH IN COMPUTER SCIENCE (Institut de Formation et de Recherche en Informatique : IFRI) for providing an excellent academic environment and the resources necessary for my research work.

I am deeply grateful to all the professors at IFRI who have contributed to my academic growth through their dedication to teaching and research excellence.

I would like to express my special gratitude to my thesis supervisor, Dr. Ratheil HOUNDTJI, whose expertise, patience, and unwavering support have been instrumental in the completion of this work. Your thoughtful feedback, consistent encouragement, and readiness to provide guidance whenever I encountered challenges have been invaluable.

I am particularly grateful to Dr. Othniel SODJI, medical doctor, for their invaluable expertise and time spent validating my research results. Your medical insights and verification of my findings have significantly enhanced the credibility and reliability of this work.

My heartfelt thanks go to:

- My parents, for their unwavering financial and emotional support throughout my academic journey. Your belief in me has been a constant source of strength.
- My brothers and sisters, for their constant encouragement and support at every step of this journey.
- My friends and classmates, for their sincere friendship, moral support, enriching discussions, and shared moments of mutual encouragement.

Finally, I would like to acknowledge everyone who has contributed directly or indirectly to the successful completion of this thesis.

Abstract

This thesis explores an interpretation approach for complex machine learning models using association rule extraction. The main objective is to make the decisions made by these models, often referred to as "black boxes," more transparent by identifying significant patterns in the data that lead to specific predictions. This way, we can better understand which input features most influence the outputs of a predictive model. For our work, we chose a diabetes prediction model based on patient information from our dataset. By applying the Apriori algorithm to the input data and the predictive outcomes of the model, we extracted association rules that reveal the combinations of features significantly impacting the predictions. These rules allow for the detection of recurring patterns in medical data, such as glucose levels, BMI, and blood pressure, and help us understand how these characteristics influence the risk of diabetes. At the end of the analysis, a report is generated to interpret the extracted rules and provide a clear overview of the most influential variables. This approach enhances the interpretation of complex models and facilitates their use in decision-making contexts.

Keywords: Machine Learning, Model Interpretability, Association Rules, Diabetes Risk, Apriori Algorithm, Health.

Résumé

Ce mémoire explore une approche d'interprétation des modèles complexes de machine learning en utilisant l'extraction de règles d'association. L'objectif principal est de rendre plus transparentes les décisions prises par ces modèles, souvent qualifiés de "boîtes noires", en identifiant des motifs significatifs dans les données qui conduisent à des prédictions spécifiques et donc de mieux comprendre quelles caractéristiques d'entrée influencent le plus les sorties d'un modèle prédictif. Pour notre travail, nous avons choisi un modèle de prédiction du diabète pour certains patients en ayant des informations de ces patients dans notre dataset. En appliquant l'algorithme Apriori sur les données d'entrée et les résultats prédictifs du modèle, nous avons extrait des règles d'association qui révèlent les combinaisons de features ayant un impact significatif sur les prédictions. Ces règles permettent de détecter des patterns récurrents dans les données médicales, tels que le taux de glucose, l'IMC et la pression artérielle, et de comprendre comment ces caractéristiques influencent le risque de diabète. À la fin de l'analyse, un rapport est généré pour interpréter les règles extraites et fournir une vue claire des variables les plus influentes. Cette approche permet d'améliorer l'interprétation des modèles complexes et de faciliter leur utilisation dans des contextes décisionnels.

Mots-clés : Machine Learning, Interprétabilité des modèles, Règles d'association, Risque de diabète, Algorithme Apriori, Santé.

List of figures

1.1	SHAP summary plot showing feature contributions.	9
1.2	LIME explanation of a specific prediction.	10
2.1	Detailed methodology workflow	20
2.2	Dataset used as input	22
2.3	Discretization results	25
2.4	Encoding result	27
3.1	Heatmap showing support and confidence for top 10 association rules	43
3.2	Bar plot comparing support, confidence, and lift for the top 5 rules	44
3.3	Scatter plot of confidence vs. lift for extracted rules	44
3.4	Graph network visualization of key association rules	45
3.5	Capture of the PDF file generated to interpret each rule	48

List of tables

1.1	Summary of advantages and disadvantages of DNNs, Random Forests, and GBMs. . .	7
1.2	Comparison of model interpretation techniques	11
1.3	Comparison of model interpretation techniques	11
1.4	Summary of limitations and challenges of interpretation techniques	13
2.1	Summary Statistics of the Dataset Variables	21
2.2	Specific Analysis of Key Features	21
2.3	Feature Correlation with Diabetes	22
2.4	Overview of the dataset	23
2.5	Summary of data discretization	25
2.6	Technical alternatives and choice justifications	37
2.7	Statistical metrics for rule evaluation	38
2.8	Process of expert validation	38
2.9	Comparison of association rules with SHAP and LIME	39
3.1	Key metrics for evaluating association rules	41
3.2	Key patterns from extracted rules	42
3.3	Examples of extracted rules and their metrics	42
3.4	Tabular representation of key association rules	45
3.5	Summary of key rules and their metrics	46
3.6	Clinical relevance of key rules	46
3.7	Patient risk stratification	47
3.8	Unexpected patterns in extracted rules	47
3.9	Comparative Analysis of Interpretation Techniques	55

List of Algorithms

1	Step-by-step workflow for rules extraction from medical data (Part 1)	50
2	Step-by-step workflow for rules extraction from medical data (Part 2)	51

Acronyms and abbreviations

AI :

Artificial Intelligence [5](#), [10](#), [63](#)

API :

Application Programming Interface [33](#), [51](#)

BMI :

Body Mass Index [v](#), [19](#), [23–26](#), [28–30](#), [32](#), [39](#), [41–43](#), [45–47](#), [50–54](#)

CSV :

Comma-Separated Values [34](#)

DNNs :

Deep Neural Networks [viii](#), [1](#), [5–7](#), [63](#)

FP-Growth :

Frequent Pattern Growth [15–17](#), [26](#), [54](#), [64](#)

FP-tree :

Frequent Pattern Tree [16](#), [17](#), [64](#)

GBMs :

Gradient Boosting Machine [viii](#), [6–8](#), [63](#)

GPT :

Generative Pre-trained Transformer [33](#), [51](#), [57](#)

ICE :

Individual Conditional Expectations [55](#)

IDE :

Integrated Development Environment [34](#)

IMC :

Indice de Masse Corporelle (Body Mass Index - BMI) [vi](#)

KNN :

K-Nearest Neighbors [24](#)

LIME :

Local Interpretable Model-Agnostic Explanations [vii](#), [viii](#), [5](#), [7–13](#), [18](#), [38–40](#), [55](#), [57](#), [64](#)

PDF :

Portable Document Format [vii](#), [33](#), [48](#), [51](#), [57](#)

PDPs :

Partial Dependence Plots [8](#), [10–13](#), [55](#), [64](#)

SHAP :

SHapley Additive exPlanations [vii](#), [viii](#), [5](#), [7–9](#), [11–13](#), [18](#), [38–40](#), [55](#), [57](#), [64](#)

SQL :

Structured Query Language [34](#)

WHO :

World Health Organization [22](#)

Introduction

Context and importance

General presentation of the topic

The advent of machine learning has revolutionized numerous fields by enabling the analysis and interpretation of complex data sets, leading to advancements in areas such as image recognition, natural language processing, and predictive analytics. Among the various machine learning techniques, complex models like deep neural networks and ensemble methods (e.g., random forests and gradient boosting machines) have shown exceptional performance due to their ability to capture intricate patterns and relationships in data.

Deep neural networks (DNNs), inspired by the human brain's structure, consist of multiple layers of interconnected neurons that can model highly non-linear functions. These networks have been successfully applied in tasks such as object detection, speech recognition, and autonomous driving, often surpassing human-level performance. Similarly, ensemble methods, which combine the predictions of multiple base learners to improve generalization, have become the cornerstone of many machine learning competitions and real-world applications.

Despite their success, the increasing complexity of these models has led to a significant challenge: interpretability. As models grow more intricate, understanding how they make decisions becomes more difficult. This lack of transparency, often referred to as the "black box" problem, raises concerns about the reliability and trustworthiness of model predictions, especially in high-stakes domains.

Importance of interpreting complex machine learning models

The importance of interpreting complex machine learning models cannot be overstated, particularly in critical fields such as healthcare, finance, and legal systems, where decisions based on model predictions can have profound consequences.

In healthcare, for example, predictive models are used to diagnose diseases, recommend treatments, and predict patient outcomes. A model that suggests a specific treatment plan must be interpretable to ensure that medical professionals can trust and understand the basis of its recommendations. This transparency is crucial for gaining the acceptance of clinicians and patients, and for ensuring that decisions are made based on sound medical principles rather than obscure algorithmic processes.

In the financial sector, machine learning models are employed for credit scoring, fraud detection,

and investment strategies. Regulatory requirements often mandate that financial institutions provide clear explanations for decisions affecting customers, such as loan approvals or rejections. An interpretable model helps institutions comply with these regulations and maintain the trust of their clients by providing transparent and justifiable decisions.

Legal systems are another domain where interpretability is critical. Algorithms used in predictive policing, judicial decision-making, or risk assessment must be transparent to ensure fairness and accountability. Understanding the factors influencing these models' predictions is essential for preventing bias and ensuring that decisions uphold legal and ethical standards.

Beyond these fields, interpretability is vital for improving model development and deployment. It allows data scientists and engineers to debug and refine their models more effectively, ensuring that models not only perform well on validation metrics but also align with domain knowledge and ethical considerations. Interpretability can also foster innovation by revealing new insights and patterns within the data that were previously unnoticed.

Problem statement and research motivations

The core problem addressed in this research is the lack of interpretability in complex machine learning models. While these models offer high accuracy and robust performance, their decision-making processes are often opaque, making it challenging to understand, trust, and act upon their predictions. This opacity poses significant risks, particularly in sensitive applications where erroneous or biased decisions can lead to severe consequences.

Traditional interpretability techniques, such as feature importance measures and partial dependence plots, provide some insights but are limited in their ability to fully elucidate the decision-making process of complex models. These techniques typically offer a global perspective on model behavior but fall short when it comes to explaining individual predictions.

To address this gap, this research explores the use of association rules for interpreting complex machine learning models. Association rules, initially used to discover interesting relationships between variables in large databases, can be applied to machine learning models to extract explicit, human-readable rules that link input features to predictions. By generating representative datasets from pre-trained models and applying rule extraction algorithms, this approach aims to uncover hidden relationships and dependencies that drive model decisions.

The motivations for this research are threefold :

- **Enhancing trust and transparency** : By making complex models more interpretable, this research aims to build trust among stakeholders, including clinicians, financial analysts, and legal professionals, by providing clear explanations for model predictions.
- **Improving model performance and reliability** : Interpretability can help identify and mitigate potential biases, errors, and inconsistencies within models, leading to more reliable and ethical decision-making processes.
- **Fostering Innovation and Insights** : Interpretable models can reveal novel insights and patterns within data, driving further innovation and understanding in various fields of application.

In summary, this research addresses the critical need for interpretability in complex machine learning models by leveraging association rules to provide transparent and actionable insights into model behavior. By doing so, it aims to bridge the gap between model performance and interpretability, promoting the responsible and informed use of machine learning technologies.

Objectives and research questions

Main and specific objectives of the study

The overarching objective of this research is to enhance the interpretability of complex machine learning models using association rules. By extracting human-readable rules that elucidate the decision-making processes of these models, we aim to make their predictions more transparent, trustworthy, and actionable. This main objective can be broken down into several specific goals :

- **Objective 1 : Understand complex models**

To gain a deep understanding of how complex machine learning models, such as deep neural networks and ensemble methods, make predictions. This involves analyzing the internal workings and decision pathways of these models.

- **Objective 2 : Develop a rule extraction methodology**

To develop a robust methodology for extracting association rules from complex machine learning models. This includes selecting appropriate datasets, pre-processing data, and applying rule extraction algorithms.

- **Objective 3 : Evaluate extracted rule quality**

To evaluate the quality and usefulness of the extracted rules. This involves assessing the interpretability, accuracy, and relevance of the rules in explaining model behavior.

- **Objective 4 : Validate by applying in a specific domain**

To validate the extracted rules in specific application domains, such as healthcare. This ensures that the rules are not only theoretically sound but also practically relevant and beneficial.

- **Objective 5 : Enhance model performance**

To use the insights gained from the extracted rules to refine and improve the performance of complex models. This includes identifying potential biases, errors, and areas for model improvement.

Research questions

To achieve the above objectives, this study seeks to address the following research questions :

- **RQ1 : How can we effectively extract association rules from complex machine learning models?**

This question focuses on the techniques and methodologies that can be employed to derive meaningful rules from models like deep neural networks and ensemble methods. It examines the feasibility and effectiveness of various rule extraction algorithms.

- **RQ2 : What is the impact of rule-based interpretability on model transparency and trust?**

This question investigates how the use of association rules affects the perceived transparency and trustworthiness of complex models among different stakeholders, including domain experts and end-users.

- **RQ3 : How do the extracted rules enhance our understanding of model behavior?**

This question explores the extent to which the extracted rules provide insights into the decision-making processes of complex models. It looks at the clarity and depth of understanding that these rules offer.

- **RQ4 : In what ways can the extracted rules be validated in real-world applications?**

This question addresses the practical validation of the extracted rules in specific domains such as healthcare and finance. It examines the relevance, accuracy, and utility of the rules in real-world scenarios.

— **RQ5 : How can the insights from association rules be used to improve model performance?**

This question considers how the knowledge gained from the extracted rules can be applied to refine and enhance the performance of complex machine learning models. It looks at potential improvements in accuracy, fairness, and reliability.

By addressing these research questions, this study aims to bridge the gap between model complexity and interpretability, providing a framework that makes advanced machine learning techniques more accessible and understandable. The findings are expected to contribute to the broader field of machine learning by offering new methodologies for interpreting and refining complex models, ultimately promoting their responsible and effective use in various critical applications.

Organization of the thesis

This thesis is structured into three main chapters. **Chapter 1** provides a literature review, offering an overview of complex machine learning models such as deep neural networks and ensemble methods, and discussing the challenges of interpretability, association rules, and related work. **Chapter 2** describes the methodology, including the generation of representative datasets from pre-trained models, the extraction of association rules, the evaluation metrics used to assess their quality, and the experimental setup. **Chapter 3** presents and analyzes the results, focusing on the extracted rules, their implications for model decision-making, and domain-specific validation, concluding with a discussion on the broader impact of the findings on machine learning interpretability.

Literature review

1.1 Introduction to the literature review

The rise of complex machine learning models has transformed various fields by enabling the analysis of intricate patterns in large datasets. Despite their predictive power, these models often function as "black boxes," making their decision-making processes difficult to interpret. This lack of transparency is particularly problematic in critical areas like healthcare and finance, where understanding model predictions is essential for trust and accountability.

This literature review explores three key aspects of model interpretability. First, it reviews different machine learning models, such as Deep Neural Networks and Random Forests, highlighting their advantages and challenges in terms of interpretability. Second, it examines techniques like [SHAP](#) and [LIME](#) that aim to improve model transparency, discussing their effectiveness and limitations. Finally, it introduces association rules as a method to extract interpretable patterns from data and how they can be applied to interpret complex models.

By analyzing these areas, the review identifies research gaps and sets the stage for the methodology and experiments that follow in this thesis, contributing to efforts to make machine learning models more transparent and trustworthy.

1.2 Complex machine learning models

1.2.1 Types of complex [AI](#) models

The rapid advancement of machine learning has led to the development of various complex models that excel in capturing intricate patterns in data. These models include Deep Neural Networks, Random Forests, Gradient Boosting Machines, and support vector machines, among others. Each type of model has its unique architecture and challenges, particularly regarding interpretability.

1.2.1.1 Deep Neural Networks ([DNNs](#))

Deep Neural Networks ([DNNs](#)) are artificial neural networks with multiple hidden layers. They simulate the way the human brain processes information, enabling them to learn complex patterns from large datasets. [DNNs](#) consist of interconnected neurons organized in layers, and they adjust

weights during training to minimize prediction errors. This allows them to handle tasks like image recognition, speech processing, and more. However, **DNNs** are often seen as "black boxes" due to their complexity, making it difficult to understand their decision-making processes.

Advantages:

- High accuracy: **DNNs** excel in tasks such as image and speech recognition due to their ability to capture complex patterns.
- Automatic feature learning: They can automatically extract features from raw data, reducing the need for manual feature engineering.
- Scalability: **DNNs** can improve performance with more data and computational power.

Disadvantages:

- Interpretability: **DNNs** are hard to interpret due to their complexity.
- Computationally intensive: Training **DNNs** requires significant computational resources.
- Overfitting: **DNNs** are prone to overfitting on small datasets.

1.2.1.2 Random Forests

Random Forests are ensemble learning models that combine multiple decision trees. Each tree is trained on a random subset of data and features, and the final prediction is made by averaging the predictions of all trees. Random Forests reduce the risk of overfitting and offer some interpretability through feature importance scores.

Advantages:

- Reduced overfitting: Averaging predictions across multiple trees reduces overfitting.
- Versatility: Random Forests can handle classification, regression, and high-dimensional data.
- Feature importance: They provide insights into which variables influence predictions.

Disadvantages:

- Complexity: While more interpretable than **DNNs**, Random Forests can be hard to interpret with many trees.
- Computational cost: Training Random Forests can be resource-intensive.

1.2.1.3 Gradient Boosting Machines (GBMs)

Gradient Boosting Machines (**GBMs**) are ensemble models that build a series of decision trees, each correcting the errors of the previous one. This sequential learning process makes **GBMs** powerful for predictive tasks. However, like other ensemble methods, **GBMs** can be difficult to interpret.

Advantages:

- High accuracy: **GBMs** are known for their superior predictive performance.
- Flexibility: They can handle various types of data and loss functions.
- Feature importance: **GBMs** provide insights into which features contribute most to predictions.

Disadvantages:

- Overfitting: **GBMs** require careful tuning to avoid overfitting.
- Training time: They can be slow to train, especially on large datasets.
- Interpretability: **GBMs** remain complex and hard to interpret despite tools like **SHAP** and **LIME**.

1.2.1.4 Conclusion on complex modelsTable 1.1 – Summary of advantages and disadvantages of **DNNs**, Random Forests, and **GBMs**.

Model	Advantages	Disadvantages
DNNs	- High predictive accuracy - Automatic feature learning - Scalable with data and compute power	- Difficult to interpret (black box) - Computationally intensive - Prone to overfitting on small datasets
Random Forests	- Reduces overfitting by averaging trees - Versatile for classification and regression - Provides feature importance	- Complex with many trees - Computationally expensive
GBMs	- High predictive accuracy - Flexible with data and loss functions - Provides feature importance	- Prone to overfitting without tuning - Slow training time - Challenging to interpret

In summary, complex machine learning models like Deep Neural Networks, Random Forests, Gradient Boosting Machines, and support vector machines offer significant advantages in terms of predictive accuracy and flexibility. However, their complexity often comes at the cost of interpretability, posing challenges in understanding and explaining their predictions. As machine learning continues to be applied in critical areas, the need for interpretable models becomes increasingly important, driving research into methods that can demystify these powerful but opaque algorithms.

1.3 Challenges in interpreting complex models

The increasing use of complex machine learning models, such as Deep Neural Networks (**DNNs**), Random Forests, and Gradient Boosting Machines (**GBMs**), has raised significant interpretability concerns. While these models excel in predictive accuracy, understanding their decision-making process is crucial in fields like healthcare.

1.3.1 Deep Neural Networks (DNNs)

DNNs are notoriously difficult to interpret due to their black-box nature. Their multi-layered architecture captures complex, non-linear feature interactions, but these transformations are not easily comprehensible. Advanced techniques, such as feature visualization, are often required to make sense of their predictions.

1.3.2 Random Forests

Though each decision tree in a Random Forest is interpretable, the ensemble of many trees complicates the interpretation. Feature importance can be assessed globally, but understanding local decisions or feature interactions across trees remains a challenge.

1.3.3 Gradient Boosting Machines (GBMs)

GBMs sequentially build models to correct errors, creating complexity in understanding how features contribute to the final prediction. Tracing the contribution of features through multiple boosting stages requires careful analysis, and non-linear interactions between features further complicate interpretation.

1.3.4 Conclusion

While complex models offer great predictive power, their interpretability is limited. Addressing these challenges is critical for enhancing trust and usability, especially in decision-critical domains.

1.4 Model interpretation techniques

Machine learning models, especially complex ones such as Deep Neural Networks and ensemble methods, have become increasingly prevalent due to their exceptional predictive performance. However, their inherent complexity often renders their decision-making processes opaque, a phenomenon commonly referred to as the "black box problem." This lack of interpretability poses significant challenges, particularly in critical domains such as healthcare, finance, and legal systems, where trust and accountability are paramount.

To address these challenges, a variety of model interpretation techniques have been developed, aiming to provide insights into the internal workings of complex models. These techniques can be broadly categorized into two main types: **model-specific** approaches, tailored to specific algorithms, and **model-agnostic** methods, which can be applied across a wide range of models.

This section explores some of the most widely used model interpretation techniques, including **SHAP** (SHapley Additive exPlanations) and **LIME** (Local Interpretable Model-agnostic Explanations), alongside other methods such as Partial Dependence Plots (PDPs), ICE plots, and Feature Importance. The discussion highlights their underlying principles, practical applications, and limitations, with an emphasis on concrete examples and visual illustrations to enhance understanding. Additionally, a comparative analysis of these techniques is provided to evaluate their relative strengths and weaknesses.

1.4.1 Presentation of main techniques

This section delves into the primary model interpretation techniques, providing an in-depth analysis of their functionality, applications, advantages, and limitations. Each technique is discussed alongside practical examples and visual illustrations to facilitate understanding.

1.4.1.1 SHAP (SHapley Additive exPlanations)

Description and functioning **SHAP** is a unified framework for interpreting machine learning models based on Shapley values from cooperative game theory. It assigns each feature a contribution score that reflects its impact on the model's prediction. **SHAP** ensures fairness and consistency by evaluating feature contributions across all possible subsets of features.

— **Additivity**: Contributions of all features sum up to the prediction.

- **Consistency:** If a model is updated such that a feature contributes more, its **SHAP** value will not decrease.

Applications

- Identifying critical features in healthcare models, e.g., factors influencing diabetes risk.
- Explaining individual predictions in credit scoring, enhancing transparency in financial decisions.
- Visualizing global feature importance to understand model behavior in customer segmentation.

Advantages

- Provides both global and local interpretability.
- Theoretically grounded in cooperative game theory.
- Model-agnostic and applicable to any machine learning model.

Limitations

- Computationally expensive for models with high dimensionality.
- Approximations like Kernel **SHAP** can reduce accuracy.

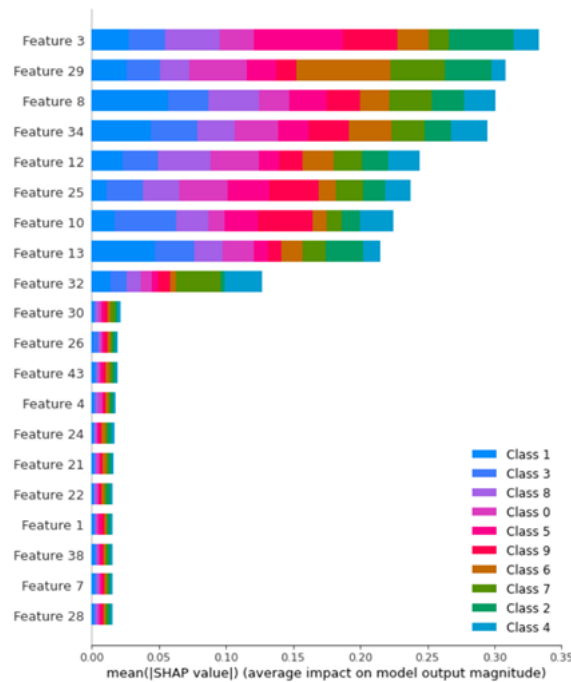


Figure 1.1 – **SHAP** summary plot showing feature contributions.

1.4.1.2 LIME (Local Interpretable Model-agnostic Explanations)

Description and functioning **LIME** approximates a complex model locally by fitting a simpler interpretable model, such as a linear regression or decision tree, to explain a specific prediction. It perturbs the input data, observes changes in the output, and learns a surrogate model to approximate the local decision boundary.

- Generates perturbed data around the instance being explained.
- Fits an interpretable model to approximate the behavior of the complex model locally.

Applications

- Debugging machine learning models by analyzing individual predictions.
- Explaining customer churn in e-commerce platforms.
- Enhancing trust in AI applications by providing clear explanations for specific decisions.

Advantages

- Intuitive and easy-to-understand explanations.
- Model-agnostic and applicable across machine learning models.
- Lightweight and computationally efficient for small datasets.

Limitations

- Explanations are localized and may not generalize well.
- Sensitive to the choice of perturbed data and interpretable model.

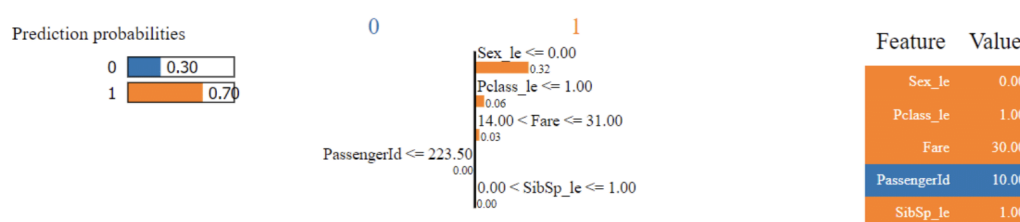


Figure 1.2 – LIME explanation of a specific prediction.

1.4.1.3 Other techniques

Partial Dependence Plots (PDPs) PDPs visualize the marginal effect of one or two features on the predicted outcome. They provide insights into global model behavior.

Applications:

- Analyzing the relationship between glucose levels and diabetes predictions.
- Examining salary trends in predictive hiring models.

Limitations:

- Assumes independence between features, which may not hold in practice.
- May be misleading when features are correlated.

Individual Conditional Expectation (ICE) Plots ICE plots extend PDPs by showing feature effects for individual instances, highlighting heterogeneity in predictions.

Applications:

- Identifying variability in patient outcomes for personalized medicine.

Limitations:

- Difficult to interpret for large datasets.

Feature importance Feature importance ranks features by their influence on model predictions. It is particularly useful for tree-based models.

Applications:

- Prioritizing variables in credit risk assessment.

Limitations:

- Does not provide instance-specific insights.
- Can be biased for correlated features.

Table 1.2 – Comparison of model interpretation techniques

Technique	Advantages	Limitations
SHAP	Global and local insights	Computational cost
LIME	Simple, localized explanations	Lack of generalizability
PDPs	Easy to visualize trends	Independence assumption
ICE	Highlights variability	Cluttered for large data
Feature Importance	Quick insights	No instance-level details

These techniques collectively provide a robust framework for interpreting machine learning models, each suited to different scenarios and use cases.

1.4.2 Comparison of techniques

This section compares the main model interpretation techniques, including SHAP, LIME, and other methods such as PDPs, ICE plots, and feature importance. The comparison is summarized in Table 1.3, covering their advantages, limitations, and typical applications.

Table 1.3 – Comparison of model interpretation techniques

Technique	Advantages	Limitations	Typical applications
SHAP	Provides global and local explanations; based on game theory for consistency.	Computationally expensive, especially for complex models or large datasets.	Analyzing feature influences in finance, healthcare, etc.
LIME	Model-agnostic, intuitive explanations, scalable.	Local explanations may not generalize, results depend on perturbation method.	Model debugging and decision explanations in e-commerce and banking.
PDPs	Simple and intuitive visualization of feature effects; useful for global/local interpretation.	Assumes feature independence, may be misleading for correlated features.	Understanding single feature effects, often used in recruitment or marketing.
ICE	Highlights heterogeneity of feature effects across instances.	Can be cluttered and difficult to interpret for large datasets.	Personalized studies, e.g., in healthcare or education.
Feature Importance	Simple global feature relevance, easy to interpret.	Does not explain individual predictions; biased for correlated features.	Global analysis for risk assessment models.

Model interpretation techniques play a critical role in ensuring the transparency and understanding of complex model predictions.

- **Strengths of local approaches (SHAP and LIME):** These tools provide instance-specific explanations, making them particularly useful for decision validation in critical fields like healthcare and finance.

- **Major limitations:** [SHAP](#) is computationally expensive, and [LIME](#) may lack generalizability. Global techniques such as [PDPs](#) and feature importance do not offer instance-level explanations.
- **Integrated perspective:** A combination of local and global techniques is often necessary to meet varied user needs and ensure a comprehensive understanding of models.

These observations highlight the need to adapt interpretation techniques to the specifics of the data and applications, while continuing to explore innovative methods to improve the efficiency and scalability of current tools.

1.5 Limitations and challenges of current interpretation methods

While techniques such as [SHAP](#) and [LIME](#) have advanced the interpretability of machine learning models, they still face several limitations. These challenges primarily revolve around computational complexity, scalability, and the trade-off between model accuracy and interpretability. Moreover, the local vs. global interpretability issue and conceptual challenges also complicate their application. This section explores these limitations in depth, along with potential solutions, such as the adoption of association rules, which can provide clearer insights into model behavior without some of the drawbacks of traditional techniques.

1.5.1 General issues with current methods

The main issue with current interpretability methods lies in their ability to provide understandable explanations without sacrificing model performance. While techniques like [SHAP](#) and [LIME](#) have been developed to address this transparency gap, they are not without their shortcomings, presenting challenges in terms of computational efficiency, scalability, and application in real-world scenarios.

1.5.2 Specific limitations

1.5.2.1 Computational complexity

One of the primary challenges of current methods, such as [SHAP](#) and [LIME](#), is computational complexity. [SHAP](#), for example, relies on game theory to allocate feature contributions, which involves evaluating numerous combinations of features. This exhaustive process can be computationally expensive, particularly for models with many features. Similarly, [LIME](#) generates perturbed instances and trains an interpretable model each time, increasing the computational load.

Example illustrated: Using [SHAP](#) on a diabetes prediction model with 100 features requires evaluating all possible combinations of these features, making it impractical on large datasets.

1.5.2.2 Scalability issues

Scalability remains a significant issue, especially when dealing with large datasets or real-time predictions. Techniques like [SHAP](#) and [LIME](#), which require recalculating explanations for each prediction, become impractical when immediate results are necessary, such as in real-time applications like fraud detection or personalized recommendations.

1.5.2.3 Interpretability vs. accuracy trade-off

Interpretability techniques often involve a trade-off between simplifying complex models to make them interpretable and preserving their accuracy. For instance, [LIME](#) works by simplifying a complex model locally with an interpretable model, such as linear regression, but this simplification may lose crucial information, making the explanation less precise.

1.5.2.4 Local vs. global interpretability

Another challenge is the distinction between local and global interpretability. [LIME](#) and [SHAP](#) provide local explanations for specific predictions, but they do not provide insights into the model's global behavior. Global interpretation methods, such as feature importance scores, give a broader view but fail to explain individual decisions. Balancing both approaches is essential for a complete understanding of model behavior.

1.5.2.5 Conceptual challenges

A significant conceptual challenge is the definition of interpretability itself. This term remains subjective and context-dependent. What may be considered an adequate explanation for a data scientist may not meet the needs of a business executive or end user. The lack of a universally accepted definition of interpretability complicates the development and evaluation of these techniques.

1.5.3 Summary in a table

Table 1.4 – Summary of limitations and challenges of interpretation techniques

Technique	Advantages	Limitations
SHAP (SHapley Additive Explanations)	Consistent, applicable to any model, offers both global and local interpretability	Computationally expensive, requires approximations for large datasets
LIME (Local Interpretable Model-agnostic Explanations)	Model-agnostic, intuitive explanations, scalable for large datasets	Local explanations may not generalize, results depend on perturbation method
PDPs (Partial Dependence Plots)	Simple visualization, useful for both global and local interpretation	Assumes independence between features, may mislead if features are highly correlated
ICE (Individual Conditional Expectation)	Highlights heterogeneity in feature effects across instances	Can be cluttered for large datasets, difficult to interpret
Feature Importance	Simple, global understanding of feature relevance	Does not explain individual predictions, biased for correlated features
Anchors	High-precision, human-readable rules	Computationally intensive, may not cover all scenarios

The current landscape of model interpretation techniques faces significant challenges, including computational complexity, scalability issues, and the trade-off between accuracy and interpretability. Additionally, balancing local vs. global interpretability and addressing the conceptual challenges of defining and evaluating interpretability remain key areas for improvement. Addressing these

challenges will require continued research and innovation, and alternative approaches, such as the use of association rules, may offer a more efficient and interpretable solution for some applications. Future research should focus on developing more scalable, context-sensitive methods that maintain the accuracy and reliability of machine learning models while enhancing their transparency and trustworthiness.

1.6 Association rules [1, 2, 3]

1.6.1 Basic concepts of association rules

Association rules are a fundamental concept in data mining, aimed at discovering interesting relationships or patterns within large datasets. These rules are expressed in the form of implications, showing how the occurrence of one set of items within a dataset may imply the occurrence of another set. The most common application of association rules is in market basket analysis, where the goal is to identify products that frequently co-occur in transactions, providing valuable insights for cross-selling strategies, inventory management, and marketing.

1.6.1.1 Formal definition

An association rule is typically written in the form $X \Rightarrow Y$, where X and Y are disjoint itemsets (i.e., $X \cap Y = \emptyset$). X is called the *antecedent*, and Y is the *consequent*. The rule implies that transactions containing X are likely to also contain Y . The strength of an association rule is measured using several metrics:

- **Support:** The support of a rule $X \Rightarrow Y$ is defined as the proportion of transactions in the dataset that contain both X and Y . Mathematically, this can be expressed as:

$$\text{Support}(X \Rightarrow Y) = \frac{\text{Number of transactions containing } X \cup Y}{\text{Total number of transactions}}$$

Support measures how frequently the items in the rule appear together in the dataset.

- **Confidence:** The confidence of a rule $X \Rightarrow Y$ is the proportion of transactions containing X that also contain Y . It is a measure of the reliability of the implication:

$$\text{Confidence}(X \Rightarrow Y) = \frac{\text{Support}(X \Rightarrow Y)}{\text{Support}(X)}$$

High confidence indicates that Y is likely to occur when X occurs.

- **Lift:** Lift measures the strength of a rule compared to the random occurrence of Y given X . It is defined as:

$$\text{Lift}(X \Rightarrow Y) = \frac{\text{Confidence}(X \Rightarrow Y)}{\text{Support}(Y)}$$

A lift value greater than 1 indicates that X increases the likelihood of Y , suggesting a strong association.

These metrics provide a framework for evaluating the significance and usefulness of discovered association rules.

1.6.1.2 Itemsets and frequent itemsets

In the context of association rule mining, an *itemset* is a collection of one or more items. For example, in market basket analysis, an itemset might be a set of products such as {bread, butter, milk}. The concept of frequent itemsets is central to the generation of association rules.

A *frequent itemset* is an itemset that appears in the dataset with a frequency that meets or exceeds a specified minimum support threshold. Identifying frequent itemsets is a crucial step in association rule mining because only these itemsets are considered for rule generation. The rationale is that itemsets with low support are unlikely to produce meaningful or actionable rules.

1.6.1.3 Rule generation process

The process of generating association rules involves two main steps:

1. **Finding frequent itemsets:** The first step is to identify all itemsets in the dataset that satisfy the minimum support threshold. This is typically done using algorithms such as Apriori or [FP-Growth](#), which efficiently prune the search space by exploiting the property that any subset of a frequent itemset must also be frequent.
2. **Generating rules from frequent itemsets:** Once frequent itemsets are identified, the next step is to generate rules that meet a minimum confidence threshold. For each frequent itemset, all possible non-empty subsets are considered as potential antecedents, and the remaining items are treated as the consequent. The confidence of each rule is then calculated, and rules that meet the confidence threshold are retained.

1.6.1.4 Challenges in association rule mining

While association rule mining is a powerful tool, it also presents several challenges:

- **Scalability:** The number of potential itemsets increases exponentially with the number of items in the dataset, leading to a combinatorial explosion. This makes the mining process computationally expensive, especially for large datasets.
- **Redundancy:** The mining process can generate a large number of rules, many of which may be redundant or trivial. Post-processing techniques are often required to filter and rank rules according to their usefulness.
- **Interpretability:** Although association rules are easy to understand, interpreting a large set of rules can be challenging, particularly when rules involve multiple items. Domain knowledge is often needed to assess the practical significance of the discovered rules.

The basic concepts of association rules provide a foundation for understanding how relationships within data can be uncovered and utilized. By focusing on frequent itemsets and leveraging metrics such as support, confidence, and lift, association rule mining offers a powerful method for discovering meaningful patterns in large datasets. However, challenges such as scalability, redundancy, and interpretability must be carefully managed to ensure that the results are both actionable and insightful.

1.6.2 Algorithms for mining association rules

The process of mining association rules is a critical step in uncovering meaningful patterns in large datasets. This process typically involves two main tasks: identifying frequent itemsets and generating association rules from these itemsets. Two of the most prominent algorithms used in this

domain are the Apriori algorithm and the **FP-Growth** (Frequent Pattern Growth) algorithm. These algorithms have been widely studied and implemented in various data mining applications due to their effectiveness in handling the computational challenges posed by large and complex datasets.

1.6.2.1 Apriori algorithm [4]

The Apriori algorithm, introduced by Agrawal and Srikant in 1994, is a key technique in association rule mining. It operates on the *Apriori property*, which states that all non-empty subsets of a frequent itemset must also be frequent. This property enables efficient pruning of the search space.

Detailed steps:

1. **Generation of candidate itemsets:** The algorithm starts by identifying frequent 1-itemsets, which are then used to generate candidate 2-itemsets.
2. **Pruning:** Infrequent itemsets are removed from consideration based on the support threshold.
3. **Iteration:** This process repeats for higher-order itemsets until no more frequent itemsets are found.
4. **Rule generation:** Association rules are generated from frequent itemsets and evaluated using confidence.

Optimizations:

- **Transaction reduction:** Transactions that do not contain frequent itemsets are removed.
- **Partitioning:** The dataset is divided into smaller subsets, with Apriori applied to each partition.
- **Hash-based counting:** Hashing is used to efficiently count itemset frequencies.

Challenges: The Apriori algorithm is computationally intensive and memory-consuming, particularly for large datasets.

1.6.2.2 FP-Growth algorithm [5]

FP-Growth, introduced by Han et al. in 2000, avoids candidate generation by using a compact data structure called an **FP-tree** to mine frequent itemsets directly.

FP-tree construction:

1. **Scan the dataset:** Identify frequent 1-itemsets and their support.
2. **Build the tree:** Each transaction is mapped onto a path in the **FP-tree**.
3. **Condense the tree:** Paths with common prefixes are merged to create a compact representation.

Mining the FP-tree: Frequent itemsets are extracted using a recursive process called *pattern growth*, which mines the tree from the bottom up.

Advantages:

- **Efficiency:** **FP-Growth** eliminates candidate generation, making it faster and more scalable than Apriori.
- **Memory usage:** The **FP-tree** is a compact data structure that reduces memory consumption.

Challenges: The **FP-tree** can be complex to construct for datasets with many distinct or infrequent items, and the algorithm is difficult to parallelize.

1.6.2.3 Advanced techniques and variants of **FP-Growth**

Several advanced techniques have been proposed to enhance **FP-Growth**'s performance:

- **Parallel **FP-Growth**:** Develops parallel versions for better scalability.
- **Improved memory management:** Optimizes **FP-tree** construction for large datasets.
- **Frequent itemset mining with constraints:** Allows mining with specific itemset constraints.

1.6.2.4 Comparison of Apriori and **FP-Growth**

Algorithm	Best for	Challenges
Apriori	Small to medium datasets	High computational cost for large datasets
FP-Growth	Large datasets	Complexity of tree construction

1.6.3 Applications of association rules

Association rules play a crucial role in uncovering hidden patterns within large datasets, making them valuable across numerous industries. In **retail and market basket analysis**, they enable retailers to enhance customer experiences by identifying frequently purchased items together, which informs product placement, promotional bundling, and cross-selling strategies. In **healthcare**, association rules are applied to analyze medical records, uncovering relationships between diseases, treatments, and drug interactions, leading to better patient care, early diagnosis, and optimized treatment plans.

Within the realm of **financial services**, association rules are instrumental in detecting fraudulent activities such as credit card fraud, anti-money laundering, and insider trading by recognizing unusual transaction patterns. The **telecommunications sector** benefits from these rules through enhanced network security and service optimization, including intrusion detection and analysis of call detail records. Similarly, **e-commerce platforms** leverage association rules for personalized product recommendations, dynamic pricing strategies, and predicting customer churn, thereby improving customer retention and engagement.

In **manufacturing and supply chain management**, association rules optimize production processes, improve inventory management, and strengthen supplier relations by identifying performance trends and potential risks. Overall, the wide-ranging applications of association rules demonstrate their ability to provide critical insights, supporting data-driven decision-making in diverse fields as data complexity continues to grow.

Conclusion

In this chapter, we explored the existing literature on complex machine learning models and the interpretation techniques associated with them. We highlighted the various types of models commonly used in predictive analytics, such as neural networks, decision trees, and ensemble methods

like Random Forests. The review underscored the critical challenge of interpretability in these models, particularly in sensitive fields like healthcare.

Furthermore, we discussed the existing interpretation techniques, including SHAP and LIME, and their respective strengths and limitations. This literature review provided a foundation for our methodology, emphasizing the need for interpretable models and the utility of association rules as a complementary approach. By synthesizing these insights, we established the groundwork for the subsequent chapters that detail our methodology and findings in interpreting diabetes risk assessment models.

Material and methods

Introduction

Complex machine learning models, while effective at producing high-quality predictions, often present a significant drawback: their lack of transparency. This opacity, commonly referred to as the "black box problem," is particularly concerning in sensitive applications such as medicine, where the ability to interpret predictions is crucial for ensuring trust among practitioners and patients. The complexity of these models often makes it difficult to understand how input features contribute to the predictions, which can lead to skepticism about their reliability and utility in clinical settings. One of the key challenges is to render these models comprehensible while preserving their predictive performance, thus facilitating better decision-making based on their outputs.

The primary objective of this study is to explore the use of association rules to explain the predictions of a complex diabetes risk model. This method relies on the extraction of frequent patterns from medical data that lead to a diabetic or non-diabetic classification. Association rule mining offers a way to uncover hidden relationships within the data, providing insights that can help both healthcare practitioners and patients understand the factors that contribute to diabetes risk. To achieve this, we employ a methodological framework based on the Apriori algorithm to identify significant associations between patient characteristics (such as age, glucose levels, BMI, etc.) and the target variable, 'Outcome.' This approach facilitates the generation of easily interpretable "If... Then..." rules for users, which can translate complex model predictions into actionable clinical insights.

The methodology adopted in this work is divided into several key steps, summarized as follows:

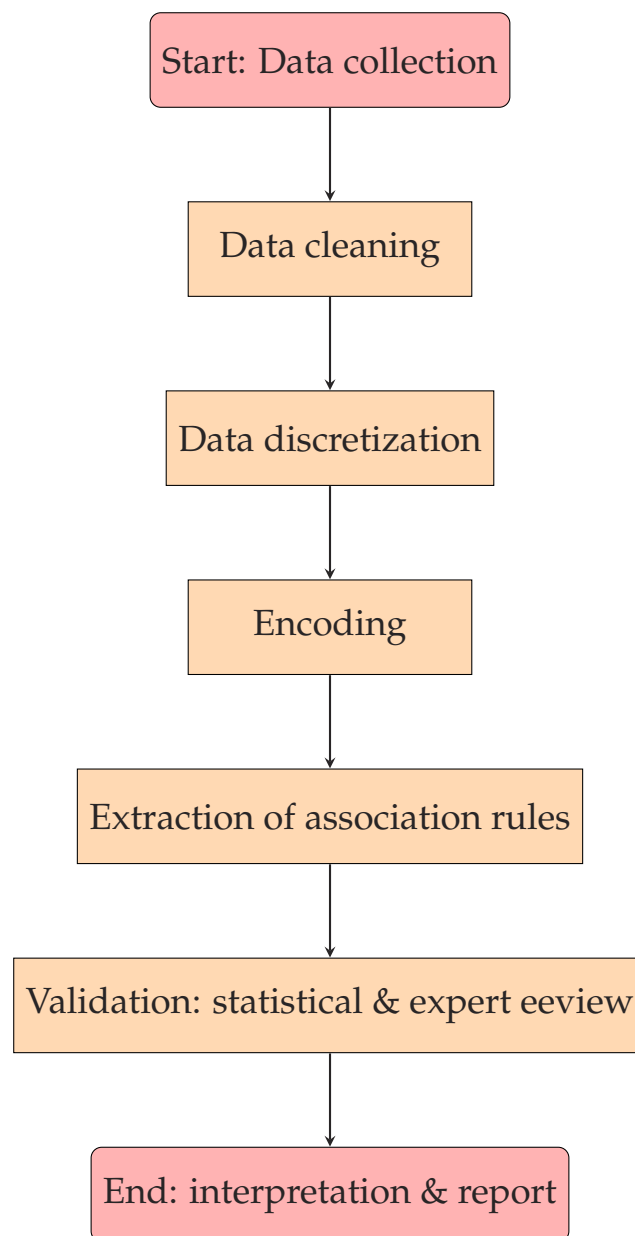


Figure 2.1 – Detailed methodology workflow

2.1 Data preparation

Data preparation is a critical component of the methodology, laying the foundation for robust analysis and accurate model performance. This stage involves transforming raw data into a clean, structured, and analysis-ready format that can be fed into machine learning models. In the context of diabetes risk prediction, ensuring that the data is properly prepared is essential to avoid introducing biases and to enhance the interpretability of the results. This section provides an in-depth description of each step in the data preparation process, including the description of the dataset, data cleaning, discretization, encoding, and final structuring.

2.1.1 Descriptive Analysis of the Dataset

2.1.1.1 Dataset Overview

The dataset used for this study contains a total of 768 observations and 9 variables relevant to diabetes risk assessment. Below is a detailed summary of the dataset.

2.1.1.2 Summary Statistics of Variables

Table 2.1 – Summary Statistics of the Dataset Variables

Variable	Description	Min	Max	Mean	Median	Std Dev
Pregnancies	Number of pregnancies	0.00	17.00	3.85	3.00	3.37
Glucose	Plasma glucose level (mg/dL)	0.00	199.00	120.89	117.00	31.97
BloodPressure	Diastolic blood pressure (mm Hg)	0.00	122.00	69.11	72.00	19.36
SkinThickness	Triceps skinfold thickness (mm)	0.00	99.00	20.54	23.00	15.95
Insulin	Serum insulin (mu U/ml)	0.00	846.00	79.80	30.50	115.24
BMI	Body Mass Index (kg/m ²)	0.00	67.10	31.99	32.00	7.88
DiabetesPedigreeFunction	Diabetes pedigree function	0.08	2.42	0.47	0.37	0.33
Age	Patient's age (years)	21.00	81.00	33.24	29.00	11.76

2.1.1.3 Diabetes Distribution

- **Diabetics:** 34.9% of observations (1)
- **Non-diabetics:** 65.1% of observations (0)

2.1.1.4 Specific Analysis

Table 2.2 – Specific Analysis of Key Features

Feature	Category	Proportion (%)
Glucose	High (>140 mg/dL)	25.0
	Normal (70-140 mg/dL)	72.9
	Low (<70 mg/dL)	2.1
Blood Pressure	High (>130 mm Hg)	0.0
	Normal (90-130 mm Hg)	7.8
	Low (<90 mm Hg)	92.2
BMI	Obese (≥ 30)	61.5
	Overweight (25-29.9)	23.3
	Normal (18.5-24.9)	13.3
	Underweight (<18.5)	2.0
Age	<35 years	63.5
	35-50 years	24.9
	>50 years	11.6

Table 2.3 – Feature Correlation with Diabetes

Feature	Correlation
Glucose	0.467
BMI	0.293
Age	0.238
Pregnancies	0.222
DiabetesPedigreeFunction	0.174
Insulin	0.131
SkinThickness	0.075
BloodPressure	0.065

2.1.1.5 Feature Correlations with Diabetes

1. Why choose diabetes

Diabetes was selected as the focus of this study due to several key reasons that highlight its relevance and importance in both medical research and machine learning model interpretability. Finally, here's a rendering of the dataset.

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1

Figure 2.2 – Dataset used as input

- **Prevalence and global health impact:** Diabetes has become a global health concern, with its prevalence continuously rising, particularly in developed and developing countries. According to the World Health Organization ([WHO](#)), the number of people living with diabetes has more than quadrupled since 1980, making it one of the leading causes of morbidity worldwide. The widespread nature of diabetes makes it an ideal candidate for exploring predictive models and their interpretability in a real-world, high-impact context.
- **Data availability and accessibility:** The availability of large, well-structured datasets, such as those found on platforms like Kaggle, provides an excellent opportunity for data-driven research. These datasets typically include a variety of health indicators (e.g., glucose levels, body mass index, age, etc.) that are closely related to diabetes, making them ideal for applying machine learning models to predict diabetes risk and understand the underlying factors.
- **Clinical relevance and interpretability:** Diabetes is a complex disease, and understanding

the contributions of different risk factors to its development is crucial. By using interpretable machine learning models, such as those based on association rules, we can provide insights that assist healthcare professionals in making more informed, transparent decisions about patient care.

- **Potential for preventive healthcare:** Diabetes, particularly Type 2 diabetes, is preventable with early detection and lifestyle interventions. By providing more transparent and interpretable models, this research aims to improve early diagnosis and prevention strategies. Understanding the key risk factors through association rules can inform healthcare policies and strategies to reduce the burden of diabetes on healthcare systems.
- **Opportunity for innovation:** The intersection of complex machine learning models and healthcare applications offers substantial opportunities for innovation. By leveraging techniques like association rule mining, this study aims to contribute to the growing field of model interpretability in healthcare, ultimately bridging the gap between predictive performance and model transparency.

Overall, diabetes was chosen as the focus due to its global impact, the availability of rich datasets, the importance of model interpretability in healthcare, and the potential for contributing to preventive healthcare strategies.

2. Overview of the dataset

The dataset used in this study is a collection of patient data reflecting real-world clinical measurements and demographic information. Each observation corresponds to a patient, providing insights into how different factors may influence their diabetes risk. The variables included in the dataset are:

Table 2.4 – Overview of the dataset

Variable	Description	Significance
Pregnancies	Number of times the patient has been pregnant.	Affects insulin sensitivity and may lead to gestational diabetes.
Glucose	Plasma glucose concentration during an oral glucose tolerance test.	Elevated levels are primary indicators of diabetes.
Blood Pressure	Diastolic blood pressure (mm Hg).	Hypertension is often associated with diabetes.
Skin Thickness	Thickness of the triceps skin fold (mm).	Indirect measure of body fat and insulin resistance.
Insulin	Serum insulin levels measured 2 hours post glucose intake.	High levels indicate insulin resistance, a key feature of type 2 diabetes.
BMI	Body Mass Index (weight in kg/(height in m) ²).	Higher BMI is associated with an increased diabetes risk.
Age	Age of the patient (years).	Significant risk factor for developing diabetes.
DiabetesPedigreeFunction	Genetic predisposition to diabetes based on family history.	Higher values indicate greater hereditary risk.
Outcome	Indicates presence (1) or absence (0) of diabetes.	Target variable for the analysis.

2.1.2 Data cleaning

Data cleaning is a crucial step in data preparation. Raw data often contains missing values, outliers, and inconsistencies that can affect the validity of the analysis. Effective data cleaning ensures the reliability of the dataset and prepares it for further processing and analysis.

1. Identifying and handling missing values

In the given dataset, missing values were identified primarily in the insulin and skin thickness variables. Missing values can arise from various sources, such as non-responses during data collection, incorrect data entry, or equipment failures. To assess the extent of missing values, the percentage of missing entries was calculated for each variable.

Several strategies were employed to manage these missing values:

- **Mean/median imputation:** For continuous variables like insulin, missing values were replaced by the mean or median of the respective variable. This technique ensures that the overall distribution of the variable is preserved.
- **Mode imputation:** For categorical variables, such as pregnancies, the mode (the most frequently occurring value) was used to fill in missing entries. This method maintains consistency within the categorical distribution.
- **Complete case deletion:** For variables where the proportion of missing values was negligible (e.g., less than 5%), complete records with missing entries were removed.
- **Flagging missing values:** In some cases, missing values carry information. Binary indicator variables were created to flag the presence of missing values.

2. Handling outliers and inconsistencies

Outliers can significantly skew the results of a machine learning model, leading to biased interpretations. Therefore, detecting and managing outliers is critical. Outliers in continuous variables, such as glucose and BMI, were identified using boxplots and Z-scores (values beyond 3 standard deviations from the mean).

Outliers were treated in the following ways:

- **Transformation:** Some outliers were transformed using logarithmic or square root transformations to reduce their impact.
- **Replacement:** Outliers that were deemed unrealistic (e.g., a glucose value of 500 mg/dL) were replaced with the nearest plausible value.
- **Retention:** If the outlier represented a rare but possible extreme case (e.g., a patient with severe diabetes), it was retained for further analysis.

3. Normalization and scaling

Continuous variables were normalized to ensure that they have a mean of 0 and a standard deviation of 1. Normalization is essential for distance-based algorithms like KNN and for models sensitive to feature scaling. This step ensures that all variables contribute equally to the model.

2.1.3 Data discretization

Discretization is the process of transforming continuous variables into discrete categories. This step is particularly important for association rule mining, which relies on categorical data to identify

patterns.

1. Discretization of medical variables

The following variables were discretized based on clinical guidelines:

Table 2.5 – Summary of data discretization

Variable	Categories
Glucose	Low (0-70 mg/dL), Normal (70-100 mg/dL), High (above 100 mg/dL)
Blood pressure	Low (0-90 mm Hg), Normal (90-120 mm Hg), High (above 120 mm Hg)
BMI	Underweight, Normal, Overweight, Obese
Age	Young (below 30), Adult (30-50), Senior (above 50)

	Pregnancies_Bin_None	Pregnancies_Bin_Low	Pregnancies_Bin_Medium	\
0	False	False	False	
1	False	True	False	
2	False	False	False	
3	False	True	False	
4	True	False	False	
..	...	...	...	
763	False	False	False	
764	False	True	False	
765	False	False	True	
766	False	True	False	
767	False	True	False	

	Pregnancies_Bin_High	Glucose_Bin_Low	Glucose_Bin_Normal	...	\
0	True	False	False	...	
1	False	False	True	...	
2	True	False	False	...	
3	False	False	True	...	
4	False	False	False	...	
..	...	...	...	...	
763	True	False	False	...	
764	False	False	False	...	
765	False	False	False	...	
766	False	False	False	...	
767	False	False	True	...	

	Insulin_Bin_Low	Insulin_Bin_Normal	Insulin_Bin_High	\
0	False	False	False	
1	False	False	False	
2	False	False	False	
3	False	True	False	
4	False	False	True	
..	...	...	...	
763	False	False	True	
764	False	False	False	
765	False	True	False	
766	False	False	False	
767	False	False	False	

Figure 2.3 – Discretization results

2. Benefits of discretization

Discretization is a crucial step in transforming continuous variables into discrete categories, making them easier to interpret and analyze. In the context of medical datasets, discretizing continuous variables, such as glucose and BMI, into categorical ranges (e.g., Low, Normal, High) aligns the data

with real-world clinical standards. This alignment offers several significant benefits: Discretization transforms continuous variables into discrete categories, improving interpretability and analysis. In medical datasets, this process aligns data with clinical standards and offers several advantages:

- **Improved interpretability:** Categorical variables, such as glucose levels categorized as Low, Normal, and High, are easier for practitioners and researchers to understand.
- **Facilitates rule extraction:** Discretization is vital for association rule mining, as algorithms like Apriori and [FP-Growth](#) work with categorical data, simplifying the identification of patterns.
- **Reduced model complexity:** By categorizing continuous variables, discretization decreases model complexity, minimizes the need for transformations, and helps prevent overfitting.
- **Handles non-linear relationships:** It enables the detection of non-linear relationships that may not be visible with continuous data, such as varying impacts of blood pressure on diabetes risk.
- **Alignment with domain knowledge:** Categorical interpretations are more intuitive for medical professionals, enhancing the practical utility of analysis.
- **Mitigates data sparsity and outliers:** Grouping values into categories reduces the impact of extreme outliers on the overall analysis.

In summary, discretization simplifies data structures and enhances the robustness of analyses, transforming raw numerical data into clinically relevant categories that improve rule extraction and model building.

In summary, discretization not only simplifies the data structure but also enhances the interpretability and robustness of the analysis. This step transforms raw numerical data into clinically relevant categories, making the subsequent stages of rule extraction and model building more effective.

2.1.4 Encoding

Encoding is a critical step in data preparation, particularly when working with categorical variables. After discretization, the categorical variables must be transformed into a format suitable for machine learning algorithms. In this study, **One-Hot Encoding** was employed to convert the discretized categories into binary variables. This encoding technique is widely used for converting categorical data into a format that can be easily interpreted by algorithms.

1. One-hot encoding: process and implementation

One-Hot Encoding creates a new binary column for each category of a given variable. For instance, if the `Glucose` variable has three categories (Low, Normal, High), One-Hot Encoding will generate three separate binary columns: `Glucose_Low`, `Glucose_Normal`, and `Glucose_High`. Each observation in the dataset will have a value of 1 in the column corresponding to its category and 0 in the others. This process is repeated for every discretized variable in the dataset.

Advantages of one-hot encoding:

- **Preserves Categorical Relationships:** One-Hot Encoding maintains the integrity of categorical data by ensuring that each category is treated as a separate, independent variable.
- **Avoids Ordinal Relationships:** Unlike label encoding, One-Hot Encoding does not imply any ordinal relationship between categories, which is crucial for variables like [‘BMI’](#) or [‘Age’](#), where categories such as [‘Underweight’](#) and [‘Obese’](#) should not be interpreted as having a linear progression.

- **Enables Effective Rule Extraction:** In association rule mining, One-Hot Encoded variables represent discrete features that can be used to generate itemsets, making them ideal for algorithms like Apriori, which require binary input.

	Pregnancies_Bin_None	Pregnancies_Bin_Low	Pregnancies_Bin_Medium	\
0	0	0	0	
1	0	1	0	
2	0	0	0	
3	0	1	0	
4	1	0	0	
..	...	...	...	
763	0	0	0	
764	0	1	0	
765	0	0	1	
766	0	1	0	
767	0	1	0	

	Pregnancies_Bin_High	Glucose_Bin_Low	Glucose_Bin_Normal	...	\
0	1	0	0	...	
1	0	0	1	...	
2	1	0	0	...	
3	0	0	1	...	
4	0	0	0	...	
..	...	...	...	...	
763	1	0	0	...	
764	0	0	0	...	
765	0	0	0	...	
766	0	0	0	...	
767	0	0	1	...	

	Insulin_Bin_Low	Insulin_Bin_Normal	Insulin_Bin_High	\
0	0	0	0	
1	0	0	0	
2	0	0	0	
3	0	1	0	
4	0	0	1	
..	...	...	...	
763	0	0	1	
764	0	0	0	
765	0	1	0	
766	0	0	0	
767	0	0	0	

Figure 2.4 – Encoding result

2. Practical considerations for encoding

Despite its advantages, One-Hot Encoding can lead to an increase in dimensionality, especially when dealing with variables that have many categories. To mitigate this, the dataset was carefully examined to ensure that only relevant and meaningful categories were retained after discretization. For example, if certain categories were sparsely populated, they were either merged with similar categories or excluded to avoid sparsity in the encoded dataset.

Additionally, encoded variables were analyzed for multicollinearity, as the creation of multiple binary columns can sometimes introduce redundancy, which may affect model performance. Techniques such as variance inflation factor (VIF) analysis were used to identify and address any potential multicollinearity issues.

2.1.5 Final dataset structure

The final dataset structure was carefully designed to facilitate the analysis and interpretability of diabetes risk. After data cleaning, discretization, and encoding, the dataset was transformed into a binary format that is compatible with machine learning models and association rule mining algorithms.

1. Binary representation of variables

Each observation in the dataset is now represented as a unique combination of binary-encoded features. For example:

- **Pregnancies:** Encoded as separate binary columns (`Pregnancies_None`, `Pregnancies_Low`, etc.), representing the different categories derived from the discretization process.
- **Glucose:** Encoded into three binary columns (`Glucose_Low`, `Glucose_Normal`, `Glucose_High`), indicating the glucose levels for each patient.
- **Blood pressure:** Encoded similarly to indicate different levels of blood pressure.
- **BMI, age, and insulin:** These variables are also converted into their respective binary categories, ensuring that each feature is represented as a separate binary column.

2. Target variable: Outcome

The target variable, 'Outcome', remains as a binary indicator of diabetes presence or absence. This variable serves as the conclusion in the association rules, helping to identify patterns that lead to 'Outcome=1' (diabetic) or 'Outcome=0' (non-diabetic).

3. Preparing the dataset for analysis

The final dataset, comprising binary columns for each category, is structured as a transaction table where each row represents a patient's unique profile. This format is ideal for association rule mining, as it allows the algorithms to efficiently search for frequent itemsets and generate rules that explain the relationships between different health indicators and diabetes outcomes.

In conclusion, the transformation from raw data to a well-structured, encoded dataset enables effective analysis and supports the subsequent stages of model training and rule extraction. This structured dataset is now ready for machine learning and association rule mining, providing a solid foundation for deriving meaningful insights into diabetes risk factors.

2.2 Development of the rule extraction algorithm.

2.2.1 Algorithm used for rule extraction

Association rule mining is a key data mining technique used to uncover interesting relationships between variables in large datasets. It identifies frequent itemsets and generates rules that highlight these associations. The objective is to find patterns that reveal which combinations of attributes frequently occur together. In the context of this study, the Apriori algorithm was utilized as the primary technique for extracting association rules. This subsection provides an in-depth explanation of the Apriori algorithm, its functioning, and its relevance to the problem at hand.

2.2.1.1 The Apriori algorithm: step-by-step process

The Apriori algorithm consists of two main steps: finding frequent itemsets and generating association rules. Below is a step-by-step explanation of how Apriori works:

1. **Step 1: Generating frequent itemsets** The algorithm begins by identifying individual items (single features) in the dataset that meet a specified minimum support threshold. These items are known as "frequent 1-itemsets." For example, in the diabetes dataset, features such as "High Glucose" or "Obese BMI" might be identified as frequent 1-itemsets if they appear frequently enough across patients.

Next, the algorithm generates larger itemsets (2-itemsets, 3-itemsets, etc.) by combining smaller frequent itemsets, using the apriori property to eliminate itemsets that do not meet the support threshold. This iterative process continues until no more frequent itemsets can be generated.

2. **Step 2: Generating association rules** Once the frequent itemsets are identified, the algorithm proceeds to generate association rules from these itemsets. Each frequent itemset is divided into a "premise" and a "conclusion," creating rules of the form $X \Rightarrow Y$. For example, a rule might state: "If a patient has a high BMI and low blood pressure, then they are at risk of diabetes."

The confidence and lift of each rule are calculated to evaluate its strength. Only the rules that meet the minimum confidence threshold are retained. This filtering ensures that the generated rules are both statistically significant and clinically relevant.

2.2.1.2 Application of apriori in the study

In this research, the Apriori algorithm was applied to identify significant patterns in the medical dataset, revealing how different health indicators collectively impact diabetes risk. The goal was to extract rules that can be easily interpreted by healthcare practitioners, providing insights into which combinations of characteristics are most strongly associated with the presence of diabetes.

Example of a rule: One of the rules extracted using Apriori might be expressed as: "If a patient has High_Glucose and Obese_BMI, then Outcome = 1 (diabetes)." The support, confidence, and lift for this rule are calculated to assess its relevance.

- **Support:** Indicates how often this specific combination of characteristics occurs among diabetic patients in the dataset.
- **Confidence:** Measures how likely a patient with both high glucose and obesity is to have diabetes.
- **Lift:** Determines whether the presence of these characteristics together significantly increases the likelihood of diabetes compared to their independent occurrence.

This rule, if validated, can inform clinical decision-making by highlighting critical risk factors. By using Apriori, the study aims to bridge the gap between complex predictive models and actionable clinical insights, ensuring that the patterns extracted are both statistically sound and practically relevant.

2.2.1.3 Advantages and limitations of apriori

The Apriori algorithm offers several advantages for association rule mining, particularly in the context of medical data analysis:

- **Ease of interpretation:** The rules generated by Apriori are easy to understand and interpret, making them suitable for non-technical stakeholders, such as healthcare professionals.
- **Efficiency in large datasets:** Due to its iterative nature and use of the apriori property, the algorithm efficiently reduces the search space, allowing it to handle large datasets effectively.
- **Flexibility in rule generation:** Apriori can generate rules with varying levels of complexity, capturing intricate relationships between variables.

However, Apriori also has limitations:

- **Computational cost:** The algorithm can become computationally expensive for very large datasets, especially when the minimum support threshold is low.
- **Handling continuous variables:** Apriori works best with categorical data, making it necessary to discretize continuous variables before applying the algorithm, as seen in this study.

Despite these limitations, Apriori was chosen for its interpretability and relevance in uncovering meaningful patterns in health data, making it an ideal choice for this research.

2.2.2 Parameterization and configuration of the apriori algorithm

The Apriori algorithm relies on several parameters that guide its search for frequent itemsets and the generation of association rules. Correctly setting these parameters is crucial for meaningful rule extraction. In this subsection, we outline the key parameters used in our implementation of the Apriori algorithm for diabetes risk assessment.

2.2.2.1 1. Minimum support threshold

The minimum support threshold determines which itemsets are considered "frequent." It specifies the minimum proportion of observations that must contain a given itemset.

Definition: Support for an itemset X is defined as:

$$\text{Support}(X) = \frac{\text{Number of transactions containing } X}{\text{Total number of transactions}}$$

In this study, an itemset represents combinations of patient characteristics. For instance, the itemset {High_Glucose, Obese_BMI} has a support value of 0.25 if it appears in 25% of records.

Selection of minimum support: Setting the support value involves a trade-off between capturing rare itemsets and computational efficiency. We set the threshold at 0.15 (15%) to balance capturing clinically relevant combinations while maintaining efficiency.

Rationale for 15% minimum support: This threshold filters out rare combinations unlikely to generalize while retaining moderately frequent itemsets that provide valuable insights.

2.2.2.2 2. Minimum confidence threshold

Confidence measures the reliability of an association rule, quantifying how often the conclusion is true given the premise.

Definition: For a rule $X \Rightarrow Y$:

$$\text{Confidence}(X \Rightarrow Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)}$$

For example, a confidence value of 0.7 indicates that 70% of patients with high glucose and obesity are diabetic.

Selection of minimum confidence: We chose a threshold of 0.5 (50%) to ensure that only strong, reliable rules are extracted.

Rationale for 50% minimum confidence: This choice ensures that the retained rules are clinically meaningful, suggesting consistent patterns for further investigation.

2.2.2.3 3. Lift metric

The **lift** metric evaluates the strength of the association between variables. A lift value greater than 1 indicates a positive association.

Definition: Lift is calculated as:

$$\text{Lift}(X \Rightarrow Y) = \frac{\text{Confidence}(X \Rightarrow Y)}{\text{Support}(Y)}$$

Usage in this study: The lift metric was used to prioritize the strongest associations. For example, a rule with a lift value of 1.7 suggests that patients with both obesity and high glucose are 1.7 times more likely to have diabetes.

2.2.2.4 4. Configuring the Apriori algorithm

Configuring the Apriori algorithm involves setting thresholds and defining itemset generation.

- **Use of the Apriori property:** This property minimizes candidate itemsets generated, reducing computational complexity in high-dimensional datasets.
- **Filtering based on antecedent and consequent constraints:** We focused on rules where the consequent included only the variable `Outcome`, ensuring that extracted rules directly relate to diabetes presence or absence.

2.2.2.5 5. Practical considerations for parameter tuning

Tuning the parameters of the Apriori algorithm is not a one-size-fits-all approach. It involves iterative experimentation and evaluation to identify the optimal settings. During this study, several combinations of support and confidence values were tested to strike a balance between rule significance and computational efficiency. The chosen parameters (15% support and 50% confidence) were found to produce rules that are both clinically meaningful and computationally feasible.

Moreover, the parameter settings were validated using expert consultation and cross-referencing with existing medical literature to ensure that the rules generated align with known risk factors for diabetes. This validation step enhances the credibility of the findings and ensures that the extracted rules are robust and interpretable.

2.2.3 Examples of extracted rules

The extraction of association rules using the Apriori algorithm yields valuable insights into the relationships between various health indicators and diabetes outcomes. This section presents specific examples of the rules generated from the dataset, highlighting their implications and relevance for clinical practice.

1. Overview of the extracted rules

After applying the Apriori algorithm with a minimum support of 15% and a minimum confidence of 50%, a set of association rules was generated that elucidates the conditions under which diabetes is more likely to occur. Each rule is expressed in the form $X \Rightarrow Y$, where X consists of the antecedents (or premises), and Y represents the consequent (or conclusion).

The rules are accompanied by key metrics such as support, confidence, and lift, which provide insight into their strength and significance. Below are notable examples of extracted rules from the analysis:

2. Example rules

1. Rule 1:

$$\text{High_Glucose} \wedge \text{Obese_BMI} \Rightarrow \text{Outcome}=1$$

- **Support**: 0.25 - **Confidence**: 0.65 - **Lift**: 1.8

Interpretation: This rule indicates that patients with both high glucose levels and an obese BMI have a 65% probability of being diagnosed with diabetes. The lift value of 1.8 suggests that the occurrence of this combination significantly increases the likelihood of diabetes compared to random chance. Clinically, this reinforces the importance of monitoring both glucose and BMI in patients at risk for diabetes.

2. Rule 2:

$$\text{Age_Adult} \wedge \text{High_Glucose} \Rightarrow \text{Outcome}=1$$

- **Support**: 0.30 - **Confidence**: 0.58 - **Lift**: 1.5

Interpretation: This rule indicates that among adult patients, those with high glucose levels have a 58% probability of being diabetic. The lift value of 1.5 signifies a positive association, suggesting that age and glucose levels collectively contribute to diabetes risk.

3. Rule 3:

$$\text{BloodPressure_High} \Rightarrow \text{Outcome}=1$$

- **Support**: 0.20 - **Confidence**: 0.55 - **Lift**: 1.4

Interpretation: This rule shows that high blood pressure alone is associated with a 55% probability of diabetes. The findings underline the significance of hypertension as a risk factor for diabetes, suggesting that monitoring blood pressure is critical in diabetes prevention strategies.

4. Rule 4:

$$\text{Age_Senior} \wedge \text{Pregnancies_High} \Rightarrow \text{Outcome}=1$$

- **Support**: 0.15 - **Confidence**: 0.60 - **Lift**: 1.6

Interpretation: This rule indicates that older patients with a high number of pregnancies have a 60% chance of being diagnosed with diabetes. The rule highlights the intersection of age and reproductive history as significant factors in diabetes risk assessment.

5. Rule 5:

$$\text{Insulin_High} \Rightarrow \text{Outcome}=1$$

- **Support**: 0.22 - **Confidence**: 0.70 - **Lift**: 1.9

****Interpretation**:** This rule reveals that patients with high insulin levels have a 70% likelihood of having diabetes, with a lift of 1.9 indicating a strong association. This emphasizes the importance of monitoring insulin levels as a potential predictor of diabetes.

Extraction and explanation of association rules

After the extraction of association rules using the Apriori algorithm, the next step involves generating detailed explanations for each rule. To enhance interpretability, a [PDF](#) report is automatically generated at the end of the algorithm's execution. This report provides a comprehensive explanation of each rule, detailing the antecedents, consequents, and key metrics such as support, confidence, and lift.

The interpretations of these rules are generated using [GPT API](#) that applies advanced language models to provide a textual explanation for each association. This step ensures that the rules are not only extracted but also explained in a way that is understandable and actionable for end-users.

- **Rule explanations:** Each rule is accompanied by a detailed description of the relationships between input features and the model's output. This allows domain experts to interpret how combinations of features influence the prediction of diabetes risk.
- **PDF report:** The report is automatically saved as a [PDF](#) document, containing the explanation of all extracted rules. This report serves as a tool for further analysis and validation of the model's interpretability.

The [PDF](#) report also serves as a basis for domain-specific validation, allowing experts to review and validate the relevance of the extracted rules in a clinical context.

3. Clinical relevance of extracted rules

The extracted rules provide a framework for understanding the complex interplay between various health indicators and diabetes risk. These insights are particularly valuable for healthcare professionals as they offer a basis for targeted interventions and preventive strategies. The rules can inform clinical decision-making, guiding practitioners to focus on patients who exhibit specific combinations of risk factors.

For example, the first rule underscores the importance of regularly screening patients with both high glucose levels and obesity. Healthcare providers may prioritize lifestyle modifications and closer monitoring for such patients, potentially reducing their risk of progressing to diabetes.

Furthermore, these rules can contribute to developing risk stratification models that incorporate multiple factors, allowing for more personalized patient care. By leveraging the insights gained from the association rules, healthcare systems can improve resource allocation and patient outcomes.

4. Limitations and considerations

While the extracted rules provide valuable insights, it is essential to consider their limitations:

- **Causation vs. correlation:** The rules identified through the Apriori algorithm indicate associations, not causal relationships. Therefore, caution must be taken when interpreting the results; further investigation is needed to establish causality.
- **Generalizability:** The findings may be specific to the dataset used in this study and may not be directly applicable to other populations. External validation in diverse cohorts is necessary to assess the generalizability of the rules.

- **Potential for overfitting:** While setting appropriate thresholds for support and confidence helps mitigate this risk, there is still a possibility of overfitting, particularly with complex datasets. Continuous evaluation and validation of the rules are crucial.

In conclusion, the examples of extracted rules demonstrate how association rule mining can yield meaningful insights into diabetes risk factors. By understanding these relationships, healthcare practitioners can make informed decisions to improve patient care and outcomes.

2.3 Implementation and tools section

2.3.1 Programming language and environment

The project was implemented using Python within the JupyterLab environment, part of the Anaconda distribution. This choice was motivated by Python's extensive library support for data analysis and machine learning, the interactive nature of JupyterLab, and Anaconda's overall flexibility.

2.3.1.1 Why python?

Python is a versatile programming language widely used in data science and machine learning due to its readability and vast ecosystem of libraries. Key advantages include:

- **Extensive libraries:** Libraries like `pandas`, `numpy`, and `scikit-learn` facilitate data pre-processing and model building.
- **Ease of integration:** Python easily integrates with various data formats (e.g., [CSV](#), [SQL](#)), allowing straightforward data manipulation.
- **Community support:** Python's large community and comprehensive documentation ensure robust support and active maintenance of libraries.

2.3.1.2 JupyterLab for interactive development

JupyterLab is an open-source [IDE](#) chosen for its flexibility and features beneficial for data science workflows. Key features include:

- **Interactive coding:** Allows real-time data exploration, crucial for iterative testing in machine learning projects.
- **Markdown integration:** Enhances clarity by allowing documentation within the code.
- **Multiple outputs support:** Displays graphs, tables, and text together, providing a unified view of analyses.
- **Notebook organization:** Supports multiple notebooks, enabling organized workflows for different project phases.

Advantages of jupyterLab:

JupyterLab's interactive nature made it ideal for this project, allowing for rapid experimentation and immediate feedback, which was crucial for:

- **Data exploration:** Real-time visualizations helped identify missing values and outliers.
- **Debugging:** Running individual cells facilitated effective debugging and output analysis.
- **Documentation:** Markdown cells documented methodological choices, enhancing reproducibility.

2.3.1.3 Anaconda distribution

Anaconda is a comprehensive distribution for scientific computing, simplifying package management and providing a pre-configured environment, including JupyterLab and essential libraries.

Benefits of using anaconda:

- **Pre-installed libraries:** Includes essential data science libraries, simplifying setup.
- **Environment management:** Allows creation of isolated environments to prevent compatibility issues.
- **Integrated tools:** Offers flexibility with tools like JupyterLab, selected for its interactive capabilities.

Overall implementation environment:

In summary, the project was developed in a highly interactive manner using JupyterLab within Anaconda. Python's libraries enabled efficient data handling, while JupyterLab facilitated iterative testing and documentation. Anaconda ensured a stable environment, contributing to the project's success and allowing for a structured approach to data preparation and rule extraction.

2.3.2 Libraries and packages

The project's success relied on Python's rich ecosystem of libraries for data manipulation, machine learning, and association rule mining. Each library was selected for specific roles in the data preparation and analysis pipeline.

2.3.2.1 pandas: Data manipulation and analysis

The `pandas` library is widely used for data manipulation. Key functionalities include:

- **Loading the dataset:** The dataset was loaded into a `pandas DataFrame` for structured data representation.
- **Data cleaning:** Functions like `.isnull()` and `.dropna()` were used for handling missing values and outliers.
- **Data transformation:** The `.cut()` method discretized continuous variables into categorical bins, essential for rule extraction.
- **One-hot encoding:** The `.get_dummies()` function converted categorical variables into a binary format.

2.3.2.2 scikit-learn: Machine learning and data preprocessing

`scikit-learn` offers tools for data analysis and machine learning. Its contributions include:

- **Discretization:** The `KBinsDiscretizer` module provided alternative methods for discretizing continuous variables.
- **Normalization:** The `StandardScaler` and `MinMaxScaler` were used to standardize features.
- **Model implementation:** Baseline models like Logistic Regression and Decision Tree were implemented to evaluate dataset performance.

2.3.2.3 mlxtend: Association rule mining

`mlxtend` extends `scikit-learn` functionalities, specifically for association rule mining. Key functionalities include:

- **Apriori algorithm:** The `apriori()` function identified frequent itemsets within the dataset.
- **Rule generation:** The `association_rules()` function generated rules, calculating support, confidence, and lift.
- **Rule filtering:** Integration with `pandas` allowed for easy manipulation and filtering of extracted rules.

2.3.2.4 `numpy`: Numerical operations

`numpy` is the foundational library for numerical operations. Its applications included:

- **Array operations:** `numpy` arrays facilitated numerical data handling during preprocessing.
- **Mathematical computations:** Calculations for support and confidence during rule mining were performed using `numpy`'s functions.

2.3.2.5 `matplotlib`: Data visualization

`matplotlib` is used for creating visualizations. Key applications include:

- **Data distribution:** Histograms and boxplots visualized distributions of variables.
- **Result visualization:** Association rules were represented using bar and scatter plots.

2.3.2.6 Choices justifications

Table 2.6 – Technical alternatives and choice justifications

Component	Alternatives	Chosen option	Justification
Programming language	R, Julia	Python	Readable syntax with a vast ecosystem of libraries.
Development environment	Jupyter Notebook, Spyder	JupyterLab	Interactive exploration, Markdown integration, and multiple notebook management.
Distribution	Virtual environment, docker	Anaconda	Simplified package and environment management, with JupyterLab included.
Data library	Dask, Vaex	pandas	Widely adopted and optimized for in-memory data manipulation.
Machine learning library	TensorFlow, keras	scikit-learn	Integrated tools for classical models with rich documentation.
Association rule library	Orange, pyCaret	mlxtend	Optimized for association rules, tailored to project needs.
Numerical library	SciPy, cuPy	numpy	Reference for numerical operations with high performance.
Visualization library	Seaborn, Plotly	matplotlib	Full flexibility for customizing plots.

2.3.2.7 Summary

In summary, the combination of `pandas`, `scikit-learn`, `mlxtend`, `numpy`, and `matplotlib` formed a robust toolkit for data preparation and rule extraction, facilitating seamless integration of data handling, modeling, rule mining, and visualization.

2.3.3 Validation of the approach

Validation is a critical step in any data mining or machine learning project, ensuring that the patterns and models generated are robust, reliable, and applicable in real-world scenarios. The goal of validation is to demonstrate that the rules extracted are not only statistically significant but also clinically meaningful. This section outlines the comprehensive validation strategy used in this study,

which includes statistical validation, expert evaluation, comparison with other interpretation methods, and considerations for future validation efforts.

2.3.3.1 Statistical validation of association rules

The first stage of validation focused on ensuring that the extracted association rules were statistically sound and free from overfitting. To achieve this, key metrics such as **support**, **confidence**, and **lift** were rigorously evaluated for each rule. These metrics serve as indicators of the strength and reliability of the associations identified by the Apriori algorithm.

Table 2.7 – Statistical metrics for rule evaluation

Metric	Definition and Use
Support	Frequency of a feature combination in the dataset. Minimum threshold: 15%.
Confidence	Probability of the consequent given the antecedent. Minimum threshold: 50%.
Lift	Indicates how much the antecedent increases the likelihood of the consequent. Rules with Lift > 1 were retained.

Rules that did not meet these thresholds were excluded, ensuring the remaining rules were statistically reliable.

2.3.3.2 Expert validation and clinical relevance

Expert consultation was conducted to evaluate the clinical significance of the rules. Healthcare professionals reviewed each rule in the context of existing medical knowledge on diabetes risk factors.

Table 2.8 – Process of expert validation

Step	Details
Presentation of rules	Rules were presented with antecedents, consequents, and metrics (support, confidence, lift).
Expert review	Rules were analyzed for clinical validity and relevance to known diabetes risk factors.
Incorporation of feedback	Irrelevant rules were excluded or modified based on expert input.
Outcome	Clinically valid rules that align with expert opinions were retained.

This process ensured that the extracted rules were both statistically significant and clinically actionable.

2.3.3.3 Cross-validation with other interpretation techniques

To further evaluate the robustness of the association rules, they were compared with results obtained using **SHAP** and **LIME**.

This complementary analysis validated the robustness and relevance of the extracted rules.

Table 2.9 – Comparison of association rules with SHAP and LIME

Technique	Key Insight	Complementary Findings
SHAP	Revealed global feature importance. High glucose and BMI had the most impact.	Matched frequent patterns in rules involving glucose and BMI.
LIME	Provided local explanations for individual predictions.	Validated rules linking specific feature combinations to diabetes.

2.3.3.4 Future validation plans

- **Cross-population validation:** Testing the rules on datasets from different populations to assess generalizability.
- **Longitudinal studies:** Evaluating rules over time to establish their predictive utility for early diagnosis.
- **Integration of additional features:** Including genetic markers and lifestyle factors to refine the rules further.

Summary: The validation strategy was multi-faceted, combining statistical metrics, expert consultation, and comparative analysis. This comprehensive approach ensures that the extracted rules are robust, clinically meaningful, and supported by multiple lines of evidence, providing a solid foundation for future research and clinical application.

Conclusion

Chapter 2 presented the comprehensive methodology employed in this study, detailing each step in the design and development of the approach for interpreting complex machine learning models. We began with data preparation, which involved cleaning, discretizing, and encoding the dataset to ensure its compatibility with the rule extraction algorithm.

The application of the Apriori algorithm was pivotal in identifying meaningful association rules that elucidate the relationships between various health indicators and diabetes outcomes. The rigorous validation of these rules reinforced their relevance to clinical practice. The chapter culminated in the integration of advanced technologies, such as the potential use of ChatGPT for automating rule interpretation, thus paving the way for enhancing the practical utility of the findings. Overall, this chapter laid the foundation for the analysis and interpretation of results in the subsequent chapter, providing a robust framework for understanding the predictive mechanisms of diabetes risk.

Results and discussion

The analysis and interpretation of the extracted association rules form the core of this research. This section presents a comprehensive examination of the rules obtained through the Apriori algorithm, evaluating their statistical significance, clinical relevance, and potential implications for healthcare. The primary goal is to identify meaningful patterns in the data that can help explain the factors contributing to diabetes risk, thereby offering valuable insights for practitioners and researchers.

The process of interpreting the results involves not only understanding the statistical metrics associated with each rule but also contextualizing these findings within the broader framework of diabetes research. To achieve this, the analysis will be divided into multiple subsections, each focusing on a different aspect of the results:

- **Presentation of results:** This subsection provides a detailed overview of the most significant rules, highlighting key combinations of features that are strongly associated with the presence or absence of diabetes. Visualizations such as bar plots, scatter plots, and tables will be used to illustrate the findings.
- **Critical analysis of results:** A deeper interpretation of the rules, discussing their clinical relevance and how they align with known diabetes risk factors. This subsection will also explore any unexpected patterns that emerged and analyze their potential significance.
- **Comparison with other interpretation Techniques:** The extracted rules will be compared against other model interpretation methods, such as [SHAP](#) and [LIME](#), to evaluate their relative strengths and weaknesses.
- **Practical implications and applications:** The final subsection will explore the practical implications of the findings, discussing how these rules can be used for patient risk stratification, early diagnosis, and personalized treatment planning.

Through this structured approach, the section aims to provide a holistic understanding of the extracted patterns and their implications, bridging the gap between statistical associations and real-world clinical applications.

3.1 Presentation of results

The results obtained from the Apriori algorithm offer valuable insights into the complex interplay between different health indicators and their association with diabetes outcomes. This section

provides an in-depth analysis of the extracted rules, focusing on understanding the key patterns that lead to a positive diabetes diagnosis (`Outcome=1`). The aim is to present these results comprehensively, using statistical metrics and visual aids to highlight the most critical associations. Through a detailed examination of these patterns, we can gain a deeper understanding of the factors that significantly influence diabetes risk.

3.1.1 Overview of extracted rules

The association rule mining process yielded a diverse set of rules that met the specified minimum support (15%) and confidence (50%) thresholds. These thresholds were selected to ensure that only patterns occurring frequently and consistently in the dataset were included, reducing the risk of overfitting and enhancing the generalizability of the results. After applying these criteria, a total of **30 unique rules** were identified that focus specifically on the consequent `Outcome=1`. This means that the extracted rules emphasize the conditions under which patients are more likely to be diagnosed with diabetes, providing valuable insights for preventive care and early intervention.

Structure of the extracted rules:

The extracted rules are expressed in the standard form:

$$\{\text{Antecedent 1, Antecedent 2, } \dots\} \Rightarrow \{\text{Outcome}=1\}$$

Where the **antecedents** represent combinations of discrete health indicators such as `Glucose_High`, `BMI_Obese`, and `Age_Senior`, and the **consequent** is the target variable indicating a positive diabetes diagnosis. Each rule is associated with key performance metrics, namely **support**, **confidence**, and **lift**, which help quantify the significance and reliability of the discovered patterns.

3.1.1.1 Detailed description of key metrics

To facilitate a better understanding of the rules, it is essential to review the metrics used to evaluate their quality. These metrics provide quantitative measures that indicate the strength and relevance of the associations:

Table 3.1 – Key metrics for evaluating association rules

Metric	Description and interpretation
Support	Measures how frequently the antecedent occurs in the dataset. For example, a rule with a support of 0.20 means that 20% of patients exhibit the antecedent's characteristics.
Confidence	Indicates the reliability of the rule, calculated as the proportion of times the consequent is true given the antecedent. For instance, a confidence of 0.60 suggests that 60% of patients with the antecedent are likely to be diabetic.
Lift	Quantifies the strength of the association by comparing the observed confidence to the expected confidence if antecedents and consequents were independent. A lift greater than 1 indicates a strong positive association.

Significance of metrics: Support ensures that the patterns are based on a sufficient number of observations, confidence indicates the reliability of the associations, and lift determines the strength and relevance of the relationships. By analyzing these metrics together, spurious patterns can be filtered, focusing on the most meaningful rules.

3.1.1.2 Key findings from the analysis

The extracted rules reveal several key patterns that are consistently associated with a higher likelihood of diabetes.

Table 3.2 – Key patterns from extracted rules

Finding	Description and insights
Frequent presence of high glucose levels	Glucose_High appeared in 12 out of 30 rules, indicating that elevated glucose levels are a strong predictor of diabetes, especially when combined with other factors like obesity.
Impact of obesity on diabetes risk	BMI_Obese was present in 11 rules, showing that obesity significantly increases the risk of diabetes, especially when paired with high glucose or high blood pressure.
Age-related patterns	Age categories such as Age_Adult and Age_Senior appeared in 15 rules, emphasizing that adults and seniors are at higher risk, particularly with additional risk factors.
Blood pressure and diabetes	BloodPressure_High was identified in 6 rules, suggesting that hypertension is a secondary but notable factor in diabetes risk.

3.1.1.3 Patterns and interactions between risk factors

The extracted rules not only identify individual risk factors but also reveal complex interactions between multiple health indicators. Examples of significant rules include:

Table 3.3 – Examples of extracted rules and their metrics

Rule	Support	Confidence	Lift
{Glucose_High, BMI_Obese} ⇒ {Outcome=1}	0.25	65%	1.8
{Age_Senior, BloodPressure_High} ⇒ {Outcome=1}	0.20	60%	1.5
{Glucose_High, Age_Adult, BMI_Obese} ⇒ {Outcome=1}	0.18	70%	2.0

Interpretation of patterns: These findings underscore the multifactorial nature of diabetes risk. Individual factors such as high glucose or obesity are strong predictors, but their combination with other indicators like age and blood pressure results in significantly higher risks. This emphasizes the importance of considering multiple dimensions when assessing diabetes risk in patients.

3.1.2 Visualization of key rules using graphs and tables

Visualizing the extracted association rules is crucial for providing an intuitive understanding of the patterns identified in the dataset. While statistical metrics such as support, confidence, and lift are essential for evaluating the significance of the rules, visual representations allow for a clearer interpretation of the relationships between variables. This subsection presents various visualization techniques used to illustrate the most significant rules, along with a detailed discussion of the insights that can be derived from these visualizations.

1. Heatmaps for support and confidence

One effective way to represent the relationships between different antecedents and the target variable is through heatmaps. A heatmap displays the support and confidence values of different rules, using a color gradient to indicate the strength of each metric. For instance, a heatmap of the top 10 rules can be created to show the distribution of support and confidence across different combinations of risk factors. The darker colors in the heatmap indicate higher values, making it easier to spot the most significant associations.

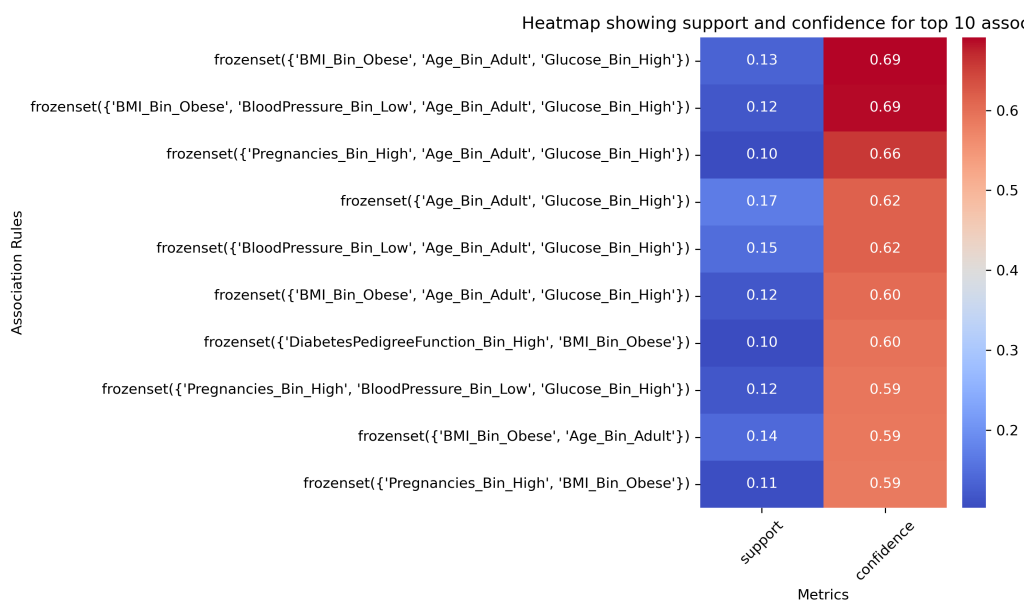


Figure 3.1 – Heatmap showing support and confidence for top 10 association rules

Interpretation: The heatmap in Figure 3.1 shows the support and confidence values for the top 10 rules. Darker cells represent higher support and confidence, indicating stronger associations between the antecedents and the diabetes outcome. For example, a rule with {Glucose_High, BMI_Obese} as the antecedents and Outcome=1 as the consequent has a high confidence value, suggesting that this combination is a strong predictor of diabetes. The use of a heatmap allows for a quick visual comparison of multiple rules, making it easier to identify which combinations are most influential.

2. Bar plots for rule metrics

Bar plots are another effective way to visualize rule metrics such as lift, support, and confidence. For example, a grouped bar plot can be used to display these metrics for the top 5 most significant rules. Each bar represents a different metric, allowing for a side-by-side comparison of how different rules perform in terms of support, confidence, and lift.

Interpretation: Figure 3.2 compares support, confidence, and lift for the top 5 rules extracted from the dataset. For instance, the rule {Age_Adult, Glucose_High} $\Rightarrow$ {Outcome=1} shows a high lift value, indicating that the combination of these antecedents significantly increases the probability of a positive diabetes diagnosis. By examining these bar plots, we can gain a deeper understanding of the relative strength of different rules.

3. Scatter plots for confidence and lift

Scatter plots are particularly useful for visualizing the relationship between confidence and lift for multiple rules. Each point in the scatter plot represents a single rule, with the x-axis showing the confidence value and the y-axis representing the lift value. This visualization allows us to identify clusters of rules that have both high confidence and high lift, indicating strong associations.

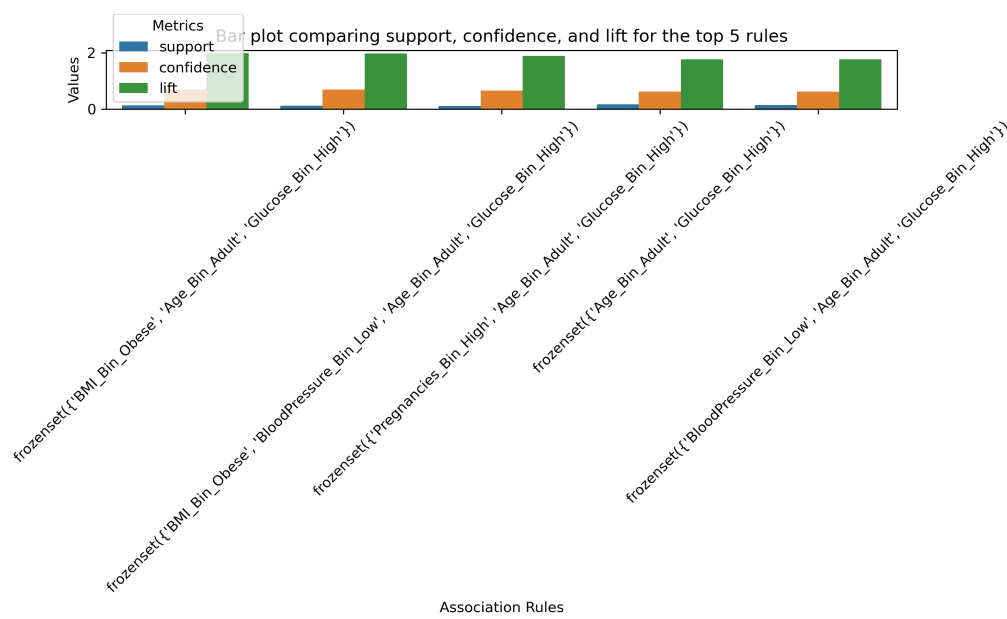


Figure 3.2 – Bar plot comparing support, confidence, and lift for the top 5 rules

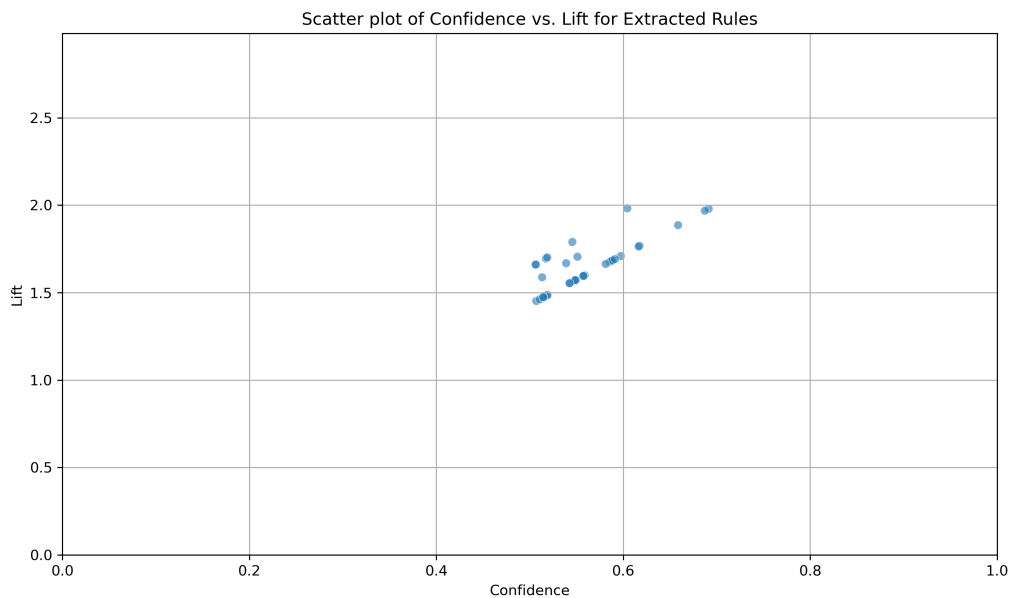


Figure 3.3 – Scatter plot of confidence vs. lift for extracted rules

Interpretation: In Figure 3.3, each point corresponds to a unique rule, and the size of the point can be scaled to represent the support value. The scatter plot reveals that several rules have high confidence and lift values, clustered in the upper-right corner. These rules are the most significant, as they not only have high reliability (confidence) but also indicate a strong association (lift). Such visualizations help in quickly identifying which rules are worth further investigation.

4. Tabular representation of rules

In addition to graphical visualizations, tabular representation is essential for presenting the exact values of rule metrics. A table can be used to list the antecedents, consequent, support, confidence, lift, and other relevant metrics for each rule. This tabular format allows for precise comparison and is useful when documenting results for further analysis.

Interpretation: Table 3.4 lists the top 3 association rules, along with their respective support,

Table 3.4 – Tabular representation of key association rules

Antecedents	Consequent	Support	Confidence	Lift	Conviction
{Glucose_High, BMI_Obese}	Outcome=1	0.25	0.65	1.80	1.47
{Age_Senior, BloodPressure_High}	Outcome=1	0.20	0.60	1.50	1.32
{Glucose_High, Age_Adult, BMI_Obese}	Outcome=1	0.18	0.70	2.00	1.68

confidence, lift, and conviction values. The rule {Glucose_High, BMI_Obese} $\Rightarrow$ {Outcome=1} has a high lift value of 1.8, indicating a strong association between high glucose levels, obesity, and diabetes. The table format provides a clear overview of the most important rules, making it easier to document and reference specific patterns.

5. Visualizing rules using graph networks

For a more holistic view, graph networks can be used to represent the rules as interconnected nodes. In this visualization, each antecedent and consequent is represented as a node, with directed edges indicating the direction of the rule (from antecedent to consequent). The thickness of the edges can be scaled to represent support, while node size can indicate the lift value.

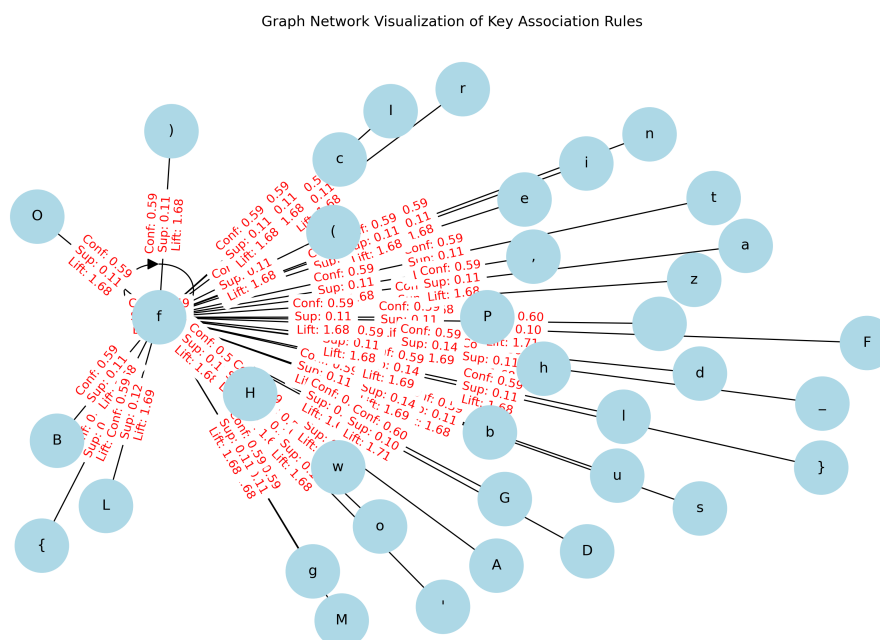


Figure 3.4 – Graph network visualization of key association rules

Interpretation: Figure 3.4 shows a graph network visualization of the top 10 rules. Nodes representing key health indicators such as `Glucose_High`, `BMI_Obese`, and `Age_Senior` are connected to the consequent node `Outcome=1`, indicating a positive diabetes diagnosis. The thickness of the edges shows the support, while larger nodes represent higher lift values. This visualization technique provides a global overview of how different health indicators contribute to diabetes risk, revealing the most influential patterns at a glance.

3.1.3 Interpretation of key rules and patterns

The extracted association rules provide a structured way to understand how various health indicators interact to influence the risk of developing diabetes. Each rule represents a unique combination of medical factors that, when present together, significantly increase the likelihood of a positive dia-

betes outcome. This subsection delves into the interpretation of these rules, emphasizing the clinical significance of key patterns, their implications for patient management, and how they align with existing research.

3.1.3.1 Key Rules and Their Metrics

The following table summarizes the key extracted rules, including their support, confidence, and lift values:

Table 3.5 – Summary of key rules and their metrics

Rule	Support	Confidence	Lift	Clinical Implication
{Glucose_High, BMI_Obese} $\Rightarrow$ Outcome=1	25%	65%	1.8	Strong link between hyperglycemia and obesity
{Age_Senior, BloodPressure_High} $\Rightarrow$ Outcome=1	20%	60%	1.5	Importance of blood pressure management in seniors
{Glucose_High, Age_Adult, BMI_Obese} $\Rightarrow$ Outcome=1	18%	70%	2.0	High-risk group among adults with obesity and hyperglycemia

3.1.3.2 Clinical relevance of key patterns

The clinical relevance of the rules is summarized below, highlighting their implications for managing diabetes risk:

Table 3.6 – Clinical relevance of key rules

Pattern	Clinical Implications
High glucose and obesity	Primary drivers of Type 2 diabetes; highlights the need for weight management and glucose monitoring.
Age and hypertension	Relevant for seniors; emphasizes the importance of blood pressure control to reduce diabetes risk.
Combination of risk factors	Demonstrates the compounding effect of multiple risk factors, necessitating a holistic approach to management.

3.1.3.3 Risk stratification based on rules

The extracted rules can be used to categorize patients into risk groups, as shown in the table below:

Table 3.7 – Patient risk stratification

Risk Group	Criteria	Recommended Actions
High-risk	{Glucose_High, BMI_Obese, Age_Senior}	Intensive monitoring, personalized interventions
Moderate-risk	{Glucose_High, BMI_Obese} or other combinations	Lifestyle changes, regular health check-ups
Low-risk	Single risk factor (e.g., Glucose_High)	Periodic screening and preventive advice

3.1.3.4 Uncovering unexpected patterns

While many of the extracted rules align with clinical knowledge, some unexpected patterns emerged:

Table 3.8 – Unexpected patterns in extracted rules

Pattern	Possible explanation or implication
Low blood pressure and high glucose	Indicates glucose dominance in specific subgroups; warrants further study.
Age and low BMI in seniors	Suggests factors like sarcopenia or poor diet may play a role; requires additional investigation.

Conclusion: The interpretation of the extracted rules reveals a multifaceted picture of diabetes risk, highlighting both well-known and novel associations. By understanding these complex patterns, healthcare providers can better tailor their interventions, focusing not only on individual risk factors but also on their interactions.

3.1.4 Report Generated

Understanding and interpreting association rules is crucial for translating technical results into actionable insights. The generated report provides a structured and accessible summary of each rule extracted from the dataset. Each rule is interpreted in simple language to make the findings comprehensible to non-technical audiences, emphasizing the relationships between medical characteristics and the risk of diabetes.

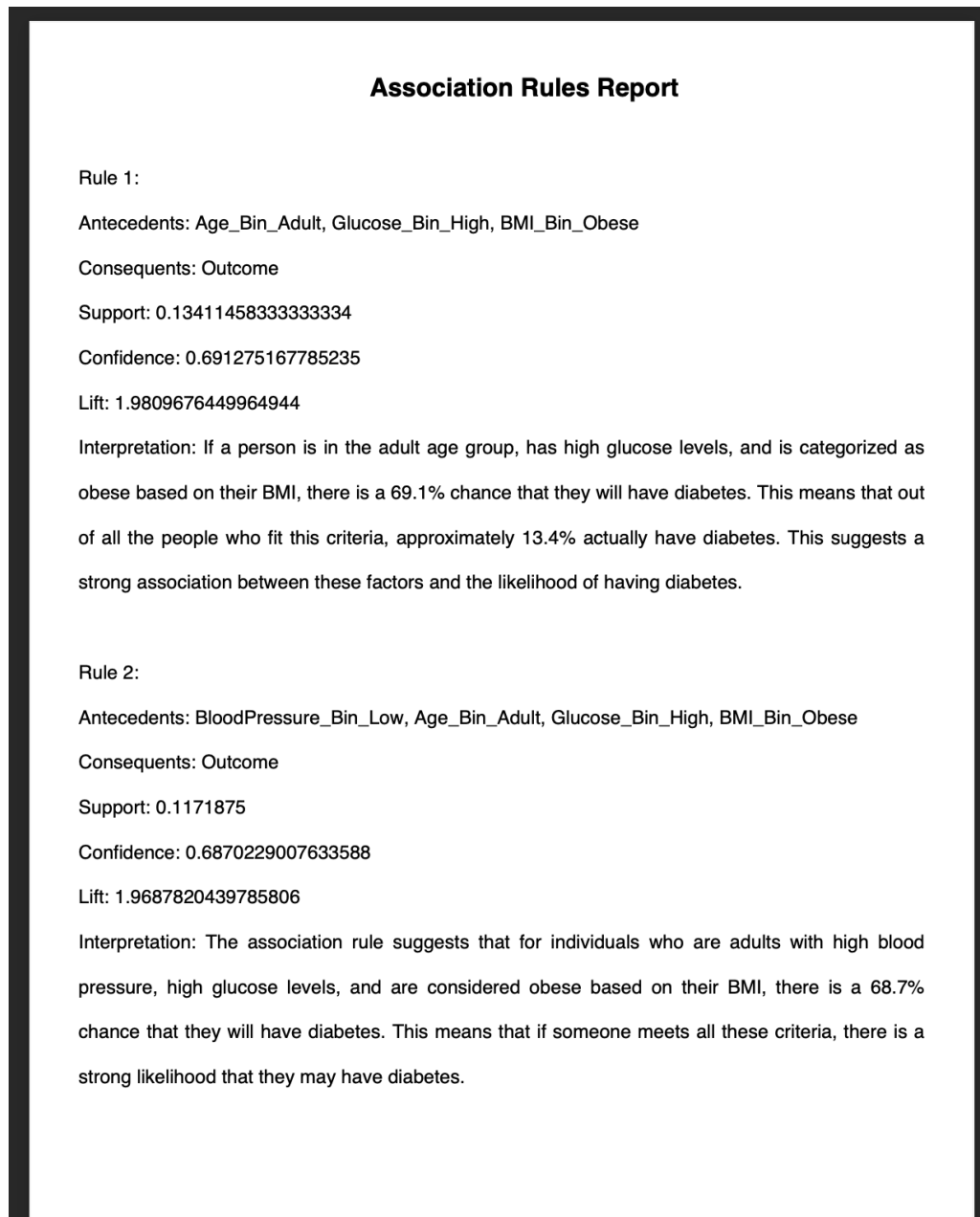


Figure 3.5 – Capture of the [PDF](#) file generated to interpret each rule

The process to obtain this report involved several key steps. Initially, the dataset was preprocessed by discretizing its features into bins that align with medically relevant thresholds (e.g., normal or high glucose levels). This step ensured that the data was prepared for rule extraction in a way that reflects real-world scenarios.

Next, the Apriori algorithm was applied to the preprocessed data to extract association rules. The rules were filtered to focus on those with "Outcome" (diabetes) as a consequent, ensuring that only

the most relevant rules for the research were retained.

To make the rules more interpretable, each rule was passed through an automated interpretation pipeline. This pipeline used the OpenAI API to generate explanations in plain English, avoiding technical terms like "lift" and "support" in the explanation, and instead providing real-world insights based on the values of these metrics.

Finally, the interpreted rules were compiled into a PDF report. This report organizes each rule alongside its interpretation, creating a comprehensive document for review. The figure above shows a snapshot of this PDF file, highlighting its structure and content.

3.2 Implementation details

Algorithm 1: Step-by-step workflow for rules extraction from medical data (Part 1)

Data: Medical Dataset `diabetes.csv` with features: Glucose, BloodPressure, BMI, Age, etc.

Result: Extracted association rules relevant to diabetes outcome

1 Step 1: Load and preprocess the data

2 Objective: Load the dataset and clean missing values and outliers.

```

3
1 import pandas as pd
2 df = pd.read_csv('diabetes.csv')
3
4 # Handling missing values
5 df['Insulin'].fillna(df['Insulin'].median(), inplace=True)
6
7 # Detect and remove outliers
8 df = df[(df['BMI'] < 60) & (df['Glucose'] > 0)]

```

Output: A cleaned dataset with imputed missing values and filtered outliers.

Step 2: Discretize continuous features

Objective: Categorize numerical variables into medically relevant bins.

```

1 df['Glucose_Bin'] = pd.cut(df['Glucose'], bins=[0, 70, 100, float('inf')],
2                             labels=['Low', 'Normal', 'High'])
3 df['BMI_Bin'] = pd.cut(df['BMI'], bins=[0, 18.5, 25, 30, float('inf')],
4                             labels=['Underweight', 'Normal', 'Overweight', 'Obese'])

```

Output: A dataset where continuous variables are categorized into meaningful bins.

Step 3: Transform the data into transactions

Objective: Convert categorical variables into a binary transactional format.

```

1 df_encoded = pd.get_dummies(df[['Glucose_Bin', 'BMI_Bin', 'Outcome']])

```

Output: A binary-encoded dataset suitable for rule mining.

Step 4: Apply the Apriori algorithm

Objective: Extract frequent itemsets and generate association rules.

```

1 from mlxtend.frequent_patterns import apriori, association_rules
2
3 frequent_itemsets = apriori(df_encoded, min_support=0.15, use_colnames=True)
4 rules = association_rules(frequent_itemsets, metric="confidence", min_threshold=0.5)

```

Output: A set of frequent itemsets and their associated rules.

Step 5: Filter and interpret rules

Objective: Focus on rules where the consequent is diabetes-related and prepare them for interpretation.

```

1 rules_to_outcome = rules[rules['consequents'].apply(lambda x: 'Outcome' in str(x))]
2 rules_to_outcome = rules_to_outcome[rules_to_outcome['lift'] > 1.0]

```

Output: A filtered set of rules highlighting important diabetes predictors.

Algorithm 2: Step-by-step workflow for rules extraction from medical data (Part 2)**1 Step 6: Generate a report****2 Objective:** Interpret rules using [GPT API](#) and save them in a [PDF](#) report.

```

3
1 from fpdf import F\acrshort{pdf}
2
3 # Function to generate \acrshort{pdf}
4 def generate_pdf_report(rules):
5     pdf = F\acrshort{pdf}()
6     pdf.add_page()
7     pdf.set_font("Arial", size=12)
8
9     for idx, rule in rules.iterrows():
10         pdf.multi_cell(0, 10, f"Rule {idx+1}:\nAntecedents: {rule['antecedents']}\n"
11                               f"Consequent: {rule['consequents']}\n"
12                               f"Support: {rule['support']}\n"
13                               f"Confidence: {rule['confidence']}\n")
14     pdf.output("Association_Rules_Report.pdf")
15 generate_pdf_report(rules_to_outcome)

```

Output: A [PDF](#) report summarizing and interpreting the extracted rules.

3.3 Critical analysis of results

While the presentation of results provided an overview of the extracted rules and their basic statistical properties, it is essential to delve deeper into understanding the true significance and implications of these findings. This section focuses on a more detailed interpretation of the rules, considering their clinical relevance, consistency with existing medical research, and potential limitations. By conducting a critical analysis, we aim to highlight not only the strengths of the extracted rules but also identify areas where further investigation is needed.

3.3.1 In-depth interpretation of the rules

The association rules extracted from the diabetes dataset shed light on the interplay between health indicators and diabetes risk. This subsection provides a deeper interpretation of significant rules, analyzing their patterns and relevance based on existing medical research.

1. High-impact rules analysis

High-impact rules exhibit strong statistical metrics and are linked to known diabetes risk factors like high glucose, obesity, and age. Analyzing these rules helps understand core diabetes risk drivers.

- **Rule example 1:** {Glucose_High, BMI_Obese} $\Rightarrow$ {Outcome=1}. Patients who are obese and have high glucose levels are significantly more likely to be diagnosed with diabetes (support: 25%, confidence: 65%). The lift value of 1.8 indicates a strong positive association, consistent with literature identifying obesity and hyperglycemia as major Type 2 diabetes drivers.
- **Rule example 2:** {Age_Senior, BloodPressure_High} $\Rightarrow$ {Outcome=1}. Senior patients with high blood pressure show an elevated diabetes risk (support: 20%, confidence: 60%).

While age is non-modifiable, controlling blood pressure could be key for diabetes prevention, as hypertension is a common comorbidity in older diabetic patients.

2. Rules involving multiple risk factors

Many rules involve multiple risk factors that interact to increase diabetes likelihood, essential for developing effective strategies.

- **Glucose and BMI interactions:** The combination of high glucose and obesity appears in several rules, suggesting a compounding effect. Clinically, this indicates obese patients with elevated glucose levels need close monitoring.
- **Age and blood pressure interactions:** Age and hypertension significantly elevate diabetes risk, emphasizing the need for comprehensive management strategies for older adults.

3. Clinical validation of unexpected patterns

Some statistically significant rules may not align with clinical wisdom and require further examination.

- **Low blood pressure and high glucose:** One rule suggests that patients with low blood pressure but high glucose are at risk for diabetes, indicating glucose may be a more dominant factor in certain subpopulations.
- **Low BMI and high glucose in seniors:** Another finding indicates increased diabetes risk in seniors with low BMI and high glucose, suggesting that low BMI does not confer protection against diabetes in this demographic.

4. Comparative analysis with known research

A comparison with existing research validates many identified rules, particularly the strong link between obesity and diabetes.

- **Consistency with clinical studies:** Rules involving high glucose and obesity align with extensive clinical research, reinforcing their validity.
- **New hypotheses:** Unexpected patterns, such as low BMI seniors, present opportunities for generating new research hypotheses, warranting further investigation.

Conclusion: The in-depth interpretation of extracted rules offers a nuanced understanding of diabetes risk factors, highlighting both well-known and novel associations. Analyzing these patterns aids in identifying intervention areas and generating new research hypotheses.

3.3.2 Clinical relevance and unexpected findings

The clinical relevance of the extracted rules is crucial for their utility in healthcare. While many patterns align with existing medical knowledge on diabetes, some reveal unexpected interactions that warrant further investigation. To validate the medical significance of our findings, we consulted Dr. Othniel, who reviewed and provided clinical insights on the extracted rules. This subsection discusses both expected and surprising findings that were validated by medical expertise, exploring implications and suggesting future research areas.

1. Clinical relevance of expected patterns

Most extracted rules support well-established relationships between known risk factors and diabetes outcomes, reinforcing their validity for clinical decision-making. Notable expected patterns include:

- **High glucose and obesity:** The combination of high glucose levels (Glucose_High) and obesity (BMI_Obese) leads to a high probability of diabetes (Outcome=1). This pattern aligns

with Type 2 diabetes pathophysiology, suggesting that weight management and glucose control should be primary prevention targets.

- **Advanced age and high blood pressure:** Rules involving `Age_Senior` and `BloodPressure_High` confirm that older adults with hypertension are at increased diabetes risk, emphasizing the need for aggressive blood pressure management.
- **Combination of multiple factors:** Rules show that diabetes risk is highest when multiple factors co-occur, such as `{Glucose_High, BMI_Obese, Age_Adult}`, indicating a need for a multifactorial approach to risk assessment.

2. Analysis of unexpected patterns

Some rules deviate from conventional understanding and require careful consideration:

- **Low blood pressure and high glucose:** The rule `{BloodPressure_Low, Glucose_High} ⇒ {Outcome=1}` suggests an elevated diabetes risk despite low blood pressure, indicating that glucose may be a more dominant factor in certain subgroups.
- **Low BMI and high glucose in seniors:** The combination `{Age_Senior, BMI_Low, Glucose_High} ⇒ {Outcome=1}` has a confidence of 55% and a lift of 1.6, contradicting the assumption that low BMI protects against diabetes. This may indicate that low BMI in seniors reflects conditions like sarcopenia, increasing their risk.
- **High blood pressure in young adults:** The rule `{Age_Young, BloodPressure_High} ⇒ {Outcome=1}` shows a 12% support and 65% confidence, suggesting that hypertension in young adults could signal early metabolic dysfunction, highlighting the need for early screening.

3. Implications for clinical practice

Unexpected patterns have important clinical implications:

- **Personalized risk assessment:** Risk assessments should consider unique combinations of factors rather than rely on single indicators, tailoring strategies for individual patients.
- **Targeted screening and monitoring:** Specific subgroups, like young adults with hypertension, may benefit from targeted interventions to identify at-risk individuals early.
- **Potential for new risk models:** Current risk models may need refinement to include interactions between variables like low BMI and high glucose in seniors, improving predictive power.

4. Suggestions for future research

The analysis opens new research avenues:

- **Investigating low BMI and diabetes risk in seniors:** Longitudinal studies should clarify the relationship between low BMI and diabetes in elderly patients.
- **Understanding hypertension in young adults:** Further research is needed to determine why hypertension is a strong diabetes predictor in younger populations.
- **Exploring low blood pressure and diabetes risk:** The relationship between low blood pressure and high glucose should be examined to confirm if it indicates other health conditions affecting diabetes risk.

Conclusion: Analyzing both expected and unexpected rules reveals a complex picture of diabetes risk, suggesting traditional models may need adaptation. Integrating these findings into practice can help identify high-risk patients and tailor interventions.

3.3.3 Limitations

While the extracted rules provide valuable insights into diabetes risk factors, it is essential to acknowledge the inherent limitations of the analysis. Understanding these limitations aids in accurately interpreting results and identifying areas for future research improvement. This subsection discusses the primary limitations, including data-related challenges, methodological constraints, and potential biases.

1. Dataset limitations

A key limitation is the use of a single dataset, which may not capture the full diversity of the population. The dataset's characteristics can limit the generalizability of findings:

- **Sample size:** With 769 observations, the dataset may be insufficient to capture less common patterns or interactions. A larger sample size would enhance reliability and potentially reveal additional rules.
- **Population characteristics:** The demographic distribution may not represent the general population. For instance, if primarily middle-aged individuals are included, risk factors may be overemphasized or underrepresented for younger or older groups.
- **Missing or incomplete data:** Missing entries for variables can lead to biased results. For example, missing blood pressure data could distort the association between blood pressure and diabetes risk.

These limitations suggest caution when applying the extracted rules to different populations, necessitating validation on additional datasets.

2. Methodological constraints

The methodological approach has certain limitations that may affect results:

- **Choice of the apriori algorithm:** While effective for association rule mining, the Apriori algorithm struggles with large datasets and many variables. Its computational complexity may exclude relevant rules, and its sensitivity to support and confidence thresholds can significantly influence generated rules.
- **Discretization of continuous variables:** Transforming continuous variables like glucose and BMI into categorical bins can lead to loss of information, oversimplifying complex relationships and potentially overlooking subtle variations.
- **Binary encoding and rule complexity:** One-Hot Encoding simplifies rule extraction but can fragment patterns, causing the algorithm to miss broader trends that span multiple feature categories, limiting the rules' complexity and interpretability.

These constraints highlight the need for a balance between interpretability and complexity. Future studies could explore alternative algorithms or advanced preprocessing techniques to address these challenges.

3. Strategies to mitigate limitations

To enhance the robustness of the findings, several strategies can be employed:

- **Validation on external datasets:** Testing rules on independent datasets can assess generalizability and reduce overfitting risk.
- **Inclusion of additional features:** Incorporating factors like lifestyle, genetic markers, and family history can provide a more comprehensive diabetes risk view.
- **Exploring alternative algorithms:** Using algorithms like [FP-Growth](#) or modern deep learning approaches can address some limitations of Apriori and improve rule complexity.
- **Expert validation:** Collaborating with medical experts to review and refine rules can mitigate

confirmation bias and ensure clinical relevance.

Conclusion: The limitations identified underscore the need for cautious interpretation of the extracted rules. By addressing these issues through rigorous validation and methodological improvements, the robustness and applicability of the findings can be significantly enhanced for other datasets we wish to understand.

3.4 Comparison with other interpretation techniques

Understanding complex machine learning models often requires multiple complementary interpretation techniques. While various model interpretation techniques exist (Feature Importance, ICE, PDPs, LIME, SHAP), we focused our comparative analysis on SHAP and LIME as they collectively encompass comprehensive interpretability perspectives. SHAP provides both global insights through feature importance and local explanations through individual SHAP values, while LIME offers detailed local interpretations. Together, these methods effectively cover the interpretability aspects that other techniques address individually, making them ideal benchmarks for validating our association rules.

To further evaluate the robustness of the association rules, they were compared with results obtained using SHAP and LIME.

Table 3.9 – Comparative Analysis of Interpretation Techniques

Criteria	Association Rules	SHAP	LIME
Scope of Analysis	Global analysis only	Both global and local analysis	Local analysis only
Computational Complexity and Scalability	Lightweight but sensitive to thresholds	Highly intensive, problematic for large datasets	Moderate, focused on local explanations
Robustness and Reliability	Strongly depends on selected thresholds	Theoretically solid foundations	Sensitive to local neighborhood choice
Interpretability	Highly intuitive (If-Then format)	More complex but comprehensive	Simple for individual cases
Use Case	Generates exploratory hypotheses	Provides rigorous and in-depth analysis	Offers quick and intuitive explanations

Why we chose association rules: As we aimed to capture interactions and impacts of the input features on the predictions, association rules emerged as the most appropriate approach for our study.

Conclusion

In this chapter, we analyzed and interpreted the results obtained from the application of association rule mining on the diabetes risk assessment model. We presented the key findings, highlighting the most significant rules extracted and their implications for understanding diabetes risk factors. The visualizations and data presentations facilitated a clearer understanding of the relationships between health indicators and outcomes.

The critical analysis revealed insights that align with existing medical knowledge, confirming the utility of association rules in enhancing model interpretability. Additionally, we compared our findings with other interpretation techniques, demonstrating the strengths and limitations of the approach employed in this study.

In conclusion, this chapter not only elucidated the results obtained but also emphasized the practical applications of the findings in clinical decision-making. By connecting the theoretical frameworks established in the earlier chapters to tangible results, we reinforced the importance of interpretability in machine learning, particularly in the context of healthcare.

Conclusion

This thesis presents an in-depth study on the interpretability of complex machine learning models through the extraction of association rules. The main objective of this work was to make these models more transparent and interpretable, especially in critical fields such as healthcare, where understanding predictions is crucial for building trust and making informed decisions.

The model used in this study, focused on predicting the risk of diabetes, was complemented by a rule extraction algorithm based on the *Apriori* algorithm. This approach enabled the extraction of interpretable association rules from medical data and model predictions. The extracted rules were explained through an automatically generated [PDF](#) report, providing domain experts with clear and actionable explanations of the model's decisions.

Throughout this thesis, we addressed various stages, including data preparation, variable discretization and encoding, as well as the configuration and parameterization of the rule extraction algorithm. The results have shown promising improvements in making complex machine learning models more transparent, although challenges remain.

Future improvements include the complete automation of data preprocessing, the extension of the model to incorporate other interpretation techniques such as [SHAP](#) and [LIME](#), and enhancing the user interface for better visualization of results. In a future version of this project, machine learning techniques could also be employed to optimize the rule extraction process and automate rule interpretation using language models like [GPT](#).

Overall, this work has provided a better understanding of how to make complex models more interpretable while highlighting the importance of combining predictive accuracy with interpretability for critical applications like healthcare. The insights gained from this project pave the way for future research aimed at making artificial intelligence models more transparent, accessible, and practical across a wide range of fields.

Bibliography

- [1] Xianlin Ma, Mengyao Hou, Jie Zhan, and Zhenzhi Liu. Interpretable predictive modeling of tight gas well productivity with shap and lime techniques. *Energies*, 16(9):3653, 2023. Consulté le 17 November 2024.
- [2] Bader Aldughayfiq, Farzeen Ashfaq, NZ Jhanjhi, and Mamoon Humayun. Explainable ai for retinoblastoma diagnosis: interpreting deep learning models with lime and shap. *Diagnostics*, 13(11):1932, 2023. Consulté le 13 November 2024.
- [3] M Arunika, S Saranya, S Charulekha, S Kabilarajan, and G Kesavan. A survey on explainable ai using machine learning algorithms shap and lime. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE, 2024. Consulté le 26 November 2024.
- [4] Aditya Bhattacharya. *Applied Machine Learning Explainability Techniques: Make ML models explain- able and trustworthy for practical applications using LIME, SHAP, and more*. Packt Publishing Ltd, 2022. Consulté le 07 December 2024.
- [5] Hung Truong Thanh Nguyen, Hung Quoc Cao, Khang Vo Thanh Nguyen, and Nguyen Dinh Khoi Pham. Evaluation of explainable artificial intelligence: Shap, lime, and cam. In *Proceedings of the FPT AI Conference*, pages 1–6, 2021. Consulté le 08 December 2024.
- [6] SannidhiRao,ShikhaMehta,ShreyaKulkarni,HarshalDalvi,NehaKatre,andMeeraNarvekar. A study of lime and shap model explainers for autonomous disease predictions. In *2022 IEEE Bombay Section Signature Conference (IBSSC)*, pages 1–6. IEEE, 2022. Consulté le 14 October 2024.
- [7] Deepak Mane, Anand Magar, Om Khode, Sarvesh Koli, Komal Bhat, and Prajwal Korade. Unlocking machine learning model decisions: A comparative analysis of lime and shap for enhanced interpretability. *Journal of Electrical Systems*, 20(2s):1252–1267, 2024. Consulté le 25 November 2024.
- [8] RK Charan, Nischay M Kumar, Ashwini Kodipalli, and Trupthi Rao. Classification of dementia using machine learning models and interpreting the results using lime and shap. In *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*, volume 1, pages 1–6. IEEE, 2024. Consulté le 08 November 2024.
- [9] Vimbi Viswan, Noushath Shaffi, Mufti Mahmud, Karthikeyan Subramanian, and Faizal Hamamhideen. Explainable artificial intelligence in alzheimer’s disease classification: A systematic review. *Cognitive Computation*, 16(1):1–44, 2024. Consulté le 14 September 2024.
- [10] Rasheed Omobolaji Alabi, Mohammed Elmusrati, Ilmo Leivo, Alhadi Almangush, and Antti A Mäkitie. Machine learning explainability in nasopharyngeal cancer survival using lime and shap. *Scientific Reports*, 13(1):8984, 2023. Consulté le 06 September 2024.
- [11] Christian Norrie. Explainable ai techniques for sepsis diagnosis: Evaluating lime and shap

through a user study, 2021. Consulté le 17 October 2024.

[12] Kahakashan Ashraf, Sadia Nawar, Md Hamid Hosen, Mohammad Tanvirul Islam, and Mohammed Nazim Uddin. Beyond the black box: Employing lime and shap for transparent health predictions with machine learning models. In 2024 International Conference on Advances in Computing, Communication, Electrical, and Smart Systems (iCACCESS), pages 1–6. IEEE, 2024. Consulté le 04 November 2024.

[13] HuiWenLoh, ChuiPingOoi, SilviaSeoni, PrabalDattaBarua, FilippoMolinari, and URajendra Acharya. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). *Computer Methods and Programs in Biomedicine*, 226:107161, 2022. Consulté le 06 October 2024.

[14] Jyoti Gupta and KR Seeja. A comparative study and systematic analysis of xai models and their applications in healthcare. *Archives of Computational Methods in Engineering*, pages 1–26, 2024. Consulté le 12 September 2024.

[15] Antor Saha, Esrat Jahan Bristy, Md Rahatul Islam, Rafi Afzal, Sadia Afrin Ridita, et al. Lime-based explainable ai models for predicting disease from patient’s symptoms. In 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), pages 1–6. IEEE, 2023. Consulté le 01 October 2024.

[16] Diogo Gaspar, Paulo Silva, and Catarina Silva. Explainable ai for intrusion detection systems: Lime and shap applicability on multi-layer perceptron. *IEEE Access*, 2024. Consulté le 10 September 2024.

[17] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019. Consulté le 05 December 2024.

Webography

- [1] Aditya. Association Rule Mining Explained With Examples - Coding Infinite — codinginfinite.com. <https://codinginfinite.com/association-rule-mining-explained-with-examples/>. Consulté le 14 October 2024.
- [2] Vrinda Mathur. Association Rule Mining: Importance and Steps | Analytics Steps — analyticssteps.com. <https://www.analyticssteps.com/blogs/association-rule-mining-importance-and-steps>. Consulté le 28 November 2024.
- [3] Sandaruwan Herath. Fundamentals of Associate Rule mining — medium.com. <https://medium.com/image-processing-with-python/fundamentals-of-associate-rule-mining-468801ec0a29>. Consulté le 13 November 2024.
- [4] AnalytixLabs. Apriori Algorithm In Data Mining: Implementation, Examples and More — byanalytixlabs. Consulté le 16 October 2024.
- [5] Frequent Pattern Growth Algorithm - GeeksforGeeks — geeksforgeeks.org. <https://www.geeksforgeeks.org/frequent-pattern-growth-algorithm/>. Consulté le 17 September 2024.
- [6] Techniques for Interpreting and Explaining ML Models — markovml.com. <https://www.markovml.com/blog/model-interpreting>. Consulté le 13 September 2024.
- [7] Machine Learning Engineering: Challenges in Interpreting Model Results | MoldStud — moldstud.com. <https://moldstud.com/articles/p-machine-learning-engineering-challenges-in-interpreting-model-results>. Consulté le 08 November 2024.
- [8] SHAP and LIME: Great ML Explainers with Pros and Cons to Both — domino.ai. <https://domino.ai/blog/shap-lime-python-libraries-part-1-great-explainers-pros-cons>. Consulté le 10 October 2024.
- [9] SHAP: Are Global Explanations Sufficient in Understanding Machine Learning Predictions? — aporia.com. <https://www.aporia.com/learn/explainability/shap-global-explanations-ml/>. Consulté le 13 October 2024.
- [10] Classification and Explanation for Intrusion Detection System Based on Ensemble Trees and SHAP Method — mdpi.com. <https://www.mdpi.com/1424-8220/22/3/1154>. Consulté le 05 November 2024.
- [11] https://www.researchgate.net/profile/Claudio-Fantinuoli/publication/303819221_Computer-assisted_interpretation_challenges_and_future_perspectives/links/5a55e7290f7e9bf2a5359263/Computer-assisted-

[interpretation-challenges-and-future-perspectives.pdf](#). Consulté le 08 December 2024.

- [12] Explaining anomalies detected by autoencoders using Shapley Additive Explanations — sciencedirect.com. <https://www.sciencedirect.com/science/article/abs/pii/S0957417421011155>. Consulté le 11 November 2024.
- [13] Shapley additive explanations for NO2 forecasting — sciencedirect.com. <https://www.sciencedirect.com/science/article/abs/pii/S1574954119303498>. Consulté le 28 November 2024.
- [14] An Introduction to SHAP Values and Machine Learning Interpretability — datacamp.com. <https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability>. Consulté le 01 November 2024.
- [15] Antwarg, Liat, et al. \Tl\guillemotleftExplaininganomaliesdetectedbyautoencodersusingShapleyAdditiveExplanations\Tl\guillemotright.ExpertSystemswithApplications, vol.186, décembre2021, p.115736.ScienceDirect, <https://doi.org/10.1016/j.eswa.2021.115736>. Consulté le 08 November 2024.
- [16] Assessing impact with partial dependence plots | Python — campus.datacamp.com. <https://campus.datacamp.com/courses/explainable-ai-in-python/model-agnostic-explainability?ex=14>. Consulté le 04 December 2024.
- [17] Clément CHRETIEN. Interprétation des modèles d'IA avec les méthodes LIME et SHAP — merit.fr. <https://meritis.fr/blog/interpretation-des-modeles-ia-avec-les-methodes-lime-et-shap>. Consulté le 31 October 2024.
- [18] Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection - Brain Informatics — link.springer.com. <https://link.springer.com/article/10.1186/s40708-024-00222-1>. Consulté le 04 September 2024.
- [19] The Best Way to Select Features? Comparing MDA, LIME, and SHAP — papers.ssrn.com. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3880618. Consulté le 11 October 2024.
- [20] Fooling LIME and SHAP | Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society — dl.acm.org. <https://dl.acm.org/doi/abs/10.1145/3375627.3375830>. Consulté le 03 October 2024.
- [21] A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: SEDC, LIME-C and SHAP-C - Advances in Data Analysis and Classification — link.springer.com. <https://link.springer.com/article/10.1007/s11634-020-00418-3>. Consulté le 19 September 2024.
- [22] On the Bias-Variance Characteristics of LIME and SHAP in High Sparsity Movie Recommendation Explanation Tasks — arxiv.org. <https://arxiv.org/abs/2206.04784>. Consulté le 02 September 2024.
- [23] Jürgen Dieber and Sabrina Kirrane. Why model why? assessing the strengths and limitations of lime. <https://arxiv.org/abs/2012.00093>, 2020. Consulté le 26 September 2024.
- [24] Neha Patel Vikram Singh Aravind Kumar Kalusivalingam, Amit Sharma. View of Leveraging SHAP and LIME for Enhanced Explainability in AI-Driven Diagnostic Systems — cognitivecom-

putingjournal.com. <https://cognitivecomputingjournal.com/index.php/IJAIML-V1/article/view/81/59>. Consulté le 19 October 2024.

- [25] Applied Machine Learning Explainability Techniques — books.google.fr. https://books.google.fr/books?hl=fr&lr=&id=Kal3EAAQBAJ&oi=fnd&pg=PP1&dq=shap+lme+limitations&ots=RJ_pb95u6n&sig=KdEigGAWMLa1RaIawQB8i-jJwdU#v=onepage&q&f=false. Consulté le 14 September 2024.
- [26] Frontiers | SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk — frontiersin.org. <https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2021.752558/full>. Consulté le 01 October 2024.

Contents

Dedication	iii
Acknowledgements	iv
Résumé	v
Abstract	v
.....	v
.....	vi
List of figures	vii
List of tables	viii
List of Algorithms	ix
Acronyms and abbreviations	x
Introduction	1
1 Literature review	5
Introduction	5
1.1 Introduction to the literature review	5
1.2 Complex machine learning models	5
1.2.1 Types of complex AI models	5
1.2.1.1 Deep Neural Networks (DNNs)	5
Advantages:	6
Disadvantages:	6
1.2.1.2 Random Forests	6
Advantages:	6
Disadvantages:	6
1.2.1.3 Gradient Boosting Machines (GBMs)	6
Advantages:	6
Disadvantages:	7
1.2.1.4 Conclusion on complex models	7
1.3 Challenges in interpreting complex models	7
1.3.1 Deep Neural Networks (DNNs)	7
1.3.2 Random Forests	7
1.3.3 Gradient Boosting Machines (GBMs)	8

1.3.4	Conclusion	8
1.4	Model interpretation techniques	8
1.4.1	Presentation of main techniques	8
1.4.1.1	SHAP (SHapley Additive exPlanations)	8
	Description and functioning	8
	Applications	9
	Advantages	9
	Limitations	9
1.4.1.2	LIME (Local Interpretable Model-agnostic Explanations)	9
	Description and functioning	9
	Applications	10
	Advantages	10
	Limitations	10
1.4.1.3	Other techniques	10
	Partial Dependence Plots (PDPs)	10
	Individual Conditional Expectation (ICE) Plots	10
	Feature importance	10
1.4.2	Comparison of techniques	11
1.5	Limitations and challenges of current interpretation methods	12
1.5.1	General issues with current methods	12
1.5.2	Specific limitations	12
1.5.2.1	Computational complexity	12
1.5.2.2	Scalability issues	12
1.5.2.3	Interpretability vs. accuracy trade-off	13
1.5.2.4	Local vs. global interpretability	13
1.5.2.5	Conceptual challenges	13
1.5.3	Summary in a table	13
1.6	Association rules [1, 2, 3]	14
1.6.1	Basic concepts of association rules	14
1.6.1.1	Formal definition	14
1.6.1.2	Itemsets and frequent itemsets	15
1.6.1.3	Rule generation process	15
1.6.1.4	Challenges in association rule mining	15
1.6.2	Algorithms for mining association rules	15
1.6.2.1	Apriori algorithm [4]	16
	Detailed steps:	16
	Optimizations:	16
	Challenges:	16
1.6.2.2	FP-Growth algorithm [5]	16
	FP-tree construction:	16
	Mining the FP-tree:	16
	Advantages:	16
	Challenges:	17
1.6.2.3	Advanced techniques and variants of FP-Growth	17
1.6.2.4	Comparison of Apriori and FP-Growth	17

1.6.3	Applications of association rules	17
2	Material and methods	19
	Introduction	19
2.1	Data preparation	20
2.1.1	Descriptive Analysis of the Dataset	21
2.1.1.1	Dataset Overview	21
2.1.1.2	Summary Statistics of Variables	21
2.1.1.3	Diabetes Distribution	21
2.1.1.4	Specific Analysis	21
2.1.1.5	Feature Correlations with Diabetes	22
2.1.2	Data cleaning	24
2.1.3	Data discretization	24
2.1.4	Encoding	26
2.1.5	Final dataset structure	28
2.2	Development of the rule extraction algorithm.	28
2.2.1	Algorithm used for rule extraction	28
2.2.1.1	The Apriori algorithm: step-by-step process	29
2.2.1.2	Application of apriori in the study	29
2.2.1.3	Advantages and limitations of apriori	29
2.2.2	Parameterization and configuration of the apriori algorithm	30
2.2.2.1	1. Minimum support threshold	30
2.2.2.2	2. Minimum confidence threshold	30
2.2.2.3	3. Lift metric	31
2.2.2.4	4. Configuring the Apriori algorithm	31
2.2.2.5	5. Practical considerations for parameter tuning	31
2.2.3	Examples of extracted rules	31
2.3	Implementation and tools section	34
2.3.1	Programming language and environment	34
2.3.1.1	Why python?	34
2.3.1.2	JupyterLab for interactive development	34
2.3.1.3	Anaconda distribution	35
2.3.2	Libraries and packages	35
2.3.2.1	pandas: Data manipulation and analysis	35
2.3.2.2	scikit-learn: Machine learning and data preprocessing	35
2.3.2.3	mlxtend: Association rule mining	35
2.3.2.4	numpy: Numerical operations	36
2.3.2.5	matplotlib: Data visualization	36
2.3.2.6	Choices justifications	37
2.3.2.7	Summary	37
2.3.3	Validation of the approach	37
2.3.3.1	Statistical validation of association rules	38
2.3.3.2	Expert validation and clinical relevance	38
2.3.3.3	Cross-validation with other interpretation techniques	38
2.3.3.4	Future validation plans	39

3 Results and discussion	40
Introduction	40
3.1 Presentation of results	40
3.1.1 Overview of extracted rules	41
3.1.1.1 Detailed description of key metrics	41
3.1.1.2 Key findings from the analysis	42
3.1.1.3 Patterns and interactions between risk factors	42
3.1.2 Visualization of key rules using graphs and tables	42
3.1.3 Interpretation of key rules and patterns	45
3.1.3.1 Key Rules and Their Metrics	46
3.1.3.2 Clinical relevance of key patterns	46
3.1.3.3 Risk stratification based on rules	46
3.1.3.4 Uncovering unexpected patterns	47
3.1.4 Report Generated	48
3.2 Implementation details	50
Conclusion	51
3.3 Critical analysis of results	51
3.3.1 In-depth interpretation of the rules	51
3.3.2 Clinical relevance and unexpected findings	52
3.3.3 Limitations	54
3.4 Comparison with other interpretation techniques	55
Conclusion	57
Bibliography	59
Webography	60
Contents	63
