Rule Mining of Early Diabetes Symptom and Applied Supervised Machine Learning and Cross Validation Approaches based on the Most Important Features to Predict Early-Stage Diabetes

Mahade Hasan [1], Farhana Yasmin [2], Linhong Deng [3]

[1, 2] School of Computer Science and Artificial Intelligence,
[3] Institute of Biomedical Engineering and Health Sciences,
Changzhou University, Changzhou, Jiangsu 213164, China.

Published in IJIRMPS (E-ISSN: 2349-7300), Volume 11, Issue 3 (May-June 2023)

Abstract

Diabetes is one of several illnesses referred to be "chronic". It is the most prevalent disease that significantly impacts the population. Although there are numerous possible causes of diabetes, age, excessive body fat, frailty, fast weight loss, and many other conditions are the ones that occur most often. Diabetes patients are more susceptible to developing a number of diseases, including heart disease, kidney issues, damaged nerves, damaged blood vessels, and blindness. It is challenging to diagnose the ailment, and it is both costly and difficult to anticipate how it will develop. Machine learning (ML) offers tremendous potential to develop useful applications for earlier detection, diagnosis, and therapy, as well as the treatment of many disorders, which is why medical experts are particularly interested in it. This study aims to develop a model that can reliably and precisely identify diabetes. Following that, association rule mining was employed to find the common indications of diabetic symptoms. This study also presents a useful model for diabetes prediction that makes use of various machine-learning approaches to improve diabetes categorization and increase the precision of diabetes prediction. Machine learning methods utilized in the early stage diabetes prediction include Gaussian Naive Bayes, ExtraTreesClassifier, Decision Trees, K-Nearest Neighbors, Random Forest Classifier, Support Vector Machine, and Logistic Regression. The choice of the dataset's major attribute was then made after considering a total of six different ways. Then, a total of 10 alternative models utilized for early-stage diabetes prediction were applied to the previously selected and highlighted dataset. Accuracy, precision, recall, and F-measure are some of the metrics used to evaluate the various performance levels of these models. The performance matrices show that the ExtraTreesClassifier performed at the maximum level possible, earning a perfect score in each area with accuracy, recall, precision, and F1 score of 100%. Therefore, we can assert that the performance of our ExtraTreesClassifier model is superior to that of the already available work. Clinical doctors who read this article will gain new knowledge and be better able to identify early diabetes.

Keywords: Machine Learning, Mine Ruling, Supervised Learning

1. Introduction

Hyperglycemia is a hallmark of the clinical illness known as diabetes mellitus, which has multiple etiologies (mellitus being Latin for sweet). About 90% of those with diabetes have type 2, with the rest having type 1. Diabetes affects more than 422 million people globally, most of whom live in low- and middle-income countries. By keeping a healthy weight and managing blood sugar levels, diabetes can be avoided. An estimated 1.5 million individuals die from diabetes directly each year throughout the world. Throughout the recent past, there has been a continuous rise in both the overall prevalence of diabetes and the number of instances of the condition. [1]. Diabetes is a chronic condition that affects how the body burns food for energy. Your body converts the bulk of the food you consume into sugar (glucose), which is subsequently absorbed into your circulation. In reaction to a rise in blood sugar levels, the pancreas secretes insulin. In the presence of insulin, blood sugar can enter cells and be utilized as a source of energy. Insufficient or improper insulin usage by the body can result in diabetes. When there is insufficient insulin or when cells stop responding to insulin, quite so much blood sugar is left in the circulation. It may be followed by serious health problems like heart disease, kidney failure, and vision loss. More people are served by global health systems. Many deadly diseases impact people worldwide. Diabetes causes heart attacks, kidney failure, blindness, and other major health issues. Every hospital tracks illnesses. IT has transformed health care. Each machine-learning algorithm enhances disease prediction and healthcare automation. Hadoop, built on computer clusters, efficiently analyzes and stores massive datasets in the cloud. This work [2] suggests forecasting diabetes with Hadoop-based machine learning. The results imply machine-learning algorithms can accurately forecast diabetes. The effectiveness of the algorithm is investigated using the Pima Indians Diabetes Database from the National Institute of Diabetes and Digestive Disorders. A cutting-edge computing healthcare system is one of the most investigated topics in healthcare research. Computing and healthcare researchers collaborate to advance such systems. The World Health Organization (WHO) found an increase in diabetic patients and deaths. Diabetes can have long-term effects. Medical research is expanding. Technological advances and continual monitoring are needed to collect, preserve, study, and forecast such individuals' health. India's rising diabetic population is worrying. A system that records, analyzes, and searches for diabetic risks using technology is essential. Researchers are developing early detection and analysis methods. This paper [3] reviews diabetic studies and prospective frameworks.

Recently, healthcare research is popular and data-driven. Healthcare data volumes require big data analytics. Millions worldwide use various treatments. If patient care patterns for a disease are examined first, conclusions will be more informed. Healthcare improves by the AI implementation. These [4],[5], [6],[7] are the some recent examples of them. Clinicians can use machine learning to diagnose diseases early. This WEKA-based study [8] creates a diabetes classifier. Naive Bayes, Support Vector Machine, Random Forest, and Simple CART will forecast the study. The project will propose the best algorithm for diabetes prediction. Each dataset for every method was analyzed. The SVM predicted diseases best. Modern diets and erratic eating habits are increasing the prevalence of diabetes. Obesity and high blood glucose levels are major diabetes risk factors. This study [9] investigates the main causes of diabetes. Variable and feature selection has become a prominent research topic in application domains with easily accessible datasets with tens or hundreds of pieces. Machine learning shows this trend. They'll also focus on key factors to consider when assessing diabetes risk.

Medical diagnostic software development is difficult due of disease prediction. Medical diagnosis is one application of machine learning. A machine learning algorithm-based classifier system may help

medical personnel identify and predict diseases and solve health issues. Machine learning classification algorithms can improve a disease diagnosis system's efficiency, effectiveness, dependability, and precision. This article [10] discusses machine learning-based diabetes diagnosis. In addition to artificial neural networks, decision trees, random forests, naive Bayes, support vector machines, logistic regression, and k-nearest neighbors, the PIMA Indian Diabetic dataset also made use of these techniques. These analyses and their pros and cons were then examined. Predictive analytics on massive data sets generally uses machine learning methods. Predictive analytics in medicine is difficult to execute, but it can help doctors make quick choices about patient health and therapy based on massive volumes of data. This study examines healthcare predictive analytics using six machine learning algorithms. The experiment uses six machine learning algorithms on patient medical records. Several methods are compared for efficiency and accuracy. The study's machine learning approaches yielded the best diabetes prediction system. This study [11] uses machine learning to help doctors diagnose diabetes early.

Moreover, healthcare generates sensitive data. Predicting diabetes is medical goal. Machine learning methods can examine data and synthesize decision-making expertise. Data mining can provide valuable insights from enormous amounts of accessible data. Analyzing new patterns gives customers vital information. Diabetes raises heart, kidney, nerve, and eye risks. Data mining will classify and identify Diabetes dataset trends. The UCI Pima Indian diabetes database was used. A sophisticated model used to predict and diagnose diabetes. Bootstrapping resampling improves accuracy in this investigation [12]. Next, Naive Bayes, Decision Trees, and KNN algorithms analyzed each strategy. Besides that, diabetes can be brought on by old age, obesity, inactivity, family history, eating unhealthily, hypertension, etc. Diabetes raises the risk of kidney disease, stroke, visual problems, nerve damage, and other conditions. Today, hospitals utilize many tests to detect and treat diabetes. Healthcare and medicine use big data analytics. Healthcare organizations have enormous databases. Big data analytics can scan enormous databases, find new information and trends, and make predictions. Current categorization and prediction fail. This study [13] suggests using blood glucose, body mass index, age, and insulin levels to predict diabetes. This model adds diabetes risk factors. New datasets enhance categorization. Diabetes prediction improved with pipeline models.

Since the body's inability to metabolize glucose is the underlying cause of diabetes mellitus, a chronic illness that is on the rise. Based on demographic information and laboratory results from medical visits, a different study [14] developed a prediction model with great sensitivity and selectivity to better identify Canadian patients at risk of developing diabetes mellitus. Most Americans die from cardiovascular disease and diabetes. Recognizing and preparing for these diseases' patient presentations is the first step in halting their progression. They analyze the capabilities of algorithms that use machine learning to identify identified as high individuals and test outcomes using survey data, [15] uncovering data factors that contribute to the prevalence of diseases.

On the other hand, in recent years, machine learning has become an increasingly popular method for the analysis of datasets pertaining to medical subjects. The objective of this work is to develop a solution by identifying the best model for the early detection of diabetes using techniques from machine learning. The following is a list of the most important ramifications that this study has:

- The data conversion for the further simulation.

- Conducted association rule mining in order to determine the most common pattern of diabetes symptoms.
- Applied Data Preprocessing on the dataset.
- Six different models were applied to the dataset to uncover key characteristics.
- Ten different models were used to analyze the dataset.
- Analyzing those models' performances to decide which one performs the best.
- Examining the best model's performance in relation to earlier studies.

The study is divided into the following sections: Related Work, Experimentation Environment, Methodology, Data Collection, Data Conversion, Rule Mining, Data Preprocessing, Important Feature Selection, Train Test Split, Applied Models, Result Analysis, Discussion, Comparison with Existing Works, and Conclusion. In light of this, the part that comes next provides some more clarification.

2. Literature Review

This section covered the research that only makes use of the diabetes dataset, and it included a review of the relevant literature to our study. We examined the research procedures as well as the findings of the investigations when we were reviewing those works of literature.

Tripathi et al. [16] Diabetes impacts glucose. Insulin resistance causes difficulties. Undiagnosed, it destroys kidneys, nerves, and eyes. Technology enhances individualized medicine. Healthcare uses machine learning, a fast-growing predictive analysis subfield. These tools detect illnesses. Machine learning categorization and diabetes-related factors predict diabetes early in this study. It improves patient diagnosis and produces clinically useful results. Four ML algorithms predict early diabetes. LDA, KNN, SVM, Random Forest (RF). Pima Indian Diabetes Database from UC Irvine's machine learning repository is used in experiments (PIDD). Categorization systems are evaluated using the metrics sensitivity (recall), precision, specificity, F-score, and accuracy. Suitable categories. The accuracy of RF categorisation is 87.66%.

Alaa Khaleel et al. [17] Diabetics have high blood sugar levels. It's deadly. Early detection lowers diabetes severity and risk. Machine learning, especially in disease, is becoming more popular in medicine due to its ubiquity. This study serves as a diabetes diagnosis model. Using precision, recall, and F1-measure, we assess the prediction accuracy of potent machine learning (ML) systems. Diabetic symptoms were predicted by the PIDD dataset. LR, NB, and KNN exhibited 94%, 79%, and 69% accuracy, respectively. LR predicts diabetes better.

Zou et al. [18] Diabetes produces hyperglycemia. It's risky. Morbidity will cause 642 million diabetes cases by 2040. 10% get diabetes. Alarming. Medical and public health employ machine learning. Their decision trees, random forests, and neural networks predicted diabetes. Hospital patients in Luzhou, China, are examined. 14. This experiment cross-validated five models. They used top methodologies to evaluate their viability. High-performing approaches did this. 68994 healthy and diabetes patients were trained. Unbalanced data pulled 5 times more. Answer: five-test average. MRMR and PCA lowered this study's dimension (mRMR). Random forest prediction was best overall (ACC = 0.8084).

Tigga et al. [19] India has approximately 30 million diabetics many others are in danger. To avoid diabetes and its complications, early diagnosis and treatment are therefore essential. A study like this

one estimates diabetes risk from lifestyle and family history. Machine learning algorithms, which are accurate, predicted type 2 diabetes risk. Medical professionals require precision. Individuals can estimate their diabetes risk after the model is trained. For the trial, 952 participants completed an online and offline questionnaire. The 18-question survey includes sections on family history, lifestyle, and health. The same methods were used to assess the Pima Native Diabetes database. For both datasets, the Random Forest Classifier performs best.

Sisodia et al. [20] Diabetes elevates glucose (sugar). Undiagnosed diabetes can create several issues. The patient always sees a doctor at a diagnostic facility because identification takes so long. Machine learning solves this major problem. This study seeks a model that reliably predicts diabetes risk. This experiment detects early diabetes using decision trees, SVMs, and naive bayes. The Pima Indians Diabetes Database is studied by the machine learning repository at UC Irvine (PIDD). Each strategy is assessed using recall, F-measure, precision, and accuracy. Incident categorization measures accuracy. Naive Bayes' 76.30% accuracy beats others. ROC curves verify this.

Ramesh et al. [21] Millions have diabetes. It worsens organ failure and life quality. Diabetics need early detection and monitoring. Remote patient monitoring facilitates treatment. For automated diabetes risk prediction and management, this study suggests an end-to-end remote monitoring system. Smartphones, smart wearables, and health devices fuel the platform. A Pima Indian Diabetes Database Support Vector Machine predicted diabetes risk after scaling, imputation, selection, and augmentation. Tenfold stratified cross validation produced 83.20% accuracy, 87.20% sensitivity, and 79% specificity. Consistent. Smartphones and smartwatches measure vitals, slow diabetes, and connect with doctors. The unobtrusive, economical, and vendor-interoperable platform aids doctors in their decisions by using the most recent diabetes risk projections and lifestyle data.

Perveen et al. [22] Interventional programs can save time and dollars by targeting high-risk diabetics. A trusted prognostic model, the Framingham Diabetes Risk Scoring Model (FDRSM), was put to the test using a Hidden Markov Model (HMM), a machine learning method. 8-year projection of diabetes risk? FDRSM performance has not been verified by any HMM studies. HMM assessed the 8-year diabetes risk from 172,168 primary care patients' EMRs. Our 911-person sample exhibited an AROC of 86.9% with all risk factors present with follow-up data, which was higher than the 78.6% and 85% in a previous FDRSM validation analysis conducted on the same Canadian population and the Framingham study, respectively. Including all risk factors and follow-up information, 911 research participants had an AROC of 86.9%. In comparison to Canadian and Framingham FDRSM validation research, the suggested HMM discriminates better. Eight-year diabetes risk can be determined by HMM.

Maniruzzaman et al. [23] Diabetes causes high blood sugar. It can cause heart attack, kidney failure, stroke, and other serious illnesses. 422 million people had diabetes in 2014. 642 million will live on Earth by 2040. This project creates an ML-based diabetes diagnosis system. Kavakiotis et al. [24] High-throughput clinical data and genetic data from massive electronic health records have expanded thanks to biotechnology and health research (EHRs). Biosciences must make use of data mining and machine learning to evaluate all data. World health is impacted by DM. Studies have been done on diabetes management, etiopathophysiology, and other topics. Methods for diabetes research, including machine learning and data mining, will be studied for prediction, diagnosis, complications, genetic background and environment, healthcare, and treatment. Popularity is first. ML algorithms were utilized. Methods

were 85% monitored, but association rules were not. SVMs dominate. Clinical data ruled. The names of the selected publications show that extracting vital knowledge generates new ideas that improve DM comprehension and research.

Hasan et al. [25] Diabetes elevates glucose. Diabetes. Early detection reduces diabetes risk. Outliers and unlabeled data complicate diabetes prediction. Outlier rejection, data standardization, feature selection, K-fold cross-validation, several Machine Learning (ML) classifiers (k-nearest Neighbor, Decision Trees, Random Forest, AdaBoost, Naive Bayes, and XGBoost), and Multilayer Perceptron were all used in this literature's robust diabetes prediction framework (MLP). Guidelines: ML Area ROC Curve weights (AUC). This study suggests weighting diabetes prediction ML models. Grid search maximizes hyperparameter adjustment AUC. This study used the Pima Indian Diabetes Dataset and identical experimental parameters. Our recommended ensembling classifier is the most successful classifier from exhaustive testing, with a diagnostic odds ratio of 66.234, an AUC of 0.950, a sensitivity of 0.789, a specificity of 0.934, a false omission rate of 0.092, and a diagnostic odds ratio of 0.934. 2.0% lower AUC. Poor diabetes prediction. Same dataset may improve diabetes prediction systems. Diabetic prognosis.

Yahyaoui et al. [26] DSS helps doctors and nurses make clinical decisions. This is needed due to escalating deadly diseases. Diabetes kills globally. Raising blood sugar may influence other organs. Diabetes. By 2035, there will be 592 million cases of diabetes worldwide, according to the International Diabetes Federation (IDA). This research proposes a machine learning-based diabetes prediction DSS. Machine vs. deep learning. SVM and Random Forest classifiers are popular (RF). Diabetics were predicted and identified by fully convolutional neural networks (CNNs) (DL). To assess the proposed strategy, 768 samples with 8 characteristics were used from the public Pima Indians Diabetes database. 500 samples were non-diabetic, 268 were. SVM 83.67%, RF 76.81%, and DL 65.38%. RF outperforms deep learning and SVM in diabetes prediction.

Sonar et al. [27] Diabetics die. It causes blindness, urinary system problems, coronary heart disease, and more. After the consultation, the patient must drive to a diagnostic center for their reports, which takes time and money. Machine learning can now solve it. Polygenic disease is diagnosed using cutting-edge information processing. Anticipating illness allows for critical care. Data from a lot of unviewed diabetes-related information is removed. This study will improve diabetic risk prediction. SVM algorithms, naïve bayes networks, decision trees, and AI networks characterize models (ANN). 85%, 77%, and 77.3% of precision are estimated by Decision Tree, Naive Bayes, and Support Vector Machine models, respectively. Results are accurate.

Sivaranjani et al. [28] One of the most widespread and deadly diseases in the world, including India, is diabetes. Lifestyle, genetics, stress, and age can cause diabetes at any age. Untreated diabetes, regardless of cause, can have catastrophic consequences. Several methods can anticipate diabetes and its complications. Researchers employed SVM and Random Forest machine learning techniques in the suggested work (RF). These algorithms estimate diabetes risk. After data preparation, step forward and backward feature selection identifies predictive qualities. Selecting features, PCA dimensionality reduction is studied. Random Forest (RF) has an 83% prediction accuracy, compared to SVM's 81.4%.

Saha et al. [29] Diabetes, a prevalent condition, can strike at any age. These diseases activate when blood sugar rises. Predicting diabetes is crucial right now. The Indian Pima Dataset has undergone many techniques. This dataset includes Pima Indian women's 1965 research. Most academics are trying to apply difficult methods to datasets; however, a lot of in-depth research lacks easy strategies. Our study included RF, SVM, and NN (NN). They used these methods in several ways. They added several methods to the main dataset. They then identified diabetics using preprocessing methods. They compared and got the best outcomes using those methods. Neural Network was the most accurate method (80.4%).

Posonia et al. [30] Diabetes mellitus, which can cause severe birth abnormalities, affects most Indian pregnant women. Several cutting-edge blood test technologies can detect diabetes. Diabetes results from elevated blood glucose. Untreated diabetes can cause renal damage and heart attacks. Thus, discovering and studying gestational diabetes requires learning models and rigorous research. This study suggested diabetes prediction using machine learning. Calculation using a decision tree, J48. One of the best classification models is "Decision Tree." A goal column showing favorable or bad outcomes and the major 8 characteristics of 768 patients were evaluated. Our Weka experiment showed that the Decision Tree J48 calculation is more effective and faster.

Pavani et al. [31] Today's healthcare uses AI and ML. The WHO says diabetes affects the most individuals worldwide. High glucose levels induce it. Diagnosing diabetes may involve other factors. This research aims to develop a diabetes-prediction system. This study employed ML methods to predict early diabetes. Among the technologies used in machine learning include support vector machines, logistic regression, decision trees, random forests, gradient boost, K-nearest neighbor methods, and Naive Bayes. These algorithms are evaluated using precision, accuracy, recall, and F-measure. This study compares approaches to improve precision. The accuracy of the Naive Base Method and Random Forest algorithm is 80%.

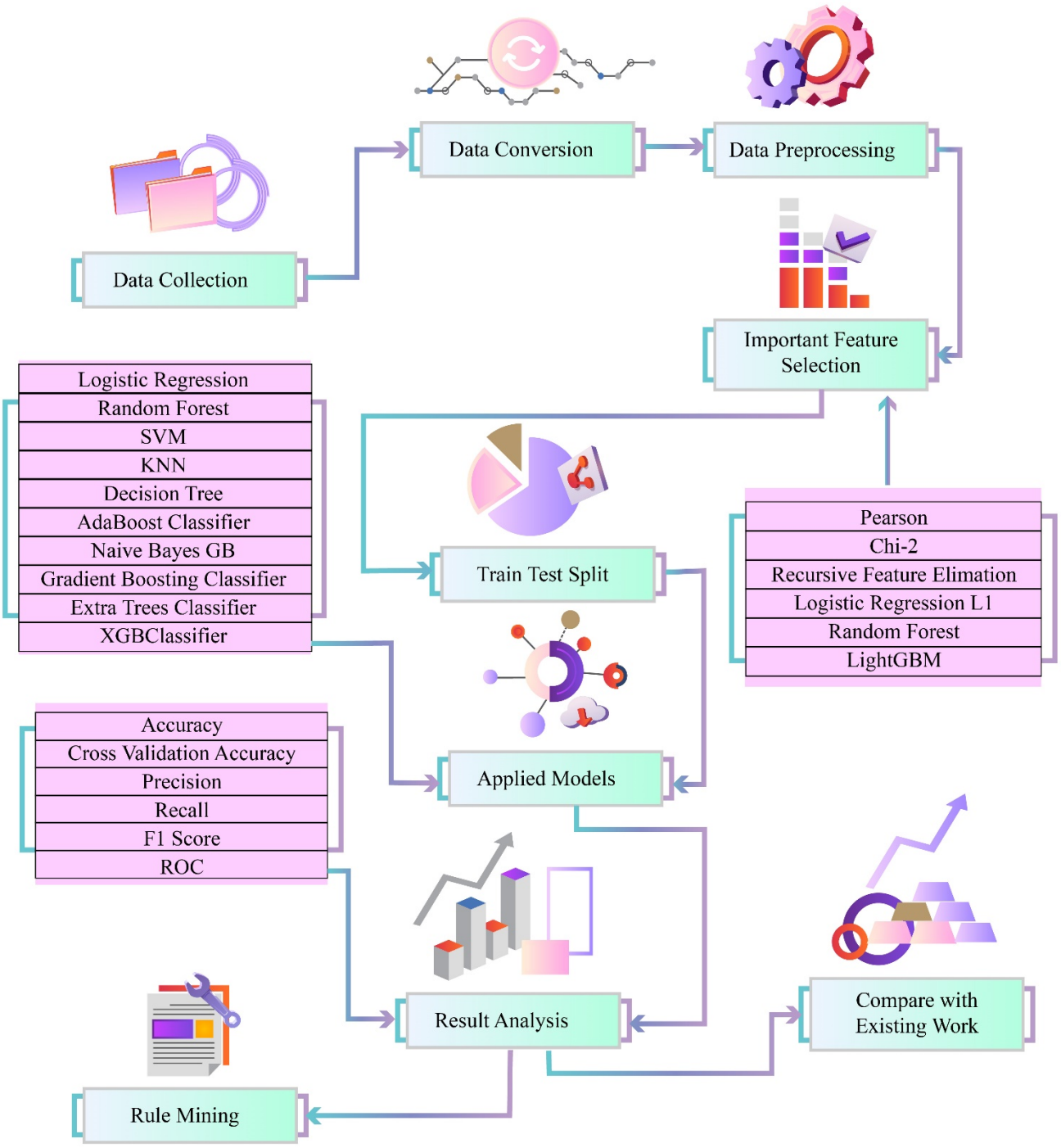
Let's quickly go over the methodology that was applied in this research now that this point has been established. The section on the methodology, which will be presented after this part, will contain additional information on this topic.

3. Methodology

The ten distinct primary sections had to be completed in order for this study to be finished. One section of the "Data Collection" section is devoted to the presentation and discussion of the specifics of the dataset's description. The data set's past has also been meticulously reviewed and dissected. The string data is converted into numerical data in the "Data Conversion" section. The report's "Data Preprocessing" section has been updated with the necessary data preprocessing techniques. Additionally, six different models have been used to determine which important features are most advantageous in the section titled "Important Feature Selection". In the "Train Test Split" portion of the document, the dataset has been split into a train set and a test set so that the experiment can be run on both of them. The "Applied Model" portion of the study contains a list of all 10 models that were used to evaluate the dataset and forecast the chance of developing early diabetes. The "Result Analysis" Section now discusses the model whose performance was determined to be the best overall after all models were evaluated and the article's section on "Rule Mining" briefly mentioned how datasets frequently correlate with one another.. In order to accurately forecast the onset of early diabetes using machine learning, the

results of the model that performed the best have been examined and contrasted with those of previously published work. The systematic method followed in this inquiry is shown in Figure 1.

Figure 1



Ten main sections were required for this study. In the "Data Collection" section, a description of the dataset is displayed and discussed. The history of the data set has also been examined. In "Data Conversion," string data is transformed into numerical data. Data preprocessing techniques are included in "Data Preprocessing". Six models were employed in the "Important Feature Selection" section to choose the top features. To enable the experiment to be run on both sets, the dataset has been split into a train set and a test set under the "Train Test Split" section. Ten models were used in the "Applied Model" section to forecast early diabetes. The evaluation of all models is followed by a discussion of the

model that performed the best overall in the "Result Analysis" and in the section Datasets frequent association was mentioned in "Rule Mining". In order to accurately predict early diabetes using machine learning, the model that performed the best was looked at and contrasted to past studies.

The dataset Section has now started the working process of this study in order to discuss the dataset's attributes in the following way.

A. Dataset

The signs and symptoms of diabetic individuals who have just received a diagnosis or who are at risk of acquiring the disease are detailed in this dataset. For this information, the patients at the Sylhet Diabetes Hospital in Sylhet, Bangladesh, completed direct questionnaires, and a medical expert approved the project before it was carried out. Diabetes is linked to 520 different patients and 16 different characteristics. Early Stage Diabetes Risk Prediction Dataset (ESDRPD) has been collected from kaggle [32]. There is one continuous attribute in addition to fifteen different categories of attributes. The dataset includes a total of 15 features, one of which is the target variable defined as class. The dataset's executive summary is displayed in Table 1.

Table 1: The Dataset's Summarized Description

Features	Count	Mean	Std.	Min.	75%	Max.
Age	520	48.02	12.15	16	57	90
Gender	520	0.63	0.48	0	1	1
Polyuria	520	0.49	0.50	0	1	1
Polydipsia	520	0.44	0.49	0	1	1
Sudden Weight Loss	520	0.41	0.49	0	1	1
Weakness	520	0.58	0.49	0	1	1
Polyphagia	520	0.45	0.49	0	1	1
Genital thrush	520	0.22	0.41	0	0	1
Visual Blurring	520	0.44	0.49	0	1	1
Itching	520	0.48	0.50	0	1	1
Irritability	520	0.24	0.42	0	0	1
Delayed Healing	520	0.45	0.49	0	1	1
Partial Paresis	520	0.43	0.49	0	1	1
Muscle Stiffness	520	0.37	0.48	0	1	1
Alopecia	520	0.34	0.47	0	1	1
Obesity	520	0.16	0.37	0	0	1
Class	520	0.61	0.48	0	1	1

The statistical data analysis of the dataset has been accomplished in the following section to ease in a better understanding of the dataset.

B. Statistical Data Analysis

As was mentioned earlier, the diabetes dataset is made up of 520 different instances and 16 different features. There is one continuous attribute in addition to fifteen different categories of attributes. An explanation of some characteristics that are associated with medicine is provided below.

(1) Age

This component determines the age of the individual.

(2) Gender

This aspect pertains to the gender of the individual who is participating in the activity [33] . There are 328 men, which is 65.60% of the total population, but there are 192 more women than men. This means that women outnumber men by 192. As a consequence of this, the proportion of females to males in the entire population is 38.40%. Positive representation of women (females make up 54.06% of the population, while males make up 45.94%) and negative representation of men (45.94% of the population) may be seen in the gender distribution (females make up 9.5%, while men make up 90.50%).

(3) Polyuria

The disorder known as polyuria causes a person to urinate more frequently than is normal and to pass excessive or unusually big amounts of urine each time [34]. More than 3 liters of urine per day are frequently passed, which is referred to as polyuria. This is in contrast to the typical daily output of 1 to 2 liters of urine for adults. This feature determines whether or not the individual had an issue with urinating an excessive amount. Distribution of Polyuria: Positive (yes = 75.94%, No = 24.06%) and Negative (yes = 7.5%, No = 92.50%)

(4) Polydipsia

The medical term for increased thirst is polydipsia. A persistent, abnormal drive to drink fluids is known as excessive thirst [34]. It is a response to your body losing fluid. This feature records whether or not the participant drank excessively or experienced excessive thirst. One of the main early indicators of diabetes is polydipsia, or excessive thirst [35]. Positive polydipsia cases account for 70.31% of all cases, while negative polydipsia cases account for 96% of all cases.

(5) Sudden Weight Loss

When a person loses a considerable amount of weight without making any changes to their eating habits or exercise routines, this is considered to be unexplained weight loss. Those with type 2 diabetes are not immune to it, although type 1 diabetics are more likely to experience it. Both a positive (Yes = 58.75%, No = 41.25%) and negative (Yes = 14.50%, No = 85.50%) distribution can be seen with abrupt weight loss.

(6) Weakness

If an individual possesses this characteristic, it can be deduced whether or not they had ever experienced a time in their life when they felt helpless or unable of doing something. On the subject of the weakness, the replies are evenly split between those who are positive (Yes = 68.12% and No: 31.87%) and those that are negative (Yes = 43.50% and No = 56.50%).

(7) Polyphagia

Polyphagia, also referred to as hyperphagia, is characterized by an overwhelming and unquenchable need to consume food. Throughout the course of the study, this trait helped researchers to determine whether or not a subject ever experienced excessive or intense hunger.Polyuria can be either positive (yes, which equals 59.06%, or negative (no, which equals 40.94%) or negative (yes, which equals 24%, or negative, which equals 76.50%). Polyphagia is characterized by an abnormally high level of hunger that leads to a considerable and continuing increase in appetite. It is a primary indicator of diabetes [36], as well as one of its main symptoms.

(8) Genital Thrush

This quality reveals whether or not the individual was suffering from a yeast infection during the course of the investigation. Thrush is the medical term for an infection caused by yeast (candida albicans) [36]. Candida albicans is able to thrive in environments that are more conducive to its growth when there is a significant amount of sugar present (Thrush, 2019). The percentage of people who have genital thrush can be split down as follows: negative (Yes, 16.5% and No, 83.5%), and positive (Yes, 25.94% and No, 74.06%)

(9) Visual Blurring

Blurred vision, often known as a loss of visual acuity, makes it impossible to see fine details clearly. This makes it impossible to read small print. Cloudiness in one's vision is usually brought on by fluctuations in one's blood sugar [37]. It is possible for a number of eye illnesses, such as nearsightedness or farsightedness, which weaken the eye's capacity to concentrate, to bring about a condition known as blurred vision. This feature will record information regarding the participant's vision, including whether or not they experienced a period of obstructed vision. The two categories of participant responses for times when they experienced foggy vision were "Positive" (Yes = 54.69% and No = 45.31%) and "Negative" (Yes = 29.00% and No = 71.00%) respectively.

(10) Itching

Whether or not the individual had an episode of itching is recorded by this feature. The two different participation categories are known as "Positive" (Yes = 48.12% and No = 51.88%) and "Negative" (Yes = 49.50% and No = 50.50%) respectively.

(11) Irritability

This feature determines whether or not the individual had a fit of irritation at any point during their participation [37]. The two distinct groups of people who took part in the study are referred to as "Positive" (where Yes = 34.38% and No = 65.62%) and "Negative" (where Yes = 8% and No = 92.00%) accordingly.

(12) Delayed Healing

This feature determines whether or not the subject experienced slowed healing after being injured and records that information [38]. The incidence of delayed healing according to: Negative (Yes = 43%, and No = 57%), and positive (Yes = 47.81%, and No = 52.19%)

(13) Partial Paresis

The disease known as paresis is characterized by a reduction in the patient's ability to move voluntarily [39]. It is possible for it to be a symptom of diabetes. Positive (Yes = 60%, No = 40%), and Negative (Yes = 16%, and No = 84%) are the proportions of people who had, respectively, had an episode of muscular weakness.

(14) Muscle Stiffness

Stiffness in the muscles is characterized by a feeling of constriction in the affected area, which frequently results in discomfort and makes it difficult to move. Muscle stiffness can be brought on by misuse of a particular muscle, or it might be an early warning sign of an underlying health problem. If a person experienced a period of muscle stiffness, this characteristic records it. Below is a breakdown of the participants' proportion who reported having a case of muscle stiffness: Positive (Yes = 42.19%, No = 57.81%) and Negative (Yes = 30%, No = 70%).

(15) Alopecia

Diabetes patients have a higher chance of acquiring alopecia areata. Any area of the body that has alopecia will experience hair loss. This factor determines whether or not the individual had hair loss during their time inside the study. Those who experienced hair loss make up a total of Positive (Yes = 24.38%, No = 75.62%) and Negative (Yes = 50.50%, No = 49.50%).

(16) Obesity

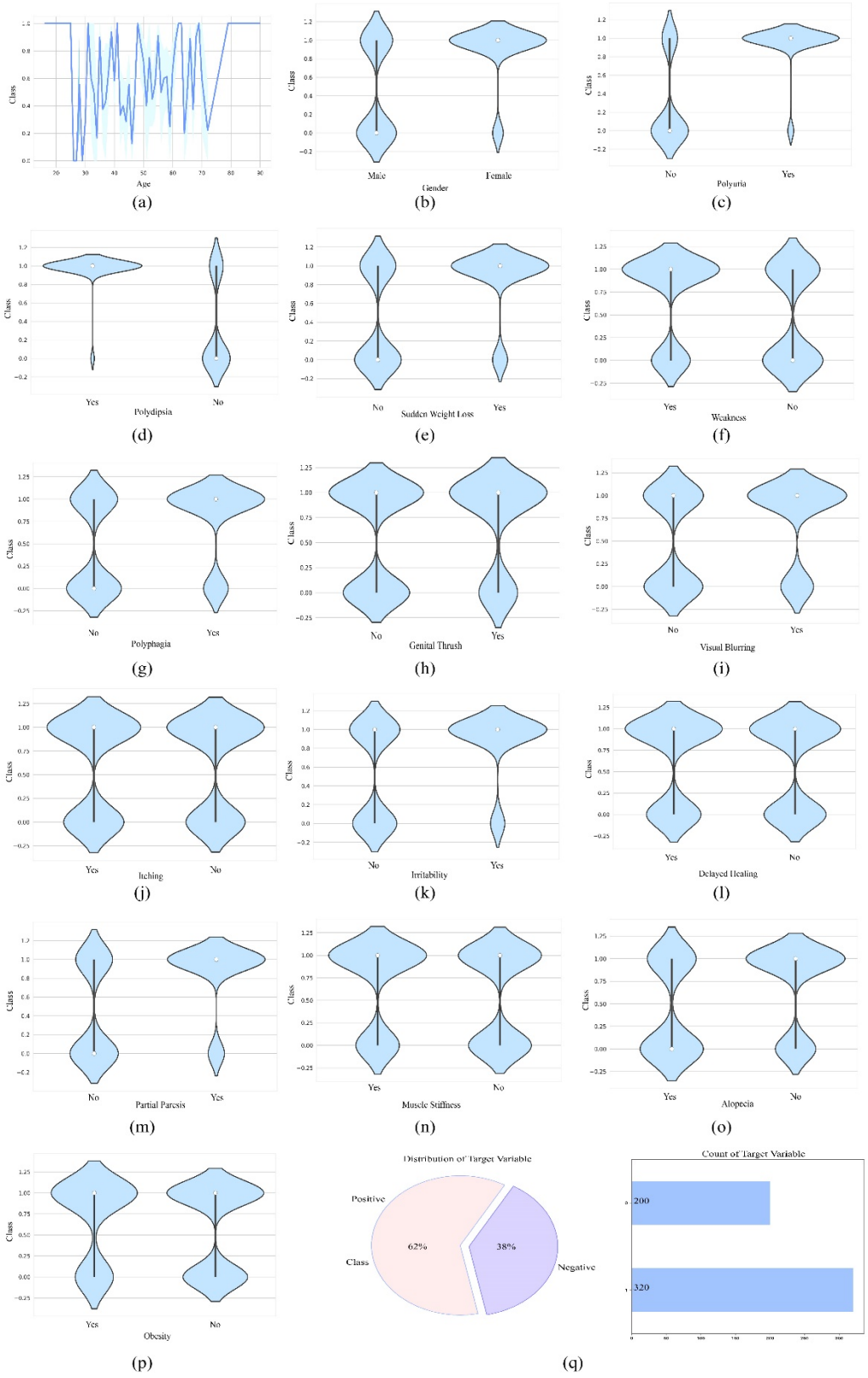
This attribute determines whether or not the individual is deemed to be obese. The percentage of participants that are positive (Yes = 19.06%, No = 80.94%) and negative (Yes = 13.50%, No = 86.50%).

(17) Class

If a person has type 2 diabetes or not can be determined by this trait. 62% of the individuals had diabetes type 2, the most prevalent form of the illness.

With the exception of age, which is a number, all the characteristics are notional. The distribution of all targeted variable is shown in Figure 2 (a) to (q).

Figure 2: (a) to (q) Distribution of All Target Variable

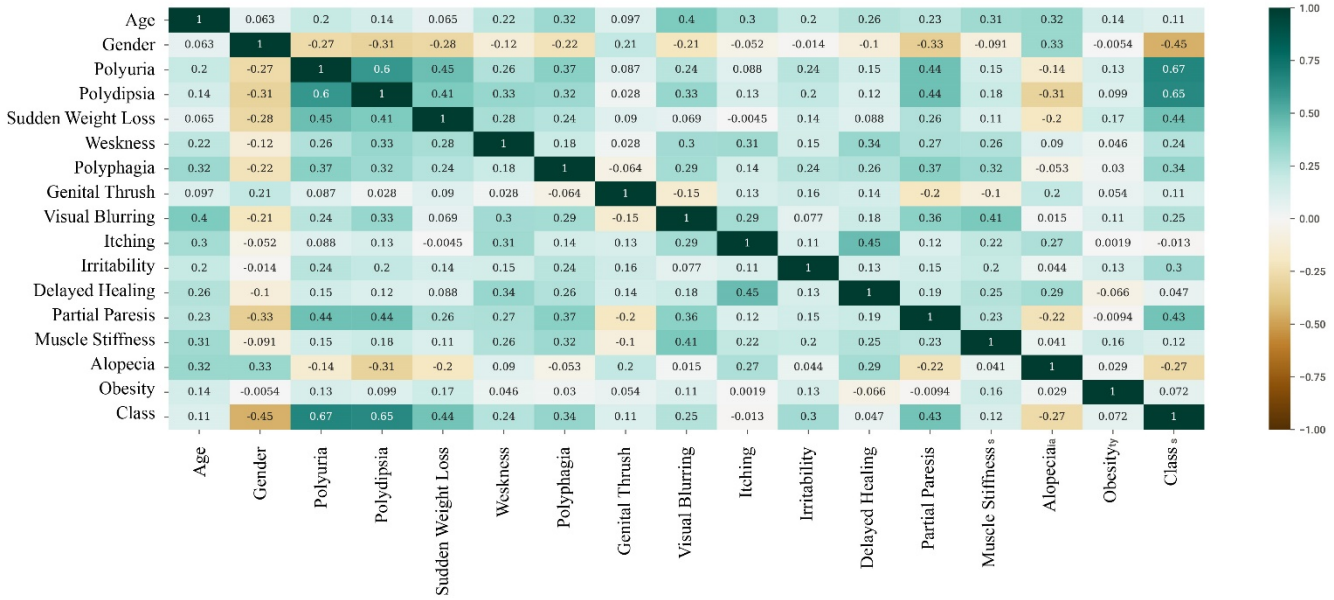


Features Correlation

The importance of the feature link with diabetes cannot be overstated when it comes to early diabetes predictions. In our data, the correlations scale from -1 to 1 for the following variables are as follows: age has a correlation of 0.10, gender has a correlation of -0.44, polyuria has a correlation of 0.66, polydipsia

has a correlation of 0.64, sudden weight loss has a correlation of 0.43, weakness has a correlation of 0.24, polyphagia has a correlation of 0.34, genital thrush has a correlation of 0.11, visual blurring has a correlation of 0. Figure 3 is an instance of the feature correlation that has been demonstrated to be associated with each others.

Figure 3: Illustration of all Features Correlation



In order to complete mining association rules, association rule mining will now be performed on the dataset. Before continuing with that step, the string data must first undergo data conversion so that it can be converted into a numerical value.

C. Data Preprocessing

In this part, data preprocessing methods have been employed in the following manner in preparation for the subsequent simulation that would predict early onset of diabetes. Converting the target's "Class" values to their corresponding numerical values, in other words, changing "positive" to "1" and "negative" to "0." Separating the Target (Class) feature from the rest of the 15 characteristics, and storing them. Furthermore, data normalization have been applied for the continues feature “Age”. The next step, which is to determine the most important feature, will include using six distinct methods, as described in the following section.

(1) Important Feature Selection

(a) Pearson

Pearson The correlation method evaluates the linear relationship that exists between two characteristics and generates a value that can range anywhere from -1 to 1 to show the degree to which the two characteristics are related to one another. This value shows the degree to which the linear relationship that exists between the two characteristics can be described as a correlation. The construction of a correlation matrix requires the use of this quantity. The construction of a correlation matrix is made possible by the use of correlation. By computing the relationship that exists between each feature and the goal variable, this determines the degree to which the two features are interdependent on one another. This determines the degree to which the two features are dependent on one another. When this stage of

the process is finished, the next step is to discover the attribute that has the greatest substantial influence on the variable that is being sought.

(b) Chi-2

The chi-2 test is a type of statistical analysis that allows you to compare an investigation's actual results to what was anticipated. This test will assess whether a disparity between actual and predicted data can be attributed to random variation or whether it can be attributed to a relationship between the variables that are the subject of the research. The k value is set to 10 for the selected dataset. This test's goal is to establish whether the relationship between the variables that are the subject of the study and the discrepancy can be determined.

(c) Recursive Feature Elimination

Recursive Feature Elimination (RFE) is a technique that involves selecting features that are appropriate for a model and gradually removing the weakest features until the necessary number of features is obtained. RFE is the common abbreviation for recursive feature elimination. In this study the parameters are set as following, estimator = LogisticRegression(), n_features_to_select = 100, step = 10, and verbose = 5.

(d) Logistic Regression L1 (LR L1)

Within the area of machine learning, the use of L1 regularized logistic regression is currently considered to be standard practice. This method is applicable to a wide range of classification issues, in particular those that involve a considerable number of distinct attributes. It is vital to discover a solution to a problem involving convex optimization before employing L1 regularized logistic regression since this type of problem requires it. Hence the parameters are set to the penalty = l2, and threshold = 1.25 × median.

(e) Random Forest (RF)

A random selection of the features and the observations from the dataset are used to build each decision tree that makes up a random forest, which can have anywhere between 400 and 2,000 decision trees. In a random forest, there could be anywhere between 400 and 2,000 decision trees. A random forest can have anything between 400 and 12,000 decision trees at any given time. Therefore the parameters are set to the n_estimators = 100, and threshold = 1.25 × median.

(f) LightGBM

LightGBM, a gradient boosting framework that prioritizes teaching using the tree-based learning approach. The LightGBM creates trees vertically, in contrast to previous techniques. Most tree-growing algorithms develop their trees horizontally. This suggests that, unlike other methods, the LightGBM approach constructs trees leaf-wise rather than level-wise. So the parameters are set to the n_estimators = 500, learning_rate = 0.05, num_leaves = 32, colsample_bytree = 0.2, reg_alpha = 3, reg_lambda = 1, min_split_gain = 0.01, and min_child_weight = 40.

Table 2: The Summarized Output of the Six Model for the Important Feature Selection

SL	Feature	Pearson	Chi-2	RFE	LR L1	RF	LightGBM	Total
1	Polyuria	True	True	True	True	True	True	6
2	Polydipsia	True	True	True	True	True	False	5
3	Gender	True	True	True	True	True	False	5
4	Weakness	True	True	True	False	False	True	4
5	Visual Blurring	True	True	True	False	False	True	4
6	Sudden Weight Loss	True	True	True	False	True	False	4
7	Partial Paresis	True	True	True	False	True	False	4
8	Itching	True	False	True	True	False	True	4
9	Irritability	True	True	True	True	False	False	4
10	Age	True	False	True	False	True	True	4
11	Delayed Healing	True	False	True	False	False	True	3
12	Polyphagia	True	True	True	False	False	False	3
13	Genital Thrush	True	False	True	True	False	False	3
14	Alopecia	True	True	True	False	False	False	3
15	Muscle Stiffness	True	False	True	False	False	False	2
16	Obesity	True	False	True	False	False	False	2

From Table 2, the first ten elements with a value that is more than or equal to four in total have been chosen as essential features that will be placed to models to predict early diabetes. These characteristics will be used to predict whether a person will get diabetes or not. These are Polyuria, Polydipsia, Gender, weakness, visual blurring, sudden weight loss, partial paresis, Itching, Irritability, and Age.

(2) Train Test Split

In order to carry out the application of the models and perform the task of predicting early diabetes, A train set and a test set of the 10 significant characteristics that were chosen have been created. As a result, 20% of the dataset was used for testing, while 80% of the dataset was used for training. The summarized description of the train test split is shown in Table 3.

Table 3: Description of the Train Test Split Dataset

Name	Description
Proportion of train set	80%
Percentage of the test set	20%
Amount of occasions of a train set	416
Instances in the test set	104
Number of patients overall	510

(3) K-Fold Cross Validation

The drawbacks of the hold-out method can be minimized by using k-Fold cross-validation. The "test just once bottleneck" can be avoided with the use of k-Fold, which offers a fresh approach to dataset segmentation.

1. Choose k folds, the number of folds. (If at all possible, the dataset should be split into k equal halves).
2. Then, k-1 folds should be used as the practice set. The test set will consist of the remaining fold.
3. Use training set to put the model through its paces.
4. In order to use cross-validation, a new model must be trained separately from the model that was trained in the prior iteration.
5. Verify your findings using the test set.
6. Maintain a record of the validation's results.
7. Steps 3-6 must be repeated K times.

In this analysis, K was found to have a value of 8 for each of the 10 different models.

D. Applied Models

(1) Decision Tree (DT)

An example of a decision support tool is a decision tree, which uses a tree-like model to represent decisions and the likely effects of those actions. The results of random events, resource costs, and resource utility are a few examples of these potential implications. This can be used to display an algorithm that is nothing more than a set of conditional control statements. Decision trees, or more precisely decision analysis, are a common tool in the field of operations research for identifying the strategy that has the highest likelihood of success. In the area of machine learning, decision trees are a common tool. In this case the criterion is set to 'gini'.

(2) Random Forest Classifier (RFC)

A random forest is a classification technique composed of numerous independent decision trees. It attempts to produce an uncorrelated forest of trees whose forecast by committee is more accurate than that of any individual tree using bagging and feature randomness when generating each individual tree. It is thought that doing this will result in a prognosis that is more precise. Therefore the parameters are set as follows, where criterion was 'gini', and n_estimators was 100.

(3) Support Vector Machine (SVM)

The support vector machine is a widely used and flexible supervised machine learning technique (SVM). With its aid, activities involving classification and regression can both be completed. However, the categorization task will be the focus of debate in this thread. It is typically seen as being optimal for medium- and small-sized data sets. Finding the ideal hyperplane that divides the data points into two components linearly while also maximizing the margin is the major goal of the support vector machine (SVM). So, the kernelis set to the 'linear', and random_state is set to the 0 for the better performance.

(4) XGBoost Classifier (XGBC)

The gradient boosted trees technique is extensively used and is successfully implemented in a piece of open-source software called XGBoost. Gradient boosting, a method of supervised learning, combines the forecasts of a number of less reliable and simpler models in an effort to provide an accurate forecast

of a target variable. Regression trees act as the weak learners when gradient boosting is employed for regression. In each of these trees, a leaf that stores a continuous score is connected to each input data point. By using a convex loss function that is based on the difference between the anticipated and target outputs plus a penalty term for the model's complexity, XGBoost can minimize a regularized (L1 and L2) objective function (in other words, the regression tree functions). By adding new trees to the mix that make predictions regarding the residuals or errors produced by prior trees, the training operation is carried out iteratively. The final prediction is then created by combining these new trees with the previous trees. Gradient boosting is the name of the technique, and it refers to the way it lessens the amount of information that is lost as new models are added. Therefore the parameters are set as follows, objective = reg:linear, colsample_bytree = 0.3, learning_rate = 0.1, max_depth = 5, alpha = 10, and n_estimators = 10.

(5) K-Nearest Neighbor (KNN)

The KNN algorithm, a type of supervised machine learning technique, is straightforward and effective for both classification and regression issues. Although it is easy to build and comprehend, it has a big drawback in that it becomes substantially slower as data usage increases. So n_neighbors is set to the1-10, metric is set to the 'minkowski', and p is set to 2.

(6) Gaussian Naïve Bayes (GNB)

The Gaussian Naive Bayes algorithm is used to illustrate a probabilistic classification method. This strategy is based on the Bayes theorem and strong independence presumptions. When working with continuous data, a typical approach is to assume that the continuous values associated with each class are distributed according to a normal (or Gaussian) distribution. This is carried out to facilitate working with continuous data. On the likelihood of the qualities, we'll proceed with the following assumption: Continuous valued features and models are believed to individually correspond to a Gaussian distribution in the Gaussian Naive Bayes approach (also known as a normal distribution).

(7) AdaBoost Classifier (AdaBC)

A meta-estimator called an AdaBoost [34] classifier operates by first fitting a classifier on the initial dataset, and then fitting additional copies of the classifier on the same dataset with the weights of incorrectly classified instances adjusted in a way that causes subsequent classifiers to focus more on challenging cases. This process is repeated till the precision is achieved. Up until the required level of categorization accuracy, this process is repeated as often as required. Until the most accurate classifier is found, this process is done many times.

(8) Logistic Regression (LR)

The appropriate regression analysis to utilize when a dependent variable is dichotomous is logistic regression (binary). The logistic regression, like all regression studies, is a predictive analysis. We apply logistic regression to characterize the data and to explain the association between one dependent binary variable and one or more independent nominal, ordinal, interval, or ratio-level variables. Where random_state is set to the Zero, and penalty is set to 12.

(9) Gradient Boosting Classifier (GBC)

The fields of regression and classification are just two examples of potential applications for the machine learning approach known as "gradient boosting." Other potential applications include many

more. It offers a prediction model in the shape of a collection of simple prediction models, the bulk of which are decision trees. The accuracy of models of this type is typically considered to be somewhat lacking in common consensus [40],[41]. When a decision tree is the weak learner, the technique that results is known as gradient-boosted trees, and it frequently outperforms random forest. When a decision tree is the weak learner, gradient-boosted trees are created. When a decision tree is determined to be a poor learner, a gradient-boosted tree will be constructed [40],[41],[42]. It is created in a similar stage-by-stage fashion to existing boosting techniques, but it generalizes other techniques by enabling optimization of any differentiable loss function. It is made in a similar way to previous boosting techniques. As a result of performing this technique, a gradient-boosted trees model is created.

(10) ExtraTrees Classifier (ETC)

The program known as Train With AutoML is an application that puts into action an approach to ensemble supervised machine learning known as additional trees (short for excessively randomized trees). Excessively randomized trees is what the term "extra trees" refers to in its longer form. This method is also sometimes referred to by the term "extremely randomized trees," which is a shortcut for the longer phrase "highly randomized trees." This tactic makes use of decision trees, and the strategy's shorter term, "additional trees," alludes to the decision trees that are implemented in the tactic. Therefore the parameters are set as follows, n_estimators = 100, and random_state = 0.

E. Performance Metrics

In this study, we make use of four performance indicators that are very common: Accuracy, Precision, Recall, F1-Score (1-4).

$$A = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{---------- (1)}$$

$$P = \frac{TP}{TP+FP} \quad \text{---------- (2)}$$

$$R = \frac{TP}{TP+FN} \quad \text{---------- (3)}$$

$$F1-Score = 2 \times \frac{P \times R}{P+R} \quad \text{---------- (4)}$$

where TP, TN, FP, FN, and FPFN represent, respectively, true positive, true negative, false positive, and false negative. P, R, and A, which stand for Precision, Recall, and Accuracy, respectively, are similar to P, R, and A [43].

4. Results Analysis

A. Association Rule Mining (Unsupervised Technique)

Using the applymap function resulted in replacing "Yes" with 1 and "No" with 0. This was done since all categorical features were specified with either "Yes" or "No." Because of this, the dataset is now prepared for the application of association rule mining. The method of "association rule mining" includes identifying intriguing connections and correlations among enormous numbers of data objects. This can be accomplished by using various techniques. By revealing information about its occurrence

frequency, this rule provides insight into the frequency with which a particular item collection can be discovered in a dataset. To measure association, the following matrices used as follows:

(a) Support: The prior probability of P and Q serves as the rule's support [44]. The following is the equation for support:

$$(P \rightarrow Q) = \frac{(|P \cup Q|)}{n} \quad \text{---------- (1)}$$

The support in this case is denoted by "Sup". The overall number of transactions is "n".

(b) Confidence: The conditional likelihood that the consequent will occur is the antecedent. The conditional probability of Q given P is the rule's confidence constraint [45].

Here, con = Confidence.

$$con(P \rightarrow Q) = \frac{(P \cup Q)}{(P)} \quad \text{---------- (2)}$$

(c) Lift: By lift value, a rule's importance is determined. In essence, the rule's filters may be used to specify lift range. By dividing the actual and predicted confidence values of the rule, life is determined [46].

Here, Li = Lift.

$$Li(P \rightarrow Q) = \frac{(Tr\ contains\ both\ P \wedge Q)(Tr\ contains\ P)}{Fu\ of\ Tr\ contains\ Q} \quad \text{---------- (3)}$$

Here, Tr = Transactions, Fu = Function.

The rules applied to the model for diabetes prediction considering all features as all the consequences ultimate result is diabetes. The apriori rule mining technique was utilized in this study so that the association rule could be mined from our dataset. To explore the unsupervised technique on the selected data set as well as to open up the further future work door, this study analysis the association rule mining among the features. In order to accommodate this, the minimum support has been set to 0.1, and the minimum threshold has been established at 0.7. A total of 1150 rules were generated for the dataset as a result of the apriori rule mining technique. The top 50 rules from our dataset are displayed in Table 4.

Table 4: Top 50 Rules of Our Dataset

SL	Rule #	Antecedents	Consequents	Support	Confidence	Lift
1	861	Polyphagia, Delayed Healing, Sudden Weight Loss, Partial Paresis	Polyuria	0.102	0.981	1.978
2	912	Sudden Weight Loss, Polydipsia, Visual Blurring, Partial Paresis	Weakness	0.131	0.971	1.656
3	530	Muscle Stiffness, Sudden Weight Loss,	Weakness	0.121	0.969	1.652

		Partial Paresis				
4	1048	Polyphagia, Muscle Stiffness, Visual Blurring, Itching	Delayed Healing	0.119	0.969	2.108
5	1073	Polyuria, Visual Blurring, Sudden Weight Loss, Partial Paresis, Weakness	Polydipsia	0.119	0.969	2.162
6	1074	Polyuria, Visual Blurring, Sudden Weight Loss, Partial Paresis, Polydipsia	Weakness	0.119	0.969	1.652
7	972	Muscle Stiffness, Polydipsia, Visual Blurring, Partial Paresis	Weakness	0.115	0.968	1.65
8	1102	Polyuria, Visual Blurring, Polyphagia, Partial Paresis, Weakness	Polydipsia	0.112	0.967	2.157
9	851	Polyuria, Muscle Stiffness, Sudden Weight Loss, Partial Paresis	Weakness	0.112	0.967	1.648
10	701	Delayed Healing, Sudden Weight Loss, Partial Paresis, Polydipsia	Polyuria	0.112	0.967	1.948
11	659	Polyuria, Polyphagia, Visual Blurring, Sudden Weight Loss	Polydipsia	0.11	0.966	2.156
12	929	Muscle Stiffness, Polydipsia, Sudden Weight Loss, Partial Paresis	Weakness	0.108	0.966	1.646
13	840	Polyuria, Muscle Stiffness, Visual Blurring, Sudden Weight Loss	Weakness	0.106	0.965	1.645
14	688	Polyuria, Muscle Stiffness, Visual Blurring, Sudden Weight Loss	Polydipsia	0.106	0.965	2.153
15	1050	Polyuria, Visual Blurring, Sudden Weight Loss, Polyphagia, Weakness	Polydipsia	0.104	0.964	2.152
16	1087	Polyuria, Visual Blurring, Sudden Weight Loss, Weakness, Muscle Stiffness	Polydipsia	0.102	0.964	2.151
17	1088	Muscle Stiffness, Polyuria, Visual Blurring, Sudden Weight Loss, Polydipsia	Weakness	0.102	0.964	1.643
18	695	Polyuria, Itching, Sudden Weight Loss, Partial Paresis	Polydipsia	0.102	0.964	2.151
19	664	Polyuria, Polyphagia, Sudden Weight Loss, Itching	Polydipsia	0.102	0.964	2.151
20	627	Polyuria, Polydipsia, Visual Blurring, Sudden Weight Loss	Weakness	0.146	0.962	1.64
21	477	Muscle Stiffness, Partial Paresis, Polydipsia	Weakness	0.142	0.961	1.638
22	520	Visual Blurring, Sudden Weight Loss, Partial Paresis	Weakness	0.138	0.96	1.637
23	682	Polyuria, Visual Blurring, Partial Paresis, Sudden Weight Loss	Polydipsia	0.123	0.955	2.132
24	834	Polyuria, Visual Blurring, Partial Paresis,	Weakness	0.123	0.955	1.629

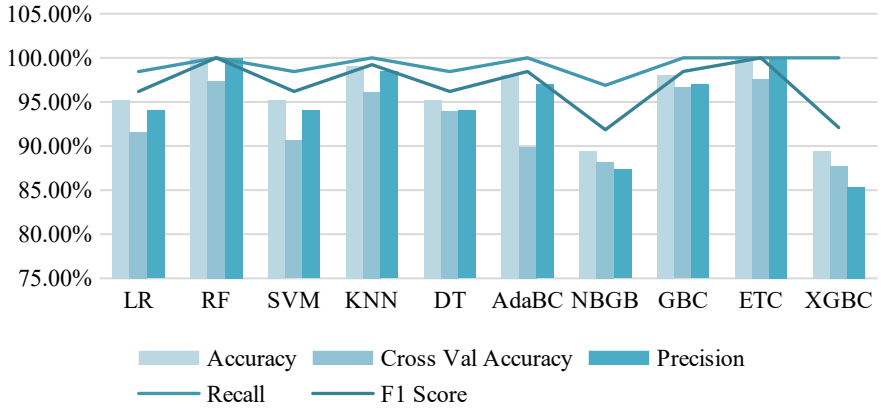
		Sudden Weight Loss				
25	774	Polyuria, Muscle Stiffness, Partial Paresis, Polydipsia	Weakness	0.119	0.954	1.626
26	751	Polyuria, Muscle Stiffness, Visual Blurring, Polydipsia	Weakness	0.117	0.953	1.625
27	285	Polyuria, Visual Blurring, Sudden Weight Loss	Weakness	0.154	0.952	1.624
28	626	Polyuria, Weakness, Visual Blurring, Sudden Weight Loss	Polydipsia	0.146	0.95	2.12
29	823	Polyuria, Polyphagia, Visual Blurring, Sudden Weight Loss	Weakness	0.108	0.949	1.618
30	957	Muscle Stiffness, Polyphagia, Polydipsia, Partial Paresis	Weakness	0.108	0.949	1.618
31	924	Itching, Weakness, Sudden Weight Loss, Partial Paresis	Polydipsia	0.104	0.947	2.114
32	1051	Polyuria, Visual Blurring, Sudden Weight Loss, Polyphagia, Polydipsia	Weakness	0.104	0.947	1.615
33	1121	Polyuria, Delayed Healing, Partial Paresis, Weakness, Itching	Polydipsia	0.102	0.946	2.112
34	320	Delayed Healing, Sudden Weight Loss, Partial Paresis	Polyuria	0.131	0.944	1.904
35	911	Sudden Weight Loss, Weakness, Visual Blurring, Partial Paresis	Polydipsia	0.131	0.944	2.108
36	308	Polyphagia, Sudden Weight Loss, Partial Paresis	Polyuria	0.16	0.943	1.901
37	422	Polyphagia, Sudden Weight Loss, Partial Paresis	Polydipsia	0.16	0.943	2.105
38	199	Polyuria, Visual Blurring, Sudden Weight Loss	Polydipsia	0.152	0.94	2.099
39	674	Polyuria, Polyphagia, Sudden Weight Loss, Partial Paresis	Polydipsia	0.15	0.94	2.097
40	678	Polyphagia, Polydipsia, Sudden Weight Loss, Partial Paresis	Polyuria	0.15	0.94	1.894
41	436	Itching, Sudden Weight Loss, Partial Paresis	Polydipsia	0.117	0.938	2.094
42	671	Polyphagia, Delayed Healing, Sudden Weight Loss, Polydipsia	Polyuria	0.112	0.935	1.885
43	832	Polyphagia, Weakness, Sudden Weight Loss, Partial Paresis	Polyuria	0.135	0.933	1.885
44	429	Visual Blurring, Sudden Weight Loss, Partial Paresis	Polydipsia	0.135	0.933	2.083
45	905	Polyphagia, Weakness, Sudden Weight	Polydipsia	0.135	0.933	2.083

		Loss, Partial Paresis				
46	743	Polyuria, Weakness, Visual Blurring, Partial Paresis	Polydipsia	0.156	0.931	2.078
47	849	Weakness, Delayed Healing, Sudden Weight Loss, Partial Paresis	Polyuria	0.104	0.931	1.877
48	1139	Delayed Healing, Polyphagia, Partial Paresis, Itching, Polydipsia	Polyuria	0.104	0.931	1.877
49	1055	Visual Blurring, Sudden Weight Loss, Polyphagia, Weakness, Polydipsia	Polydipsia	0.104	0.931	1.877
50	438	Itching, Muscle Stiffness, Sudden Weight Loss	Polydipsia	0.102	0.93	2.075

B. Applied Models Performance (Supervised Technique)

The results of the simulation demonstrated that the Logistic Regression model had an accuracy of 95.19%, an accuracy of 91.59% when subjected to cross-validation, a precision of 94.03%, a recall of 98.44%, and an F1 Score of 96.18%. Following that, the Random Forest model accomplished the following results: an accuracy of 100%, a cross-validation accuracy of 97.36%, a precision of 100%, a recall of 100%, and an F1 Score of 100%. After this, the model SVM attained an accuracy of 95.19%, an accuracy of 90.63% in cross validation, a precision of 94.03%, a recall of 98.44%, and an F1 score of 96.18%. Then The KNN model attained an accuracy of 99.04%, an accuracy of 96.15% in cross-validation, 98.46% in precision, 100.00% in recall, and a score of 99.22% on the F1 scale. After that, the Decision Tree model was able to achieve accuracy scores of 95.19%, cross-validation accuracy scores of 93.99%, precision scores of 94.03%, recall scores of 98.44%, and F1 values of 96.18%. Following this, the AdaBoostClassifier model accomplished an accuracy of 98.08%, a cross-validation accuracy of 89.90%, a precision of 96.97%, a recall of 100.00%, and an F1 Score of 98.46%. After that, the Naive Bayes GB model achieved an accuracy of 89.42%, an accuracy of 88.22% in cross-validation, a precision of 87.32%, a recall of 96.88%, and an F1 Score of 91.85%. After that, the Gradient Boosting Classifier model reached an accuracy of 98.08%, an accuracy of 96.63% in cross validation, a precision of 96.97%, a recall of 100.00%, and an F1 Score of 98.46%. Following that, the ExtraTreesClassifier model achieved an accuracy of 100.00%, a cross-validation accuracy of 97.60%, a precision of 100.00%, a recall of 100.00%, and an F1 Score of 100.00%. After that, the model ExtraTreesClassifier was able to achieve accuracy scores of 89.42%, cross-validation accuracy scores of 87.74%, precision scores of 85.33%, recall scores of 100.00%, and F1 Scores of 92.09

Figure 4: The Illustrative Example of How Well the Model Performed



The accuracy is represented by the color light blue, the accuracy of the cross-validation is represented by the color teal blue, the precision is represented by the color ocean blue, the recall is represented by the color blue, and the F1 score is represented by the color deep blue.

Figure 4 demonstrates that the model with the highest performance is ExtraTreesClassifier. The accuracy of this model is 100%, its cross-validation accuracy is 97.60%, its precision is also 100%, its recall is also 100%, and its F1 score is also 100%. In spite of the fact that Random Forest has also accomplished nearly the same level of success as ExtraTreesClassifier, with value accuracy of 100%, accuracy in cross-validation of 97.36%, precision of 100%, recall of 100%, and F1 Score of 100%. The total performance of all of the models is summarized in Table 5, which provides an overview of the situation.

Table 5: The Overall Performance of All of the Models Summed Up

Model	Accuracy	Cross Val Accuracy	Precision	Recall	F1 Score	ROC
LR	95.19%	91.59%	94.03%	98.44%	96.18%	94.22%
RF	100%	97.36%	100%	100%	100%	100%
SVM	95.19%	90.63%	94.03%	98.44%	96.18%	94.22%
KNN	99.04%	96.15%	98.46%	100%	99.22%	98.75%
DT	95.19%	93.99%	94.03%	98.44%	96.18%	94.22%
AdaBC	98.08%	89.90%	96.97%	100%	98.46%	97.50%
NBGB	89.42%	88.22%	87.32%	96.88%	91.85%	87.19%
GBC	98.08%	96.63%	96.97%	100%	98.46%	97.50%
ETC	100%	97.60%	100%	100%	100%	100%
XGBC	89.42%	87.74%	85.33%	100%	92.09%	86.25%

Following this, a brief discussion on the confusion matrix as well as the roc for each of the applied models will be presented in the following part.

5. Discussion

Figure 5 is a graphical representation of the confusion matrix that describes the anticipated results of an experiment. In addition to that, the ROC-AUC assessments of the model's performance have been shown

in this study. The term "receiver operating characteristic curve" (often abbreviated as "ROC curve") refers to a curve that shows the genuine positive rate on the ordinate of the graph and the false positive rate on the abscissa. Another common abbreviation for this type of curve is "ROC curve." It is the end result of combining the border values of numerous different areas into a single one. A measure of the likelihood that the computed score of the positive sample will be higher than the calculated value of the negative sample, the area under the ROC curve (AUC) is also known as the area under the receiver operating characteristic (ROC) curve. When samples are selected at random, it is possible to investigate both the benefits and drawbacks of using the prediction model. Figure 7, which once more depicts the ROC curve for an experiment, reveals that the average AUC value for our model ExtraTreesClassifier is 100%.

Figure 5: Graphical Representation of All Models Confusion Matrix

The ROC Curve for the Experiment

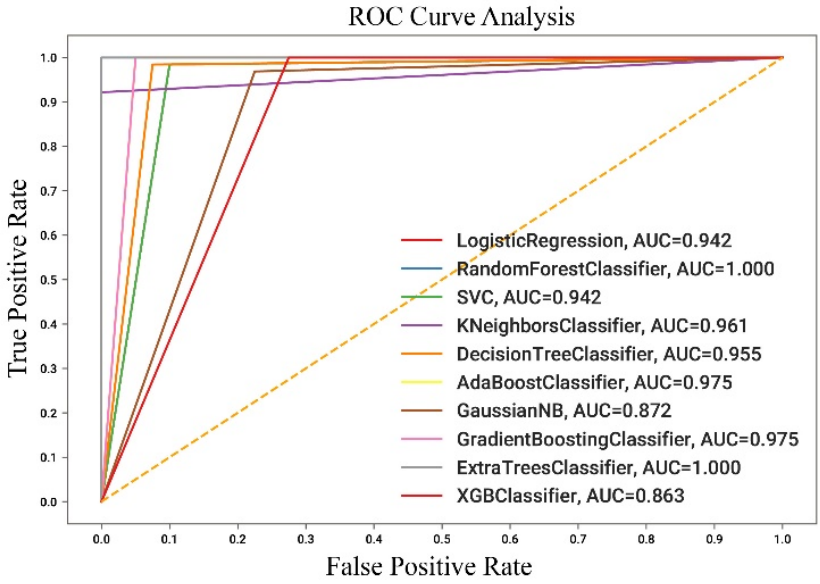


Figure 6 demonstrates that the Logistic Regression model was successful in achieving a ROC value of 94.22%. The next step was for the Random Forest model to reach a ROC value of 100%. After that, the ROC value for the SVM model was found to be 94.22%, while the ROC value for the KNN model was found to be 98.75%. The next model, Decision Tree, managed to get a ROC value of 94.22%. Following that, the AdaBoostClassifier model achieved a value of 97.50% for its ROC, whereas the Naive Bayes GB model achieved a value of 87.19% for its ROC. The subsequent Gradient Boosting Classifier model acquired a ROC value of 97.50%. The subsequent model, ExtraTreesClassifier, attained a ROC value of 100%. The subsequent XGBClassifier model reached a ROC score of 86.25%.

In the following section, we are going to compare the work that has already been done with the model that we have proposed.

6. Compare with the Existing Work

Boosted Regression is the name of the model that Tripathi and colleagues [13] utilized, and as a result of using it, they were able to attain an accuracy rate of 90.91%. Next, Alaa Khaleel et al. [18] used the Logistic Regression (LR), Naive Bayes (NB), and K-nearest Neighbor (KNN) models, and they were successful in achieving an accuracy of 94% for the LR model, 79% for the NB model, and 69% for the KNN model. This was accomplished by achieving an accuracy of 94% for the LR model. After that Zou et al. [14] The accuracy level that was attained using the Random Forest Can technique was 80.84%. Sisodia et al. are next. [16] used the Naive Bayes technique and was able to get an accuracy of 76.30%. After that, when Ramesh and his colleagues [17] used the Support Vector Machine, the accuracy increased to 83.20%. Perveen et al. are next. Using the utilization of the Hidden Markov Model, [18] was successful in accomplishing an accuracy of 86.9%. Following are Kavakitis et al. [20] The Support Vector Machine was used, and an accuracy success rate of 85% was achieved. Yahyaoui and co. are next. The accuracy for SVM, Random Forest (RF), and DL was 83.67% for SVM, 76.81% for RF, and 65.38% for DL in [22], which used these methods. After that, Sonar and his colleagues [23] employed the Decision Tree, the Naive Bayes model, and the Support Vector Machine. They were able to achieve accuracy levels of 85% for the Decision Tree, 77% for the Naive Bayes model, and 77.3% for the Support Vector Machine. [23] Subsequently, Sivaranjani et al. [24] used Random Forest (RF) and

Support Vector Machines (SVM) to reach an accuracy of 83.3% for RF and 81.4% for SVM, respectively. After that, Saha et al. [25] utilized a neural network, and as a result, they were able to accomplish an accuracy of 80.4%. Pavani et al. [27] proceeded to utilize the Random Forest method in conjunction with the Naive Base Method, and the results that both of these techniques produced had an accuracy of 80%. This is a description of both the Random Forest approach and the Naive Base Method. On the other hand, the ExtraTreesClassifier model that we proposed was able to obtain a level of accuracy of 100%. This had the direct effect of making our model perform substantially better than any of the earlier efforts. The summarized comparison of the early stage diabetics risk prediction model's performance is shown is Table 6.

Table 6: The Summarized Comparison of the Early Stage Diabetics Risk Prediction Model's Performance

Method	**Dataset Name**	**Description**	**Accuracy**
Tripathi et al. [47]	Pima Indian Diabetes Dataset from the Kaggle ML repository	Boosted Regression model	90.91%
Alaa Khaleel et al. [48]	The Pima Indian Diabetes (PIDD) dataset	Logistic Regression (LR) Naïve Bayes (NB) K-nearest Neighbor (KNN)	94%, 79%, 69%
Zou et al. [49]	PIDD-Pima Indians Diabetes Dataset	Random Forest Could	80.84%
Sisodia et al. [20]	The hospital physical examination data in Luzhou, China	Naive Bayes'	76.30%
Ramesh et al. [50]	PIDD-Pima Indians Diabetes Dataset	Support Vector Machine	83.20%
Perveen et al. [22]	PIMA Indian Diabetes Database	Hidden Markov Model	86.9%
Kavakiotis et al. [24]	accessed from the University of California, Irvine ML repository	SVM	85%
Yahyaoui et al. [26]	Canadian Primary Care Sentinel Surveillance Network (CPCSSN)	SVM RF DL	83.67% 76.81% 65.38%
Sonar et al. [27]	PIDD-Pima Indians Diabetes Dataset	Decision Tree Naive Bayes Support Vector Machine	85% 77% 77.3%
Sivaranjani et al. [28]	Pima Indian Diabetes Dataset from the Kaggle ML repository	Random Forest (RF) SVM	83% 81.4%
Saha et al. [29]	PIDD-Pima Indians Diabetes Dataset	Neural Network	80.4%
Pavani et al. [31]	Global data set	Random Forest Algorithm Naive Base Method	80% 80%
Our Applied Models [32]	Early Stage Diabetes Risk Prediction Dataset (ESDRPD)	Logistic Regression Random Forest SVM KNN Decision Tree	95.19% 100% 95.19% 99.04% 95.19%

		AdaBoostClassifier Naïve Bayes GB Gradient Boosting Classifier ExtraTreesClassifier XGBClassifier	98.08% 89.42% 98.08% 100% 89.42%

7. Conclusions

Data preparation methods such as converting and normalizing the data were performed in order to provide an accurate simulation of the dataset that was utilized for this investigation. This was done in order to create an accurate representation of the results of the study. In addition, association rule mining was utilized in order to identify the typical manifestations of diabetic symptoms and it also open a new door for the future work. After that, a total of six distinct methods were utilized in order to arrive at the final decision about the dataset's primary attribute. Following that, a total of ten distinct models were applied to the previously selected and highlighted dataset. We compared the outcomes that were produced by each model for us so that we could determine which one had provided us with the finest results all around. The ExtraTreesClassifier was able to attain the best feasible level of performance, as shown by the performance matrices, achieving a perfect score in every category (100% accuracy, 100% recall, 100% precision, and 100% F1) It is possible for us to declare that our ExtraTreesClassifier model even performs better than the work that is currently accessible. This study offers clinical physicians something new as well as something that can be of assistance to them. The lack of availability of larger databases was the primary challenge that we faced with our efforts. Yet, in order to optimize a model to its fullest potential, one must initially have access to a sizeable dataset. This is a prerequisite for the process. In the future, we are going to continue our investigation into the problems that are occurring right at this very moment.

Data Availability

Early Stage Diabetes Risk Prediction Dataset (ESDRPD) has been collected from kaggle: https://www.kaggle.com/datasets/ishandutta/early-stage-diabetes-risk-prediction-dataset

References

1. “Diabetes”, World Health Organization (WHO). https://www.who.int/health-topics/diabetes (Accessed on 4 March 2023).
2. N. Yuvaraj and K.R. SriPreethaa, “Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster”, Cluster Comput, vol. 22, no. 1, pp. 1–9, January 2019. https://doi.org/10.1007/s10586-017-1532-x
3. N. Sharma and A. Singh, “Diabetes Detection and Prediction Using Machine Learning/IoT: A Survey”, in Advanced Informatics for Computing Research, A.K. Luhach, D. Singh, P.-A. Hsiung, K.B.G. Hawari, P. Lingras and P.K. Singh, Eds., in Communications in Computer and Information Science. Singapore: Springer, pp. 471–479, 2019. https://doi.org/10.1007/978-981-13-3140-4_42
4. Md. M. Hassan et al., “A comparative assessment of machine learning algorithms with the Least Absolute Shrinkage and Selection Operator for breast cancer detection and prediction”, Decision Analytics Journal, vol. 7, p. 100245, June 2023. https://doi.org/10.1016/j.dajour.2023.100245
5. “A Comparative Study, Prediction and Development of Chronic Kidney Disease Using Machine Learning on Patients Clinical Records”, SpringerLink. https://link.springer.com/article/10.1007/s44230-023-00017-3 (Accessed on 16 June 2023).

6. F. Yasmin, M.M. Hassan, S. Zaman, S.T. Aung, A. Karim and S. Azam, "A Forecasting Prognosis of the Monkeypox Outbreak Based on a Comprehensive Statistical and Regression Analysis", Computation, vol. 10, no. 10, Art. no. 10, October 2022. https://doi.org/10.3390/computation10100177
7. Md. Mehedi Hassan, S. Mollick and F. Yasmin, "An unsupervised cluster-based feature grouping model for early diabetes detection", Healthcare Analytics, vol. 2, p. 100112, November 2022. https://doi.org/10.1016/j.health.2022.100112
8. A. Mir and S.N. Dhage, "Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare", in 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), pp. 1–6, August 2018. https://doi.org/10.1109/ICCUBEA.2018.8697439
9. D. Dutta, D. Paul and P. Ghosh, "Analysing Feature Importances for Diabetes Prediction using Machine Learning", in 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 924–928, November 2018. https://doi.org/10.1109/IEMCON.2018.8614871
10. A. Choudhury and D. Gupta, "A Survey on Medical Diagnosis of Diabetes Using Machine Learning Techniques", in Recent Developments in Machine Learning and Data Analytics, J. Kalita, V. E. Balas, S. Borah and R. Pradhan, Eds., in Advances in Intelligent Systems and Computing. Singapore: Springer, 2019, pp. 67–78. https://doi.org/10.1007/978-981-13-1280-9_6
11. M. A. Sarwar, N. Kamal, W. Hamid and M.A. Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare", in 2018 24th International Conference on Automation and Computing (ICAC), pp. 1–6, September 2018. https://doi.org/10.23919/IConAC.2018.8748992
12. I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research", Computational and Structural Biotechnology Journal, vol. 15, pp. 104–116, January 2017. https://doi.org/10.1016/j.csbj.2016.12.005
13. G. Tripathi and R. Kumar, "Early Prediction of Diabetes Mellitus Using Machine Learning", in 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), June 2020, pp. 1009–1014. https://doi.org/10.1109/ICRITO48877.2020.9197832
14. H. Lai, H. Huang, K. Keshavjee, A. Guergachi and X. Gao, "Predictive models for diabetes mellitus using machine learning techniques", BMC Endocr Disord, vol. 19, no. 1, Art. no. 1, October 2019. https://doi.org/10.1186/s12902-019-0436-6
15. A. Dinh, S. Miertschin, A. Young and S.D. Mohanty, "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning", BMC Medical Informatics and Decision Making, vol. 19, no. 1, Art. no. 1, November 2019. https://doi.org/10.1186/s12911-019-0918-5
16. D. Sisodia and D.S. Sisodia, "Prediction of Diabetes using Classification Algorithms", Procedia Computer Science, vol. 132, pp. 1578–1585, January 2018. https://doi.org/10.1016/j.procs.2018.05.122
17. K. Qu, Y. Luo, D. Yin, Y. Ju and H. Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques", Frontiers in Genetics, vol. 9, 2018. https://www.frontiersin.org/articles/10.3389/fgene.2018.00515 [Accessed on 22 October 2022]
18. F. Alaa Khaleel and A.M. Al-Bakry, "Diagnosis of diabetes using machine learning algorithms", Materials Today: Proceedings, July 2021. https://doi.org/10.1016/j.matpr.2021.07.196

19. N.P. Tigga and S. Garg, “Prediction of Type 2 Diabetes using Machine Learning Classification Methods”, Procedia Computer Science, vol. 167, pp. 706–716, January 2020. https://doi.org/10.1016/j.procs.2020.03.336
20. D. Sisodia and D.S. Sisodia, “Prediction of Diabetes using Classification Algorithms”, Procedia Computer Science, vol. 132, pp. 1578–1585, January 2018. https://doi.org/10.1016/j.procs.2018.05.122
21. S. Perveen, M. Shahbaz, K. Keshavjee and A. Guergachi, “Prognostic Modeling and Prevention of Diabetes Using Machine Learning Technique”, Sci Rep, vol. 9, no. 1, Art. no. 1, September 2019. https://doi.org/10.1038/s41598-019-49563-6
22. S. Perveen, M. Shahbaz, K. Keshavjee and A. Guergachi, “Prognostic Modeling and Prevention of Diabetes Using Machine Learning Technique”, Sci Rep, vol. 9, no. 1, Art. no. 1, September 2019. https://doi.org/10.1038/s41598-019-49563-6
23. Md. Maniruzzaman, Md. J. Rahman, B. Ahammed and Md. M. Abedin, “Classification and prediction of diabetes disease using machine learning paradigm”, Health Inf Sci Syst, vol. 8, no. 1, p. 7, January 2020. https://doi.org/10.1007/s13755-019-0095-z
24. I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas and I. Chouvarda, “Machine Learning and Data Mining Methods in Diabetes Research”, Computational and Structural Biotechnology Journal, vol. 15, pp. 104–116, January 2017. https://doi.org/10.1016/j.csbj.2016.12.005
25. Md. K. Hasan, Md. A. Alam, D. Das, E. Hossain and M. Hasan, “Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers”, IEEE Access, vol. 8, pp. 76516–76531, 2020. https://doi.org/10.1109/ACCESS.2020.2989857
26. A. Yahyaoui, A. Jamil, J. Rasheed and M. Yesiltepe, “A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques”, in 2019 1st International Informatics and Software Engineering Conference (UBMYK), November 2019, pp. 1–4. https://doi.org/10.1109/UBMYK48245.2019.8965556
27. P. Sonar and K. JayaMalini, “Diabetes Prediction Using Different Machine Learning Approaches”, in 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Mar. 2019, pp. 367–371. https://doi.org/10.1109/ICCMC.2019.8819841
28. S. Sivaranjani, S. Ananya, J. Aravinth and R. Karthika, “Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction”, in 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), March 2021, pp. 141–146. https://doi.org/10.1109/ICACCS51430.2021.9441935
29. P.K. Saha, N.S. Patwary and I. Ahmed, “A Widespread Study of Diabetes Prediction Using Several Machine Learning Techniques”, in 2019 22nd International Conference on Computer and Information Technology (ICCIT), Dec. 2019, pp. 1–5. https://doi.org/10.1109/ICCIT48885.2019.9038559
30. A.M. Posonia, S. Vigneshwari and D.J. Rani, “Machine Learning based Diabetes Prediction using Decision Tree J48”, in 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), Dec. 2020, pp. 498–502. https://doi.org/10.1109/ICISS49785.2020.9316001
31. K. Pavani, P. Anjaiah, N.V. Krishna Rao, Y. Deepthi, D. Noel and V. Lokesh, “Diabetes Prediction Using Machine Learning Techniques: A Comparative Analysis”, in Energy Systems, Drives and Automations, A. Sikander, D. Acharjee, C.K. Chanda, P.K. Mondal and P. Verma, Eds., in Lecture Notes in Electrical Engineering. Singapore: Springer, 2020, pp. 419–428. https://doi.org/10.1007/978-981-15-5089-8_41

32. "Early Stage Diabetes Risk Prediction Dataset". https://www.kaggle.com/datasets/ishandutta/early-stage-diabetes-risk-prediction-dataset (Accessed on 16 June 2023).
33. "Sex and Gender Differences in Prevention of Type 2 Diabetes", Frontiers in Endocrinology. https://www.frontiersin.org/articles/10.3389/fendo.2018.00220/full (Accessed on 18 March 2023).
34. "Initial Evaluation of Polydipsia and Polyuria", SpringerLink. https://link.springer.com/chapter/10.1007/978-3-030-52215-5_17 (Accessed on 18 March 2023).
35. "Why Am I Drinking So Much Water?", Verywell Health. https://www.verywellhealth.com/polydipsia-4783881 (Accessed on 18 March 2023).
36. "How pharmacists can encourage patient adherence to medicines", The Pharmaceutical Journal. https://pharmaceutical-journal.com/article/ld/how-pharmacists-can-encourage-patient-adherence-to-medicines (Accessed on 16 June 2023).
37. Paula C. Barata, Susan Holtzman, Shannon Cunningham, Brian P. O'Connor, Donna E. Stewart, "Building a Definition of Irritability From Academic Definitions and Lay Descriptions", 2016. https://journals.sagepub.com/doi/abs/10.1177/1754073915576228?journalCode=emra (Accessed on 18 March 2023).
38. R. Blakytny and E. Jude, "The molecular biology of chronic wounds and delayed healing in diabetes", Diabet Med, vol. 23, no. 6, Art. no. 6, June 2006. https://doi.org/10.1111/j.1464-5491.2006.01773.x
39. "sklearn.ensemble.AdaBoostClassifier", SciKit Learn. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html (Accessed on 18 March 2023).
40. "Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index", Journal of Infrastructure Systems, vol 26, no 1. https://ascelibrary.org/doi/abs/10.1061/(ASCE)IS.1943-555X.0000512 (Accessed on 16 June 2023).
41. "ESLII.pdf". https://hastie.su.domains/Papers/ESLII.pdf [Accessed on 16 June 2023]
42. "Using Machine Learning to Examine Impact of Type of Performance Indicator on Flexible Pavement Deterioration Modeling", Journal of Infrastructure Systems, vol 27, no 2. https://ascelibrary.org/doi/abs/10.1061/(ASCE)IS.1943-555X.0000602 (Accessed on 16 June 2023).
43. F. Yasmin et al., "PoxNet22: A Fine-Tuned Model for the Classification of Monkeypox Disease Using Transfer Learning", IEEE Access, vol. 11, pp. 24053–24076, 2023. https://doi.org/10.1109/ACCESS.2023.3253868
44. Kwang Hyeon Kim, Byung-Jou Lee and Hae-Won Koo, "Analysis of the Risk Factors for De Novo Subdural Hygroma in Patients with Traumatic Brain Injury Using Predictive Modeling and Association Rule Mining", Applied Sciences. https://www.mdpi.com/2076-3417/13/3/1243 (Accessed on 16 June 2023).
45. S. Sinisterra-Sierra, S. Godoy-Calderón and M. Pescador-Rojas, "COVID-19 Data Analysis with a Multi-Objective Evolutionary Algorithm for Causal Association Rule Mining", Mathematical and Computational Applications, vol. 28, no. 1, Art. no. 1, February 2023. https://doi.org/10.3390/mca28010012
46. J.J. Sonia, P. Jayachandran, A.Q. Md, S. Mohan, A.K. Sivaraman and K.F. Tee, "Machine-Learning-Based Diabetes Mellitus Risk Prediction Using Multi-Layer Neural Network No-Prop Algorithm", Diagnostics, vol. 13, no. 4, Art. no. 4, Jan. 2023. https://doi.org/10.3390/diagnostics13040723