EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

Analysis of Data Mining Techniques and Algorithms on Diabetes Dataset

Youssef FAKIR[1], Salim KHALIL[1], Weam FAKIR[1]

[1]Department of Computer Science, Faculty of science and Technology, University of Sultan Moulay Slimane, 23000, Beni Mellal, Morocco

Abstract

The fundamental goal of this work is to prepare and carry out diabetes prediction using various Machine Learning techniques and conduct output analysis of those techniques to find the best classifier with the highest accuracy. This study use the Pima Indian Diabetes Dataset and applied the Machine Learning classification methods like Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR) for diabetes prediction. The performance of each algorithm is analysed to determine the one with the best accuracy. The dataset includes details like pregnancies, glucose levels, blood pressure, and other important health information. The focus of this study is to unify FP-Growth algorithm with ML algorithm in order to predict diabetes. The FP-Growth is used to extract the frequent items for data pre-processing before prediction. LR algorithm stands out with high accuracy, showing promise in predicting type 2 diabetes when using the risk factors identified by FP-Growth algorithm. The results help guide future research and make it easier to choose the best algorithms, especially ones that are fast, for medical decision support systems. LR algorithm stands out with high accuracy, showing promise in predicting type 2 diabetes when using the risk factors identified by FP-Growth algorithm.

Keywords: Data Mining Technics, Association Rule Mining Algorithms Diabetes, FP-Growth, SVM, Logistic Regression, Random Forest.

Received on 07 March 2024, accepted on 26 September 2025, published on 20 October 2025

doi: 10.4108/eetpht.11.5317

*Corresponding author. Email: info.dec07@yahoo.fr

1. Introduction

Globally, an estimated 422 million individual's grapple with the repercussions of diabetes, with a notable concentration observed in less economically developed countries, as reported by the World Health Organization (WHO). Recent data from 2023, provided by WHO, paints a stark picture with an annual death toll of approximately 1.5 million people attributed to complications stemming from diabetes. The rising prevalence of diabetes on a worldwide scale presents an increasingly formidable challenge to public health. Type 2 diabetes, predominantly affecting adults, emerges as the most widespread form. In this variant, the body encounters difficulties in either effectively utilizing insulin or producing an adequate amount, leading to elevated blood sugar levels. Over the past three decades, the incidence of type 2 diabetes has displayed a concerning upward trajectory, affecting diverse populations across countries with varying income levels.

Concurrently, another distinct type, known as type one diabetes or juvenile diabetes, is characterized by the pancreas's diminished or complete lack of insulin production. The coexistence of these two diabetes types contributes to the intricate landscape of global health challenges, underscoring the imperative need for comprehensive strategies to address the multifaceted dimensions of this pervasive and escalating health concern [1]. Various types of diabetes, which can be broadly categorized into four main types. These include:

- Type 1 (T1DM): is a chronic autoimmune condition characterized by insulin deficiency, resulting in elevated blood sugar levels (hyperglycaemia). Over the past 25 years, there has been significant progress in our understanding of type 1 diabetes, encompassing diverse aspects including its genetic factors, epidemiology, immune and β-cell phenotypes, and the overall impact of the disease [2].
- Type 2 (T2DM): arises from insulin resistance, a condition wherein cells demonstrate insufficient

responsiveness to insulin. As the disease advances, there may also be a progression towards insulin insufficiency. Previously, terms such as 'non-insulin-based diabetes mellitus' or 'adult-induced diabetes' have been utilized to characterize this condition [3].

- Gestational diabetes mellitus (GDM) [4]: characterized by elevated blood sugar levels first identified during pregnancy, represents the most prevalent medical complication during gestation. On a global scale, GDM affects approximately 15% of pregnancies, contributing to roughly 18 million births each year.

Our study expands upon existing research by conducting a thorough comparative analysis of classification techniques and association rule mining algorithms tailored specifically to diabetes datasets. While previous studies have often focused on individual algorithms or limited comparisons, our work provides a comprehensive examination of three widely used classification methods—SVM, LR, and RF— [5, 6, 7, 8, 9, 10] in the context of diabetes prediction. By assessing performance metrics and computational efficiency across these techniques, we offer valuable insights for researchers and practitioners aiming to make informed decisions regarding algorithm selection.

The structure of this article is designed to systematically delve into the realm of data mining within the context of diabetes analysis. It begins with a related work section, which provides a comprehensive overview of existing research, offering insights into prior approaches and findings. The methodology section outlines the steps taken to analyse the diabetes dataset and algorithms employed. The results and discussion section present the outcomes of the analysis, accompanied by thorough interpretation and discussion of the findings. Finally, the conclusion synthesizes the key insights gleaned from the study, underlining its implications for both the field of data mining and diabetes research.

2. State-of-the-Art

DM generally has two categories of models, the first one is the predictive category, and the second one is descriptive. Each one of these classes has various techniques, for the predictive category we have regression, classification, prediction and time series analysis [9, 10], in the other hand the major techniques of the descriptive category are clustering, association rules, discover sequences and summary analysis [11, 12].

During the past decades, various DM techniques for diabetes detection are reviewed and discussed [13, 14, 15]. In [16], a review of the application of DM techniques for diabetes, as well as the corresponding data sets, methods, software, and technologies, is carried out. Based on this review, it is concluded that DM has a key role and bright research future in the field of glycemic control. DM is used to extract valuable information from diabetes data, which ultimately helps diabetic patients in the management of their glycemic control. Likewise, in [17] a survey is conducted on the application of different DM techniques, including Artificial Neural Network (ANN), for the prediction and classification of diabetes. The survey shows that ANN outperforms the rest of the techniques with 89% of prediction accuracy. In [18] the author's presents a survey on diabetes detection, classification and prediction by evaluating different schemes on parameters like, algorithm/models, input data type, etc. they conclude that a data pre-processing is needed for accurate detection, classification and prediction. In a review of existing literature, the most used models for diabetes prediction are SVM, RF, DT, and ANN. These Machine Learning (ML) algorithms were used on small dataset. Hybrid models and ensemble methods have also been explored [19], which can further improve the predictive performance of the models. Finally, other studies have explored the use of DM such as association's rules but most researcher's work have been worked on small dataset.

3. Proposed method

The main object of this paper is the prediction of diabetes. Three DM algorithms i.e. LR, RF and SVM are applied on a massive dataset after extracting the association rules by using FP-Growth. FP-Growth is one of the most effective algorithms for mining frequent item sets, however when the data is large, this method is constrained by the resource of a single computer and cannot complete the computing jobs in a fair amount of time.

Association rules are extracted by looking for frequents item sets [20]. Frequent item set mining stands for finding frequent items associations, correlations or causative structures (the if/then form) within groups of items or objects in databases such as transaction databases and other kind of repositories that bring information. The first task is to find all subsets of items that occur mutually in numerous transactions. The second task consist to find all the rules that associate the presence of one set of items with the presence of another set of items in the transaction database. In this work, multiple techniques were used for diabetes prediction, starting from significant attributes discovery via Principal Component Analysis (PCA) method, then mining frequent pattern and association rules extraction using FP-Growth in order to discover further knowledge from the association between attributes. Furthermore, a collection of ML algorithms is used to predict diabetes from a large data set of diabetes. The flow chart of the proposed model for diabetes prediction is illustrated in Figure 1.

4. Selected ML algorithms

4.1. FP-Growth Algorithm

The process of mining frequent itemsets in FP-Growth adopts a divide-and-conquer approach. Initially, FP-Growth condenses the database, transforming frequent itemsets into an FP-tree while preserving association information. Association rules has been oriented towards two objectives:

- Determine the set of frequent itemsets that appear in the database with support greater than or identical to minsup (minimum support) where support is the proportion of transactions that contain the itemset, calculated using the formula:

Supp (X) = (Number of transactions containing X) / (Total number of transactions) (1)

- Generate the set of associative rules, from these frequent itemsets, with a confidence measure greater than or identical to minconf. This measures the strength of the association between two itemsets. It is the conditional probability of finding Y in a transaction, given that X is present. The formula is:

Conf(X => Y) = (Number of transactions containing X and Y) / (Number of transactions containing X) (2)

where X and Y are the itemsets for which the confidence of the rule X => Y (meaning "If X, then Y") is calculated.

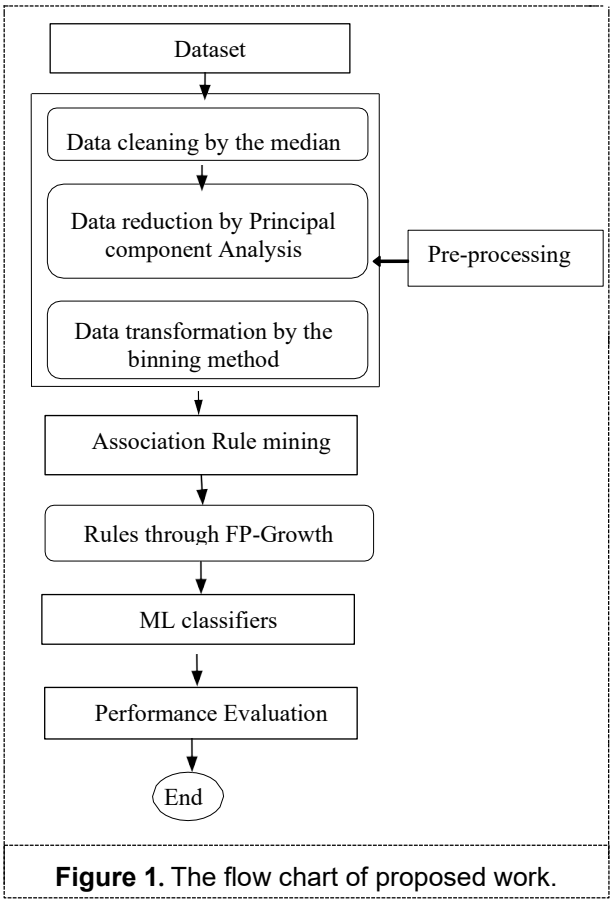

Figure 1. The flow chart of proposed work.

Afterward, the compressed database is partitioned into a series of conditional databases, each linked to a frequent item. Ultimately, each database undergoes independent mining. Critical stages in the FP-growth algorithm involve building the FP-tree and subsequently extracting valuable information from this tree structure. The algorithm is detailed in our published paper [21]. It allows the discovery of frequent itemsets without generating candidate itemsets. The FP-Growth Algorithm is described as follows:

Input: Database of transactions, minimum support threshold (minsup)
Output: Frequent itemsets, association rules

1. Construction of the FP-tree:
   Traverse transactions, count the frequency of each unique item, remove infrequent items, sort the items based on frequency, and construct an FP-tree.
2. Construction of the Header Table
   Create a header table to record the first occurrence of each item in the tree.
3. Construction of Conditional Sets:
   For each item in the header table, extract conditional paths in the tree.
4. Recurrence to Extract Frequent Sets:
   Repeat the process from Step 1 to extract frequent patterns from conditional sets.
5. Generation of Association Rules:
   Generate all possible combinations of items in frequent patterns, split into antecedent and consequent patterns, the confidence is calculated, and the rules are filtered.
6. Return the Final Set of Association Rules
   The final output is a set of association rules meeting the specified criteria.

4.2. Support Vector Machine

Support Vector Machines (SVM) [22] are renowned for their effectiveness as classification algorithms in data mining. They function in high-dimensional spaces by creating a hyperplane that separates two classes with the maximum margin. The primary goal of SVM is to find this optimal linear hyperplane, which enhances predictive accuracy and precision in classifying data points, making it a widely used method across various applications.

In SVM, the training data consists of a set of samples, $X=\{x_1,x_2,..., x_n\}$, each defined by n features, along with corresponding class labels $Y=\{y_1,y_2,...,y_n\}$, where y_i can be -1 or +1 (for example, -1 for non-diabetic and +1 for diabetic). The objective is to determine a hyperplane that maximizes the margin, which is the distance from the hyperplane to the closest points of each class, known as support vectors. The hyperplane is expressed mathematically as:

$$f(x) = w^T x+b =0 \quad (3)$$

where w represents the weight vector and b is the bias term. Maximizing the margin involves minimizing the norm of w while satisfying the constraints that ensure correct classification of the training samples. Specifically, this is done by ensuring that

$$y_i (w^T x_i+b) \geq 1 \text{ for all } i=1,2,.,n \quad (4)$$

To frame this as a convex optimization problem, we can rewrite it as minimizing $(1/2) \|w\|^2$ with the aforementioned constraints. In cases where the data is not linearly separable, kernel functions are used to transform the data into a higher-dimensional space where a linear separation is possible. The kernel function $K(x_i,x_j)$ can take the form $K(x_i,x_j)=x_i \cdot x_j$. For dataset that cannot be perfectly separated, slack variables ξ_i are introduced, allowing for some misclassifications. The constraints are then modified to account for these slack variables, leading to the formulation

$$y_i(w^T x_i+b) \geq 1-\xi_i. \quad (5)$$

Regularization is incorporated through a parameter C that balances the trade-off between maximizing the margin and minimizing misclassification. The objective function thus becomes

$$\min (1/2) \|w\|^2+C\sum\xi_i, \quad (6)$$

where a smaller C allows for a wider margin but permits more errors, while a larger C results in a narrower margin with fewer misclassifications.

To solve for the weights w and bias b, optimization techniques like the Sequential Minimal Optimization algorithm are employed. Support vectors, the points that lie on the margin boundaries, are identified by those having $\xi_i=0$. When classifying new data points, the algorithm computes $f(x) = w^T x+b$ and assigns a class label based on the sign of f(x). Specifically, if $f(x)>0$, the label is +1, and if $f(x)<0$, the label is -1.

Finally, the model's performance is evaluated using metrics like accuracy and cross-validation, ensuring it generalizes well to unseen data. Overall, SVM stands out as a powerful tool for classification tasks, effectively handling complex datasets and delivering reliable results.

4.3. Random Forest

RF, pioneered by Leo Breiman [23], is a powerful ensemble learning method that encompasses an ensemble of unpruned classification or regression trees. These trees are created by randomly sampling training data, and the induction process includes the random selection of features. Leo Breiman's RF [24] introduced in the field of ML, has gained prominence for its robustness and versatility. This ensemble learning technique involves the construction of multiple unpruned decision trees, collectively forming a "forest." The uniqueness of RF lies in its approach to tree creation, where the training data is randomly sampled, and features are selectively chosen during the induction process. The random sampling of training data ensures diversity among the individual trees, preventing overfitting and enhancing the model's generalization ability. Additionally, the random selection of features for each tree promotes the exploration of different aspects of the dataset, contributing to the overall robustness of the RF algorithm. RF works in theory as:

- **Ensemble Learning with Decision Trees:** RF is based on the idea of "bagging" (bootstrap aggregating), where multiple models (decision trees) are trained on different random samples of the data. Instead of relying on a single model, RF combines predictions from multiple trees to achieve a more robust outcome. Each tree in a random forest is a standalone model, but individually, decision trees are prone to overfitting, especially if they grow deep. RF solves this by creating numerous such trees, each one trained on a slightly different dataset subset.
- **Bootstrap Sampling:** For each tree, RF selects a random subset of the training data, where sampling is done with replacement. This means some data points may be used multiple times, while others might not be included at all. This process creates variety in the training data for each tree, reducing the risk of overfitting on any single dataset.
- **Random Feature Selection:** In addition to random sampling of data, RF also randomly selects a subset of features for each split in a tree. For instance, if there are 10 features, each split may use only a few of these, chosen randomly. This feature randomness further reduces the correlation among trees, making the ensemble more robust and less sensitive to overfitting.

By using different features and datasets for each tree, Random Forest achieves diversity among the trees. This diversity is crucial because it allows each tree to capture different aspects of the data, helping the ensemble as a whole to generalize better to new data.

- **Majority Voting or Averaging:** For classification (e.g., predicting whether a patient has diabetes or not), each tree in the forest makes a prediction, and the forest's final prediction is determined by a majority vote across all trees. This aggregation reduces the variance and makes the model more stable, as it lessens the impact of individual, potentially biased trees.

In regression, RF averages the predictions from all trees to arrive at a final prediction, thus reducing the impact of any individual tree that might have outliers or is biased.

4.4. Logistic Regression

LR [25] stands as a pivotal and widely adopted statistical and data mining technique, cherished by statisticians and researchers alike. It serves as an indispensable tool for the analysis and classification of datasets characterized by binary and proportional responses. LR has found widespread application in various fields, contributing significantly to the

exploration and interpretation of data with binary outcomes and proportional response structures. The technique's versatility makes it a cornerstone in statistical modeling predictive analytics, playing a crucial role in addressing and complex problems across diverse domains. The steps of LR are as follows:

- **Binary Classification:** LR is specifically designed to predict a binary outcome (0 or 1, diabetic or non-diabetic in this case) based on one or more predictor variables.
- **Log-Odds and Sigmoid Function:** Unlike Linear Regression, which predicts a continuous value, Logistic Regression predicts the probability of an instance belonging to a particular class (e.g., diabetic). This probability is calculated using the sigmoid (logistic) function, which outputs a value between 0 and 1, representing the likelihood of diabetes.
- **Logistic Function and Probability Prediction :** The logistic function used in LR is expressed as:

$$P(Y=1|X)=1/1+\exp-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n) \quad (7)$$

where:

- Y is the target outcome (1 for diabetic, 0 for non-diabetic).
- $X_1,X_2,\ldots,X_n$ are the feature variables (like glucose level, BMI, age).
- β_0 is the intercept, and $\beta_1,\beta_2,\ldots,\beta_n$ are the coefficients learned during training.

The model outputs a probability score between 0 and 1. By setting a threshold (e.g., 0.5), predictions are classified as diabetic if the probability is above the threshold, and non-diabetic if below.

- **Training the Model:** LR uses optimization algorithms, like Maximum Likelihood Estimation (MLE), to find the best coefficients β that maximize the likelihood of correctly classifying each instance in the training data. In practical implementations, gradient descent is often used to iteratively adjust the coefficients to minimize the error between predicted and actual values.
- **Prediction Process :** For each patient in the PIMA dataset, Logistic Regression calculates the probability of being diabetic based on the input feature.Using a set threshold (typically 0.5), the model assigns a final class (diabetic or non-diabetic). Adjusting the threshold can make the model more or less sensitive to predicting positive cases (diabetes), which can be useful depending on the goal (e.g., reducing false negatives).
- **Evaluation and Interpretation:** Evaluate the model using accuracy to understand its effectiveness on the PIMA dataset. Precision and recall are particularly important in medical data to balance false positives and false negatives.

The coefficients β can be interpreted as odds ratios, which explain the impact of each feature on the likelihood of being diabetic. For example, a positive coefficient for glucose level means that as glucose increases, the likelihood of diabetes also increases.

5. Statistical analysis

5.1. Dataset Description

The data set employed in this study is the PIMA Indian dataset, made available by the National Institute of Diabetes contains information of 768 women from a population near Phoenix, Arizona, USA. This dataset comprises eight independent variables (features) and one dependent variable (target) (Table 1).

Table 1. PIMA dataset description

Attributes	Description	Value interval	Data type
-Pregnancies (P)	It shows how many times patient is pregnant	[0-17]	numeric
-Glucose (G)	Plasma glucose concentration over 2 h in an oral glucose tolerance test.	[0-199]	numeric
-BloodPressure (BP)	It indicates the patient's Blood Pressure	[0-122]	numeric
-SkinThickness (S)	It shows skin fold thickness	[0-99]	numeric
-Insulin (I)	2-Hour serum insulin (mu U/ml).	[0-846]	numeric
-BMI	It indicates Body Mass Index	[0-67]	numeric
DiabetesPedigree Function (DPF)	It shows family history of patient.	[0-2.45]	numeric
-Age (A) -Outcome (O)	It shows age of patient 1 for diabetes and 0 for non- diabetes.	[21-81] (0,1)	numeric binary

The dataset includes various medical predictor variables alongside a target variable, "Outcome," indicating diabetes presence. Predictor variables cover aspects like the number of pregnancies, BMI, insulin levels, age, among others. Collected from the UCI Machine Learning Repository, this dataset contains 768 records, with 268 positive diabetes cases. Each row within the dataset represents a distinct individual, and the values in each column offer specific insights into that person's health and medical history. This comprehensive dataset is a valuable resource for exploring the relationships between various health-related features and the occurrence of diabetes. It provides a nuanced perspective on the health characteristics of different individuals, enabling a thorough examination of factors that may contribute to the development or prevalence of diabetes within the Pima Indian population.

5.2. Dataset Analysis

Within the statistical summary encapsulated in the "diabetes" Data Frame, the dataset's richness becomes apparent. With 768 entries, the absence of null values not only indicates the completeness of the dataset but also assures the robustness of subsequent analyses. Delving into the means of the numerical attributes offers a glimpse into the dataset's central tendencies. For instance, the average number of "Pregnancies" is approximately 3.85, providing a baseline understanding of this variable's distribution.

Examining standard deviations becomes crucial in understanding the variability around the means. Attributes such as "Insulin" and "Age" stand out with notably high standard deviations, pointing to significant data dispersion within these variables. This variability is key in assessing the diversity and distribution of values within each attribute. The range between minimum and maximum values for each attribute further unveils the diversity inherent in the dataset. However, anomalies, like the minimum "Glucose" value of 0, prompt necessary caution and signal potential missing or invalid data that necessitate careful scrutiny and validation.

Table 2. Statistical summary of PIMA diabetes dataset

	P	G	BP	S	I	BMI	DPF	A
count	768	768	768	768	768	768	768	768
min	0,00	0,00	0,00	0,00	0,00	0,00	0,08	21,00
max	17,0	199,0	122,0	99,00	846,0	67,10	2,420	81,00
1st Quartil	1.000	99.0	62.00	0.00	0.00	27.30	0.2437	24.00
Std	3,00	117,0	72,00	23,00	30,50	32,00	0,373	29,00
3rd Quartil	6.000	140.2	80.00	32.00	127.2	36.60	0.6262	41.00
Mean	3,85	120,8	69,10	20,54	79,79	31,99	0,472	33,24

The presentation of percentiles, particularly the first and third quartiles, enriches the statistical narrative by providing a nuanced view of the data distribution. These quartiles serve as crucial points, allowing for a deeper understanding of how data is spread across the dataset. Examining percentiles becomes particularly insightful in identifying potential outliers and understanding the overall structure of the numerical attributes. In summary, this detailed statistical analysis not only provides a comprehensive overview of the PIMA dataset but also lays the groundwork for informed decision-making in subsequent analyses and modeling efforts. The richness of the insights derived from measures of central tendency, variability, and distributional characteristics empowers researchers and analysts to extract meaningful conclusions and navigate the complexities embedded within the dataset presented in Table 2. These results appear to be from a study or data analysis on medical and demographic characteristics. Here is an interpretation of each column:

- **Age:** The average age of individuals in the sample is approximately 41.8 years, with a standard deviation of 22.46. The minimum age is 0.08 years (which seems to be a data entry error) and the maximum age is 80 years.
- **Hypertension:** Approximately 8% of individuals in the sample are listed as having hypertension.
- **Heart disease:** Approximately 4% of individuals in the sample are listed as having heart disease.
- **BMI (Body Mass Index):** The average BMI in the sample is approximately 27.32, with a standard deviation of 6.77. The minimum BMI is 10.01 and the maximum BMI is 95.69.
- **HbA1c (glycated hemoglobin) level:** The average HbA1c level in the sample is approximately 5.53, with a standard deviation of 1.07. The minimum level is 3.50 and the maximum level is 9.00.
- **Blood glucose level:** The average blood glucose level in the sample is approximately 138.22, with a standard deviation of 40.91. The minimum level is 80 and the maximum level is 300.
- **Diabetes:** Approximately 9% of individuals in the sample are listed as having diabetes.

Figure 2 presents the distribution of diabetes and non-diabetes outcomes in the dataset, which is visually represented in a pie chart, illustrating that approximately 34.9% of individuals have been diagnosed with diabetes, while the remaining 65.1% do not have diabetes.

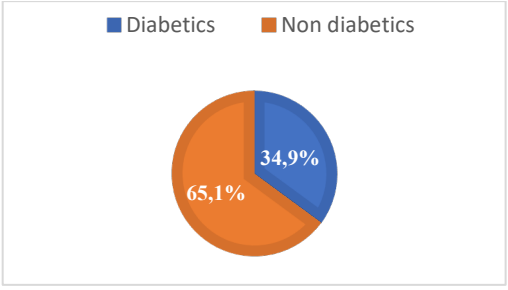

Figure 2. Distribution of Diabetics People

The histogram and the curve in Figure 3 indicate that most glucose readings are centered on 100, with fewer readings towards the extremes of the scale (0 and 200). This could represent blood glucose levels measured in a group of individuals or during repeated tests for one individual over time.

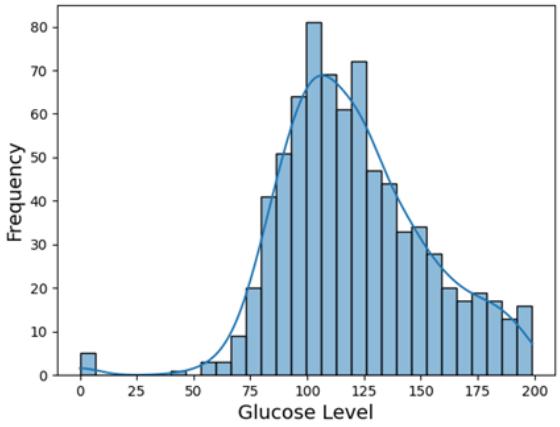

Figure 3. Histogram representation of Glucose

The histogram in Figure 4 is useful in understanding the distribution characteristics of insulin measurements within the examined dataset. It seems to indicate that the majority of the measured insulin levels are relatively low, with fewer occurrences of high insulin levels.

The histogram of Figure 6 shows that the majority of blood pressure values are between approximately 60 and 80, with a clear peak in this range. There is also a small frequency of very low values (close to 0), which could be due to measurement errors or a specific subset of the population.

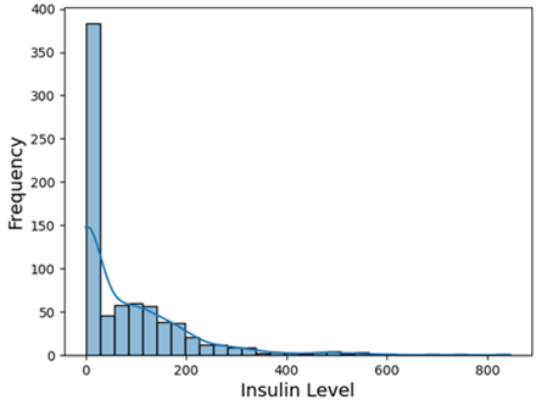

Figure 4. Representation of Insulin attribute

Figure 5 represente the histogram of BMI attribute, this histogram is useful for understanding the overall BMI distribution of a group and can be an important tool in public health analysis to assess the prevalence of underweight, normal weight, overweight, and obesity within the population sampled.

The customized pair plot, depicted in Figure 7, offers a comprehensive visual representation of the correlation a by the 'Outcome' variable. where blue color indicates diabetes and orange color represent no-diabetes.

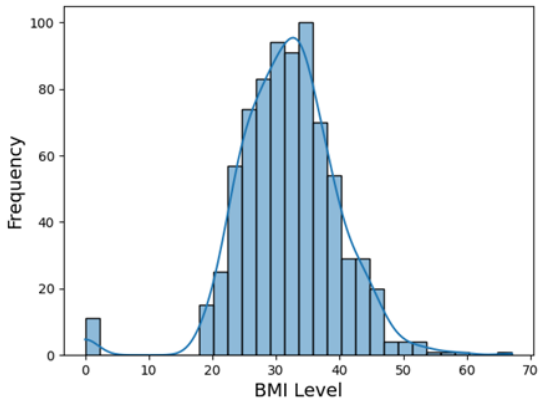

Figure 5. Histogram representation of BMI attribute

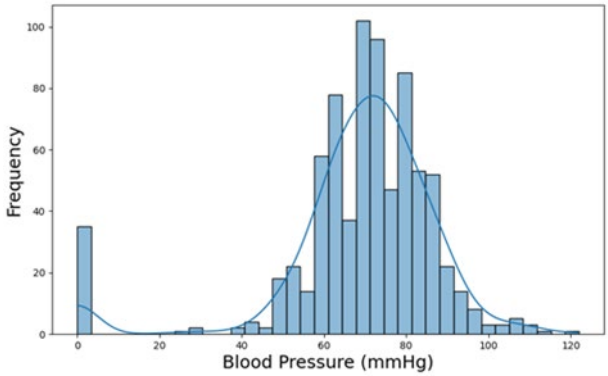

Figure 6. Histogram representation of BloodPressure attribute

From Table 3 we notice some remark such as:

- Grossness and Outcome (0.22): A weak to moderate positive correlation, suggesting that the number of pregnancies might be slightly associated with the diabetes outcome variable.
- Glucose and Outcome (0.47): Here we have a moderate positive correlation, indicating that higher glucose levels could be associated with an increased risk
- of diabetes.
- Insulin and Skin Thickness (0.44): There is a moderate positive correlation between these two variables, which can be interpreted as an association between higher insulin levels and greater skin thickness.
- Some other strong correlations are not directly related to the outcome, such as BMI and Skin Thickness (0.39), which could indicate a physiological relationship between body fat and skin thickness.

Values close to zero represent a very weak or null linear relationship between the variables. For example, BMI and Blood Pressure (0.28) show a weak positive correlation. Overall, the values generally indicate that while some variables have stronger associations with the diabetes outcome, many variables have weak correlations with each other. This suggests that the relationship between these different health measures and diabetes is complex and likely not defined by simple one-to-one correlations.

Table 3. Correlation matrix

	P	G	BP	S	I	BMI	DPF	A	O
P	1.000000	0.129459	0.141282	-0.081672	-0.073535	0.017683	-0.033523	0.544341	0.221898
G	0.129459	1.000000	0.152590	0.057328	0.331357	0.221071	0.137337	0.263514	0.466581
BP	0.141282	0.152590	1.000000	0.207371	0.088933	0.281805	0.041265	0.239528	0.065068
S	-0.081672	0.057328	0.207371	1.000000	0.436783	0.392573	0.183928	-0.113970	0.074752
I	-0.073535	0.331357	0.088933	0.436783	1.000000	0.197859	0.183928	-0.042163	0.130548
BMI	0.017683	0.221071	0.281805	0.392573	0.197859	1.000000	0.140647	0.036242	0.292695
DPF	-0.033523	0.137337	0.041265	0.183928	0.185071	0.140647	1.000000	0.033561	0.173844
A	0.544341	0.263514	0.239528	-0.113970	-0.042163	0.036242	0.033561	1.000000	0.238356
O	0.221898	0.466581	0.065068	0.074752	0.130548	0.292695	0.173844	0.238356	1.000000

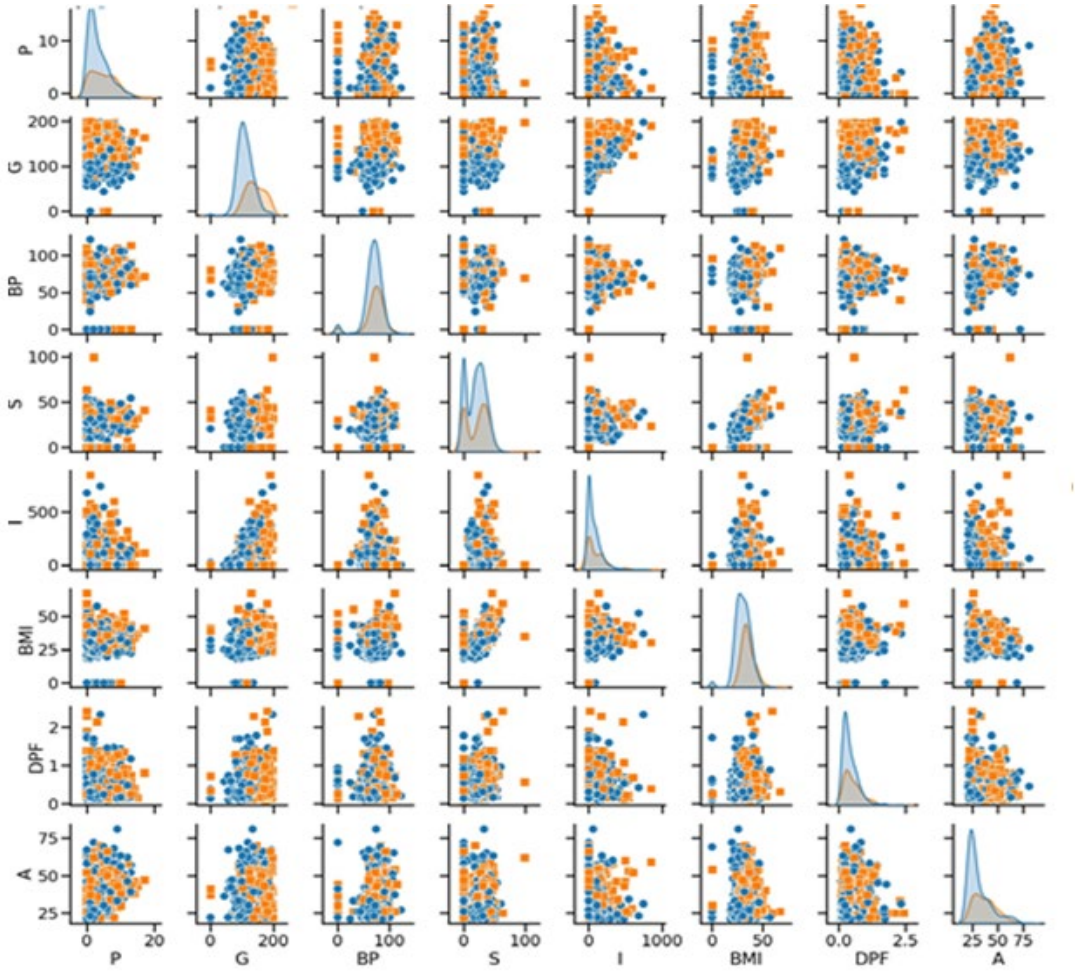

Figure 7. Correlation visualization between different attributs of diabetes dataset

The ninth significant features used for diabetes prediction are shown with their correlations in the heat map illustrated in Figure 8. A heat map is a way to show the correlation between multiple variables at once. It uses a matrix of colored cells, where each cell represents the correlation coefficient between two variables. The correlation coefficient is a numeric value that ranges from -1 to 1 and reflects the direction and magnitude of the correlation. The color and intensity of the cell indicate the value of the coefficient, with a color scale typically ranging from blue (Negative) to red (Positive). A heat map can help you quickly identify variables that are strongly or weakly correlated and spot outliers or anomalies.

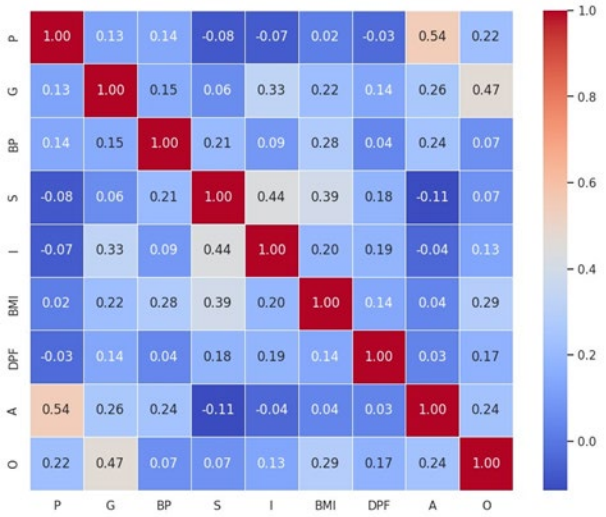

Figure 8. Diabetes features correlation using the heat map

6. Data Preparation and Transformation for Mining

Data pre-processing is an important step to consider before diving into data mining and analysis. This involves various techniques to solve issues like missing values, data reduction and so on in order to improve the quality of the dataset.

6.1. Data cleaning

The initial step is to clean the dataset, by removing duplicated lines and filling in the missing values. And null values in dataset are replaced by the median value of each attribute, and that would help resolve the problem of inconsistencies and ensures a complete dataset. This method is effective for preserving the dataset's original distribution, especially when the proportion of missing data is low.

6.2. Data reduction

This technique aims to give a reduced representation of dataset without affecting the results. Dimensionality reduction is employed to reduce the number of attributes. In this work Principal Component Analysis (PCA) is used. In general the PCA is a popular method of dimension reduction, it simplify a complex dataset from having many variables (the most and the least significant) and keeps only the significant features in the dataset. By applying PCA, the following elements are closely related:

- Pregnancy and age
- Blood glucose and blood pressure (BP)
- BMI, DPF, Insulin levels and skin thickness.

6.3. Data transformation

To effectively utilize the FP-Growth algorithm, our study resquired mmeticulous preparation and transformation of the PIMA Database. This well-curated dataset, which includes records from 768 women of PIMA descent, contains several critical attributes such as plasma glucose concentration, diastolic blood pressure, triceps skin fold thickness, 2-hour serum insulin levels, body mass index (BMI), diabetes pedigree function, the number of pregnancies, and age. These attributes are systematically documented in Table 1, providing a structured overview that highlights each variable's range. Nevertheless, it is essential to transform these continuous variables into categorical intervals, which are crucial for identifying and leveraging patterns predictive of diabetes in our algorithm.

Guided by rigorous statistical analysis and substantial domain knowledge, we transformed these attributes into categorical intervals. This strategic categorization is designed to distinguish between diabetic and non-diabetic groups based on specific attributes.

For instance, age was categorized into 'Young' [0, 30] and 'Senior' [31, 80] groups to reflect different risk profiles. This categorization is visualized in Figure 9, which illustrates the age distribution of diabetic and non-diabetic individuals. Diabetic patients are denoted with an orange colour and the symbol "0," while non-diabetic individuals are marked in blue and represented by the symbol "1."

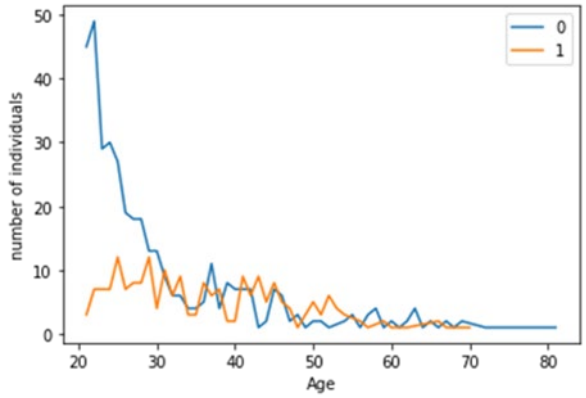

Figure 9. Age distribution of diabetic and non-diabetic individuals

Blood pressure readings were segmented into low [0, 40], medium [41, 90], and high [91, 120] categories to correlate with varying diabetes risks. This distribution is shown in Figure 10.

Similarly, glucose levels were divided into normal [0, 125] and high [126, 200], aligning with clinical thresholds for diabetes diagnosis, as depicted in Figure 11.

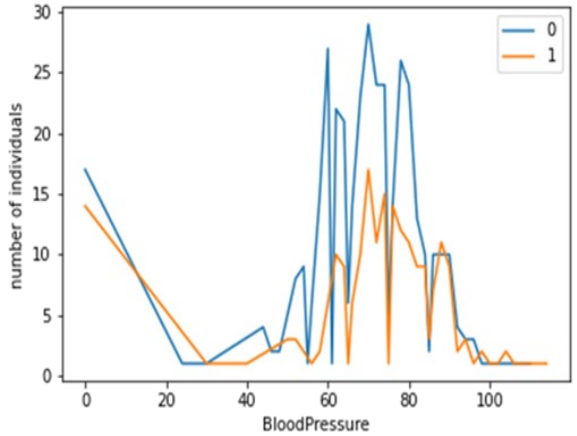

Figure 10. Blood Pressure distribution of diabetic and non-diabetic individuals

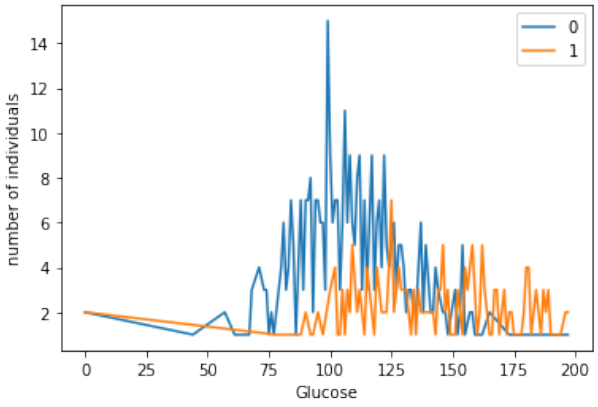

Figure 11. Glucose levels distribution of diabetic and non-diabetic individuals

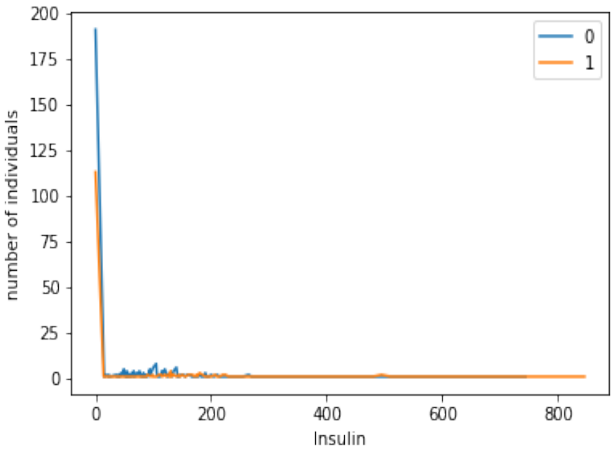

Figure 12. Insulin levels distribution of diabetic and non-diabetic Individuals

Table 4. Categorical transformations for diabetes prediction

Attributes	Categorical intervals
P	P1 {0-5}, P2{>5}
G	G1{0-125}, G2{>125}
BP	B1{0-40}, B2{40-90}, B3{>90}
S	S1{0-8}, S2{8-45}, S3{>45}
I	I1{0-30}, I2{30-150}, I3{>150}
BMI	BMI1{0-30}, BMI2{>30}
D	D1{0-0.8}, D2{>0.8}
A	A1 {0-30}, A2{>30}
O	0 for non-diabetic, and 1 for diabetic

7. Results and Discussion

This section delves into the examination of results. The outlined approach has been put into practice through the utilization of Python, within the Jupyter Notebook environment. The dataset is divided into training and testing: 70 percent of data for validation and training and 30 percent of data for the testing.

7.1. Analysis of Association Rules

By applying the FP-Growth algorithm, the top 10 association rules extracted from the diabetes dataset, generated with a minimum support of 0.4 and a confidence level of 0.5, offer vital insights into factors significantly associated with diabetes risk. These rules, detailed in Table 5, integrate various patient characteristics such as age, body mass index, glucose levels, family history, blood pressure, skin fold thickness, and insulin levels. They reveal patterns that either increase or decrease the likelihood of developing diabetes. For instance, Rule 1 indicates that younger patients (A1: age $\leq$ 30) with a BMI $\leq$ 30 (BMI1), fewer pregnancies (P1: $\leq$ 5), and a lower family diabetes history (D1: $\leq$ 0.8) are less likely to have diabetes. This suggests that traditional risk factors such as age, obesity, and family history significantly influence the

onset of diabetes. Additionally, Rule 2 highlights that patients with high glucose levels (G2: > 125 mg/dl) and a higher BMI (BMI2: > 30) are at increased risk for diabetes, aligning with established medical knowledge that links elevated glucose levels and obesity with heightened diabetes risk.

To conclude, this analysis offers important insights into the factors influencing diabetes risk:

- Protective Factors: Younger individuals (≤ 30 years), those with a lower BMI (≤ 30), fewer pregnancies (≤ 5), and a minimal family history of diabetes (≤ 0.8) are less likely to develop the condition. These findings underscore the benefit of early lifestyle management and awareness of genetic risk in lowering diabetes likelihood.
- Intermediate Risk Indicators: Moderate skinfold thickness (8-45 mm) and blood pressure (40-90 mm Hg) appear to maintain a generally low risk profile for diabetes. On their own, these factors do not strongly indicate diabetes risk, unless accompanied by other more severe factors.
- High-Risk Profiles: Higher blood glucose levels (> 125 mg/dl) alongside a BMI above 30 show a strong association with diabetes, aligning with clinical understanding of the relationship between metabolic imbalance and diabetes development.
- Compound Risk of Hypertension and Obesity: Those with very high blood pressure (> 90 mm Hg) and a higher BMI face an especially increased risk for diabetes, emphasizing the combined impact of hypertension, obesity, and metabolic health.
- Collectively, these insights enrich current medical understanding by refining predictive tools and supporting targeted treatment. They allow healthcare providers to design personalized management plans that tackle individual risk factors, thereby improving prevention and care for diabetes. Additionally, this analysis highlights the role of advanced data-mining techniques in revealing complex, multifaceted patterns within medical data.

Table 5. Key Association Rules Identifying Risk and Protective Factors for Diabetes

Rule number	Antecedent	Consequent	Confidence
1	["BMI1","A1","P1","D1"]	["0"]	0,93
2	["G2", "BMI2"]	["1"]	0,90
3	["G1","S2","P1","D1","B2"]	["0"]	0,89
4	["A2", "G2", "BMI2"]	["1"]	0,88
5	["G1","S2","D1"]	["0"]	0,85
6	["A1", "P1", "G1", "S1"]	["0"]	0,85
7	["S3", "G2"]	["1"]	0,85
8	["I3", "G2"]	["1"]	0,83
9	["B3", "BMI2"]	["1"]	0,80
10	["G2","BMI2"]	["0"]	0,85

7.2. Accuracy analysis

After association rules extraction, we conclude that factors of risk to have diabetes are mainly BMI, Glucose levels, Insulin levels, age and heredity. So we used these factors to predict diabetes by ML algorithms, the metric used in this study include accuracy and the confusion matrix. The accuracy (equation 8) is computed by summing up two correct predictions (True Positives + True Negatives) and dividing this sum by the total number of data sets (Positives + Negatives). The optimal accuracy score is 1.0, while the lowest possible score is 0.0.[25]

$$ACC = \frac{TP + TN}{TP + TN + FN + FP} \tag{8}$$

The confusion matrix (Table 6) displays the predicted values of the data, distinguishing between True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). In this context, True Positives signify correctly identified positive instances, True Negatives denote correctly identified negative instances, False Positives represent instances incorrectly classified as positive, and False Negatives indicate instances incorrectly classified as negative.

Table 6. Confusion matrix representation

Class designation		Actual Class	
		True	False
Predicted Class	Positive	TP	FP
	Negative	TN	FN

Execution time it is a crucial metric in evaluating the efficiency of an algorithm in terms of how quickly it can process input data and produce the desired output. The confusion matrices in Table 7 offer insights into the performance of three different classification techniques: SVM, LR, and RF. Looking at the SVM results, it correctly identified 96 instances as negative and 26 instances as positive. However, it made 11 false positive errors and 21 false negative errors. This suggests that while the SVM is effective in correctly classifying negatives, it struggles with false positives and false negatives. Moving to LR, it demonstrated a higher number of true negatives (98) and true positives (29) compared to the SVM. However, it made 9 false positive errors and 18 false negative errors. LR seems to have a better performance in terms of false positives but has a notable number of false negatives.

Table 7. Performance comparison using Confusion Matrices

Techniques	Confusion Matrix	
SVM	96	11
	21	26
LR	98	9
	18	29
RF	92	15
	18	29

The RF model also had a substantial number of true negatives (92) and true positives (29), but it made 15 false positive errors and 18 false negative errors. The RF model appears to have a balanced performance, with strengths and weaknesses similar to both SVM and LR. As understood and explored, we can observe that while LR achieved the highest accuracy, the SVM technique was the most efficient in terms of execution time when applied on our database (diabetes).

LR emerged as the top performer with an accuracy of 82.46%. This indicates that the technique successfully predicted the outcome variable based on the input features, displaying its effectiveness in handling the dataset. SVM, with an accuracy of 79.22%, demonstrated a slightly lower but still respectable performance. SVM is known for its ability to handle complex relationships between variables, and its performance suggests its suitability for this dataset. RF, with an accuracy of 78.57%, lagged slightly behind LR and SVM. Despite this, RF is an ensemble method known for its robustness and stability, leveraging multiple decision trees for predictions. While RF required considerably more time for computation, SVM and LR demonstrated shorter execution times. The choice among these techniques should consider the balance between computational efficiency and predictive performance. SVM's fast execution time might be advantageous in scenarios where real-time processing is crucial, while RF's time may be acceptable if its ensemble characteristics contribute significantly to accurate predictions. LR, falling in between, represents a trade-off between computational efficiency and accuracy. Figure 13 illustrates the results obtained by different ML algorithms obtained by Linshan Xie [26] and Tegga et al [27] and our proposed method. Figure 14 illustrate the comparison between K-Nearest Neighbors (KNN), Decision Tree (DT), Logistic Regression (LR), SVM, Naïve Bayes (NB) and our method. Our method uses LR and risk factors extracted by FP-Growth algorithm.

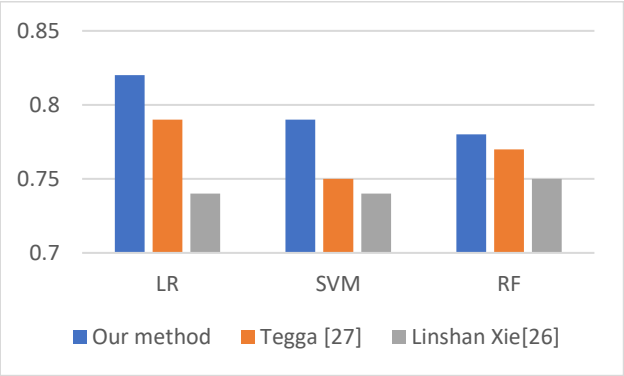

Figure 13. Accuracy comparison

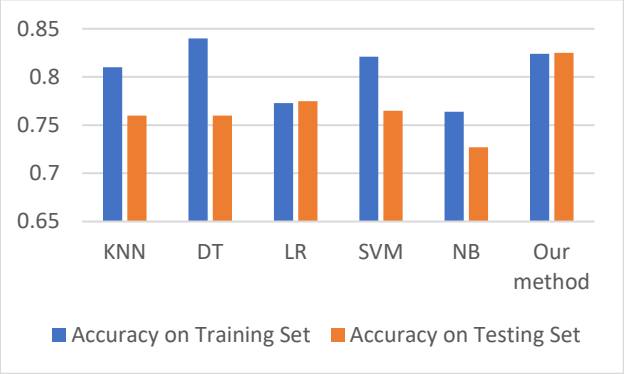

Figure 14. Accuracy comparison by different methods

8. Conclusion

In summary, the selection of a ML model or association rule-mining algorithm is contingent upon the specific goals, priorities, and constraints associated with the given task. LR, RF and SVM each offer distinct trade-offs between accuracy and computational efficiency. LR is valued for its simplicity and interpretability, RF for its robustness and ensemble capabilities, and SVM for its effectiveness in high-dimensional spaces.

When it comes to association rule mining, the choice FP-Growth depends on the desired level of computational intensity. In contrast, FP-Growth takes an approach, utilizing a tree structure to efficiently mine more streamlined association rules with a reduced computational burden.

Ultimately, the decision-making process involves a careful consideration of the benefits and drawbacks of each approach against the specific requirements of the task. Factors such as available computational resources, dataset complexity, and the balance between interpretability and computational efficiency should be taken into account. By aligning the chosen algorithm with these considerations. , practitioners can optimize their approach for the successful accomplishment of their objectives.

References

[1] F. A. Jaber and J. W. James, "Early Prediction of Diabetic Using Data Mining," *SN Comput. Sci*., *vol. 4, no. 2,* p. 169, Jan. 2023, doi: 10.1007/s42979-022-01594-z.

[2] L. A. DiMeglio, C. Evans-Molina, and R. A. Oram, "Type 1 diabetes," *The Lancet, vol. 391, no. 10138*, pp. 2449–2462, Jun. 2018, doi: 10.1016/S0140-6736(18)31320-5.

[3] H. Ikegami, Y. Hiromine, and S. Noso, "Insulin-dependent diabetes mellitus in older adults: Current status and future prospects," *Geriatr. Gerontol. Int., vol. 22, no.* 8, pp. 549–553, Aug. 2022, doi: 10.1111/ggi.14414.

[4] R. Modzelewski, M. M. Stefanowicz-Rutkowska, W. Matuszewski, and E. M. Bandurska-Stankiewicz, "Gestational Diabetes Mellitus—Recent Literature Review*," J. Clin. Med., vol. 11, no. 19*, p. 5736, Sep. 2022, doi: 10.3390/jcm11195736

[5] Yan Niu and Shenglan Ye, Data Prediction Based on Support Vector Machine (SVM)—Taking Soil Quality Improvement Test Soil Organic Matter as an Example, *IOP Conference Series: Earth and Environmental Sciences,* 295 (2019) 012021, doi:10.1088/1755-1315/295/2/012021.

[6] M. Maalouf, "Logistic regression in data analysis: an overview," *Int. J. Data Anal. Tech. Strateg., vol.* 3, no. 3, p. 281, 2011, doi: 10.1504/IJDATS.2011.041335

[7] R. D. Joshi and C. K. Dhakal, "Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches," *Int. J. Environ. Res. Public. Health,* vol. 18, no. 14, p. 7346, Jul. 2021, doi: 10.3390/ijerph18147346.

[8] S. Hegelich, "Decision Trees and Random Forests: Machine Learning Techniques to Classify Rare Events," *Eur. Policy Anal*., vol. 2, no. 1, pp. 98–120, Mar. 2016, doi: 10.18278/epa.2.1.7

[9] Leo Breiman, Random Forests, "Machine Learning, 45, pp: 5–32, 2001 *Kluwer Academic Publishers. Manufactured in The Netherlands.*", 2001.

[10] A. Verikas, A. Gelzinis, and M. Bacauskiene, "Mining data with random forests: A survey and results of new tests," *Pattern Recognition*, vol. 44, no. 2, pp. 330–349, Feb. 2011, doi: 10.1016/j.patcog.2010.08.011

[11] Youssef FAKIR, Rachid ELAYACHI , Btissam MAHI, Clustering objects for spatial data mining: a comparative study, *Journal of Big Data Research*, vol.1, issue 1, 2020

[12] B. Nigam, A. Nigam, and P. Dalal, "Comparative Study of Top 10 Algorithms for Association Rule Mining," *Int. J. Comput. Sci. Eng.,* vol. 5, no. 8, pp. 190–195, Aug. 2017, doi: 10.26438/ijcse/v5i8.190195.

[13] Victor Chang, Jozeene Bailey, Qianwen Ariel Xu, Zhili Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms - *PubMed.", part of Springer Nature* 2022

[14] C. C. Olisah L. Smith, and M. Smith, "Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective," *Computer Methods and Programs in Biomedicine,* vol. 220, p. 106773, 2022

[15] Youssef Fakir, Abdelfatah Maarouf, Rachid El Ayachi, Mining Frequents Itemset and Association Rules in Di*abetic Dataset Lecture Notes in Business Information Processing ((LNBIP*,volume 449)).

[16] U. E. Laila, K. Mahboob, A. W. Khan, F. Khan, and W. Taekeun, "An ensemble approach to predict early-stage diabetes risk using machine learning:An empirical study," *Sensors,* vol. 22, no. 14, p. 5247, 2022

[17] T. Mahesh *et al.*, "Blended ensemble learning prediction model for strengthening diagnosis and treatment of chronic diabetes disease," *Computational Intelligence and Neuroscience,* vol. 2022, 2022.

[18] B. S. Ahamed, M. S. Arya, and A. O. V. Nancy, "Diabetes Mellitus Disease Prediction Using Machine Learning Classifiers with Oversampling and Feature Augmentation," *Advances in Human-Computer Interaction,* 2022.

[19] B. Kurt *et al.*, "Prediction of gestational diabetes using deep learning and Bayesian optimization and traditional machine learning techniques," *Medical &Biological Engineering & Computing,* pp. 1-12, 2023.

[20] Ran Rong Liu , LiJun Wang, Rong Miao, A Data Mining Algorithm for Association Rules with Chronic Disease Constraints, *Hindawi Computational Intelligence and Neuroscience Volume 2022*, https://doi.org/10.1155/2022/8526256

[21] Youssef Fakir, R. El Ayachi and Mohamed Fakir, Mining Frequent Pattern by, *International Journal of Scientific Research in Computer Science Engineering and Information Technology.*, 2020, DOI: https://doi.org/10.32628/CSEIT2063230.

[22] Kaina Zhao, Zhiping Wang, Association rule mining to detect factors which contribute to heart disease in males and females, *MLMI '23: Proceedings of the 6th International Conference on Machine Learning and Machine Intelligence* October 2023 Pages 29–33, https://doi.org/10.1145/3635638.3635643

[23] Breiman, L.. Some infinity theory for predictor ensembles. *Technical Report 579, Statistics Department, University of California, Berkeley, CA 94720, 2000.*

[24] Leo Breiman , Random Forests ,Machine Learning, 45, 5–32, 2001

[25] Ghadeer Mousa, Hassan Abu Hassan and Hussein Al-Rimmawi, Prediction of Type 2 Diabetes using logistic regression techniques, *Turkish Journal of Computer and Mathematics Education Vol.15 No.1(2024*).

[26] Linshan Xie, Pima Indian Diabetes Database and Machine Learning Models for Diabetes Prediction, Highlights in Science, Engineering and Technology, Volume 88 (2024).

[27] N.P. Tigga, S. Garg, Prediction of type 2 diabetes using machine learning classification methods, Procedia Comput. Sci. 167 (2019) (2020) 706–716, https:// doi.org/10.1016/j.procs.2020.03.336.