Predicting Diabetes Risk Based on Machine Learning Methods

Abolfazl Joukar

November 17, 2024

Abstract

Nowadays a spike is seen in mortality linked to diabetes. Diabetes is known as one of the most critical human diseases in the contemporary world that has a serious impact on quality of life. Diabetes could be a chronic disease that happens either when the body cannot effectively use the insulin it produces or when the pancreas does not produce sufficient insulin. The early methods of forecasting diabetes help in avoiding health damage; however, inaccurate diabetes prediction can prove to be lethal. The machine learning algorithm can be very efficient in the prediction of diabetes. In this project, we use machine learning techniques to predict the patient having diabetes based on various health indicators. we used five classification algorithm such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machines(SVM), and K-Nearest Neighbor (KNN). The feature selection process involves selecting the most important features that contribute to the prediction of diabetes. Three methods also use for feature selection: Chi-square Test , Pearson Correlation and Principal Component Analysis (PCA). Success evaluation of classifiers was made using Accuracy, Precision, Recall, F1-Score and AUC metrics.

Contents

List of Tables

List of Figures

1 Introduction

1.1 Diabetes

1.1.1 What is Diabetes?

According to the CDC, Diabetes is a chronic (long-lasting) health condition that affects how your body turns food into energy. Diabetes mellitus (mellitus means sweet or honeylike) is a metabolic disorder caused by elevated blood glucose concentration and disordered insulin metabolism. Diabetes is known as a "siphon" in Greek, referring to the excessive passing of urine that is a symptom of untreated diabetes. Diabetes significantly increases the risk of a wide range of illnesses, including cardiovascular disease, blindness, kidney failure, and amputation of a body part such as a finger, toe, hand, foot, arm, or leg. Hemoglobin A1c (A1C), impaired fasting glucose, or oral glucose tolerance test values are used to make the diagnosis.

1.1.2 Common Symptoms of Diabetes

- Weight loss
- Blurred vision
- Fatigue
- Polydipsia or excessive thirst
- Polyphagia or excessive hunger
- Polyuria or excessive urine production
- Dry mouth (because of dehydration)
- Prone to frequent infections

1.1.3 Main types of Diabetes Mellitus

- Type 1 (previously known as IDDM or juvenile onset), 5 - 10% of diabetic population. It is usually caused by autoimmune destruction of the pancreatic beta cells, which produce and secrete insulin.
- Type 2 (previously known as NIDDM or adult onset), 90 - 95% of diabetic population. It is often asymptomatic for many years before diagnosis.

Table 1: Features of Type 1 and Type 2 Diabetes Mellitus.

Feature	**Type 1 Diabetes**	**Type 2 Diabetes**
Prevalence in diabetic population	5-10% of cases	90-95%of cases
Age of onset	<30 years	>40 years
Associated conditions	Autoimmune disease, viral infection, inherited factors	Obesity, aging, inactivity, inherited factors
Major defect	Destruction of pancreatic beta cells; insulin deficiency	Insulin resistance; insulin deficiency relative to needs
Insulin secretion	Little or none	Varies; may be normal, increased or decreased
Former names	Juvenile-onset diabetes Insulin-dependent diabetes	Adult-onset diabetes Noninsulin-dependent diabetes

1.1.4 Complications of Diabetes

Acute complications of Diabetes are a result of poorly controlled diabetes include diabetic ketoacidosis (acetone breath), in which hyperglycemia is accompanied by ketosis and acidosis. Patients with ketoacidosis may exhibit symptoms of both acidosis and dehydration. Over time, high blood glucose leads to a condition known as, chronic complications of Diabetes:

Macrovascular complications

about 20 to 30 percent of individuals with diabetes develop peripheral vascular disease (impaired blood circulation in the limbs), which increases the risk of claudication (pain while walking) and contributes to the development of foot ulcers.

Microvascular complications

- Diabetic retinopathy- resulting in visual impairments and, in rare cases, blindness.
- Diabetic nephropathy- leading to damage to kidneys specialized capillaries .
- Diabetic neuropathy- involves the peripheral nerves (peripheral neuropathy), leading to pain, numbness, tingling, or weakness in the hands, feet, or legs; and nerves that control body organs and glands (autonomic neuropathy) leading to sweating abnormalities, disturbed bladder function, delayed stomach emptying (gastroparesis), constipation, and cardiac arrhythmias.

1.1.5 Diabetes Test and Diagnosis

- A1C Test - It reflects your 3-month average blood glucose levels, but it shouldn't be used alone to diagnose diabetes.

- Fasting plasma glucose test (FPG) (fasting: no energy intake for at least 8 hours).
- Oral glucose tolerance test - determines a person's tolerance for an oral glucose load.
- Random blood sugar test it is usually done when the patient is experiencing symptoms of diabetes including polyuria, polydipsia, and unexplained weight loss and do not want to wait for an 8 hours fasting test.

Table 2: Diabetes Tests and their ranges for Type 1 and Type 2.

Result	**A1C Test**	**FPG Test**	**GTT Test**	**RBS Test**
Diabetes	6.5% or above	126 mg/dL or above	200 mg/dL or above	200 mg/dL or above
Prediabetes	5.7-6.4%	100-125 mg/dL	140-199 mg/dL	N/A
Normal	Below 5.7%	99 mg/dL or below	140 mg/dL or below	N/A

1.2 Machine Learning

Machine learning is a field of computer science focused on giving computers the capabilities to learn. The development of algorithms and statistical models enables computer systems to learn and improve from data without being explicitly programmed. The main goal of machine learning is creating algorithms and training computer systems to analyze and interpret patterns in data automatically, further make decisions and predictions, and based on input received, adapt its behavior.
Machine learning is a branch of Artificial Intelligence, broadly defined as a machines capability to imitate intelligent human behavior. Artificial intelligence systems are used to perform complex tasks in a way that is similar to how humans solve problems (Brown[1]).
Tom Mitchell, a computer scientist and professor at Carnegie Mellon University and a prominent figure that has contributed significantly to the field of machine learning, defines machine learning as:
"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E"(Mitchell [2]).
We can interpret Tom Mitchell's definition of Machine learning and apply it to the goal of credit scoring: A credit scoring model requires information on historical credit data (E) to train a model into learning to predict creditworthiness (T) and the performance measure (P) if the performance of predicting creditworthiness improves with the analysis of historical credit data.

1.2.1 Learning Methods

This section describes the three primary categories machine learning models fall into. The three categories are often referred to as machine learning problems: supervised, semisupervised, and unsupervised machine learning.

Supervised learning

We define supervised learning by its use of labeled datasets to train algorithms to predict outcomes accurately or classify data. Supervised learning are methods that attempt to discover the relationship between a target attribute (dependent variable) and input attributes (independent variable) (Maimon & Rokach [3]). When input data is fed to the model, it adjusts its weight until fitted appropriately. This happens as part of a crossvalidation process to ensure that the model avoids under or overfitting (IBM, n.d.).

There are many benefits to using supervised machine learning, and it can help organizations solve various real-world problems at scale. An example is classifying and separating spam mail from another folder in your inbox. Supervised machine learning uses a trained set of many cases consisting of features with value and resulting class. Another example is a dataset describing the color, top speed, and capacity of the trunks of a certain number of cars and the classification of the cars being a family car or not. The supervised machine learning algorithm then uses the data to infer functions relating the features of the car being a family car or not.

Reinforcement machine learning is very similar to supervised learning, but the difference is that the reinforcement machine learning algorithms are not trained using sample data. This type of model learns as it goes, by trial and error, receiving feedback in the form of rewards or penalties based on its actions. So when a sequence of successful outcomes occurs, these will be reinforced to develop the best recommendation for a specific problem.

Unsupervised learning

Unsupervised machine learning uses the algorithms to analyze and cluster unlabeled datasets. So, the algorithm is trained using just an input set, and desired results or feedback is given. The algorithm then finds the structure in the data by itself. Due to the data generated being so high, humans will not be able to analyze all the data, so an algorithm based on unsupervised learning finds hidden patterns or data groupings without the need of human intervention (IBM n.d.). This method is excellent for exploratory data analysis, and examples where it can be utilized are with customer segmentation and sales strategies. Using unsupervised learning, a seller/company can detect sales patterns and image recognitions, and behavior-based network security detection.

Semi-supervised learning

Semi-supervised learning considers the classification problem only when smaller labeled subsets of the observation have corresponding class labels. When training, it uses the smaller subset to guide classification and feature extraction from a large and unlabeled dataset. It uses labeled data to data to ground predictions and unlabeled data to learn shapes of the larger data distributions. Semi-supervised machine learning lies between the two, supervised and unsupervised learning, and using semi-supervised learning; one can achieve results with only fractions of the labeled data, which can save valuable time and money.

2 Methodology

2.1 Machine Learning Classification Techniques

In todays modern world, Machine learning is a highly advanced technical application in most industrial fields. Machine Learning is employed widely in many areas of industry, including marketing and business, social media, manufacturing, retail and customer service, cybersecurity, agriculture, finance, transportation, and healthcare (Shailaja et al. [5]). Healthcare professionals use Machine Learning to generate significant amounts of data and make useful clinical predictions and decisions. In recent years, there have been many efficient machine learning applications in the healthcare industry since medical professionals can process large datasets generated for each patient to discover patterns and learn from them.
Data mining is one of the most significant machine learning applications, and classification is a Data mining approach (Soofi & Arshad [6]). Although classification is the most used machine learning technique, it has limitations including how to handle missing data. In the project, Machine Learning models are applied to predict. These models are Logistic regression, Random forest, Decision tree, K-Nearest Neighbor (KNN) and Support Vector Machines(SVM). Classification is a sort of supervised learning.

2.1.1 Logistic Regression Classifier

Logistic regression (or logit regression) is a linear classification model used in different fields, including healthcare and social science . In this statistical classification model, there is a logistic function to model a binary dependent target using one or more independent variables. The logistic/sigmoid function used by this model to map any real value into a binary value of 0/1 is:

$$p(x) = \frac{\mathrm{e}^{\beta_0+\beta_1 X_1+\ldots+\beta_{k-1} X_k}}{\mathrm{e}^{\beta_0+\beta_1 X_1+\ldots+\beta_{k-1} X_k}+1} \tag{1}$$

where

$$y = \begin{cases} 1, & \text{if p(x)} \geq 0.5 \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

and

y : binary output (1/0, True/False, or diabetic/ non-diabetic) ,

x : input variables.

Logistic regression can be split into three different types depending on the dependent target variables: binominal (only two possible types), multinomial (three or more possible types), and ordinal (ordered categories).

2.1.2 Decision tree Classifier

A decision tree is one of the best-known classifiers due to its logical structure. A decision tree consists of nodes connecting, forming a tree with one single root node as a starting point. This node is called the "root" node and has no incoming edges. All nodes following the root node have a single incoming edge; if a node also has outgoing edges, that node is called an internal or test node. The internal node splits the data set into two or more subspaces according to a certain logic. Lastly, we have nodes called leaves, also known as terminal or decision nodes. Leaf nodes only have incoming edges, do not split like the internal nodes, and are assigned to a label based on which label is most appropriate. The classification process of a decision tree begins at the root node and continues by traversing through the internal nodes until a leaf is reached (Maimon & Rokach [3]).
The decision tree incorporates both nominal and numeric attributes. Each node is labeled with the attribute it tests, and the trees branches are labeled with its corresponding values. When using decision trees, impurity is measured to determine the split, and some parameters must be specified. The Gini Index and Entropy are the two most popular. The Gini index advantage is that it favors the production of pure over impure descendant nodes. When all possible candidate splits have been generated for one variable, the procedure is repeated for another variable. From the maximum set of possible single variable splits, the split with the largest purity is applied to generate a new partition. The Gini Index corresponds to the variance of the outcome and ranges from 1 to 0. The lower the impurity value, the more accurately each observation can be classified into the appropriate split (Kattan & Cowen [4]).
The Gini Index is defined by:

$$\text{The Gini Index} = \sum_{j=1}^{c} p_j^2 \tag{3}$$

Further parameters for the Decision tree are the determination of maximum depth, which is used to specify the maximum size of the constructed tree. A high-dimensional, deeper tree tends to have higher performance rates than less deep trees. Also, the number of features used at each split is a parameter usually determined for the decision tree. This parameter decreases the number of features used at each split, which decreases the chance of overfitting.

2.1.3 Random forest Classifier

Random forest is, as its name suggests, also a "tree" like method. The difference between the decision tree and the random forests method is that the random forests are a classifier consisting of a collection of tree-structured classifiers. i.e., the random forests classifier consists of multiple trained decision trees that together make a classification. This technique is called Bootstrap Aggregation, or "Bagging" for short. The bagging technique is to reduce the variance by averaging multiple samples, enhance accuracy when using random features and give ongoing estimates of the generalization error of the combined ensemble of trees, as well as estimates for the strength and correlation(L. Breiman [7]). The training procedure is similar to how a normal decision tree is trained, except that at each split in a tree, a random selection of features is selected, and from there, the future for the split is selected. The point of the random selection feature is to decrease the correlation between all the individual trees. Further, at least three parameters must be specified, and those are the number of features to consider at each node, the number of samples to select out of which the algorithm is constructed, and lastly, the maximum depth/maximum number of layers each individual tree can contain.

2.1.4 K-Nearest Neighbor (KNN) Classifier

K-Nearest Neighbours (KNN) is a relatively simple classification or re gression algorithm. When a new data point is introduced, its technique is to first locate the K closest data points from the training dataset and then use the class of each of those points to determine the class of the unknown data point.

The distance of these points can be calculated using any one of several distance metrics. In this analysis, the chosen distance metric is the Euclidean distance and a weighted version of majority voting is used to predict the unknown class.

This algorithm has a few important strengths including being simple to implement, flexible to feature and distance metric choices, handling multiclass problems naturally, and most importantly, it is quite accurate given a large, representative training set (Imandoust & Bolandraftar [8]). However, this algorithm has some significant flaws as well. Most importantly, since this is a lazy algorithm (work is only performed when an unknown data point is to be classified), it can be resource intensive in deployment. Every time a new point is to be classified, there is a large search problem through the full training set which must be solved to find the nearest neighbours. Another difficulty of this algorithm is selecting a meaningful distance function (Imandoust & Bolandraftar [8]). It is possible to mislead the algorithm if the function does not accurately reflect the similarity between points.

Given a classified point p, and an unknown point u with n features, the euclidean distance between them is calculated as (Imandoust & Bolandraftar [8]):

$$d(p,u) = \sqrt{\sum_{i=1}^{n}(u_i - p_i)^2}. \tag{4}$$

Given a set of K neighbours, x, and distances to those neighbours, d, the vote for class i, v_i can be computed as (Imandoust & Bolandraftar [8]):

$$d(p,u) = \sum_{j=1}^{K} \frac{d_j}{d_{\max}} [x_j \in \text{class i}]. \tag{5}$$

Note that the square brackets are denoting the Iverson bracket. This term goes to 1 if the contents are true or to 0 if the contents are false.

2.1.5 Support Vector Machines(SVM) Classifier

As a supervised learning classifier, Support Vector Machine (SVMs, also vector networks) are majorly used for classification problems, but they may also be applied to regression problems. In a multidimensional space, the SVM aims to classify the training data points by constructing an appropriate hyperplane or a set of hyperplanes. Hyperplane separates the data space or sets a decision boundary to categorize data points with the maximum margin among the classes shown on figure 9. The SVM goal is to construct the optimal margin separating the hyperplane based on the distance between the two decision boundaries. The support vectors are the data points placed closest to the SVM margin (Tigga & Garg [9]). Scikit-learn uses three classes of SVC, NuSVC, and LinearSVC to perform binary classification on datasets. Although SVC and NuSVC methods work similarly, they accept various attributes and are constructed using different mathematical formulas. In contrast to SVC and NuSVC, LinearSVCwhich is used for linear kernelshas a faster implementation.

2.2 Model Performance Measures

In this next section of the project, we will look at and discuss different statistics to assess the performance of the different models. When evaluating a model's performance, there are many different approaches and methods to consider, and each might differ in performance depending on the problem. As the goal of this projectis to compare and find which methods of machine learning perform better, and a criterion on which methods used to compare them is required.

2.2.1 Confusion Matrix

Confusion matrix is a method to visualize the accuracy using a table. To easily explain the confusion matrix, we assume a classifier that classifies instances as positive and negative. We then have four fields to be calculated in the matrix: true positive, true negative, false positive, and false negative. In a binary classification problem, a confusion matrix is typically a 2 × 2 table with four cells representing different prediction outcomes:

Table 3: Standard Form of Confusion Matrix .

	Actually Positive(1)	**Actually Negative(0)**
Predictied Positive(1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative(0)	False Negatives (FNs)	True Negatives (TNs)

- True Positive shows the models correctly predicted positive instances as positive.
- True Negative shows the models correctly predicted negative instances as negative.
- False Positive shows the models incorrectly predicted negative instances as positive.
- False Negative shows the models incorrectly predicted positive instances as negative.

There are several advantages that the confusion matrix offers. Firstly, it is efficient to use and requires little computational power. Secondly, because of its simplicity, it is easy not just for professionals to understand but also for individuals with different backgrounds. Additionally, utilizing the confusion matrix along with accuracy, can help with the verification of the model's effectiveness in predicting default and non-defaulters. Which provides a comprehensive assessment of the models predictive performance across different classes.

2.2.2 Accuracy

The possibility of calculating other statistical measures occurs when using the values found from a confusion matrix. The first one we will look at is the accuracy. A confusion matrix provides a clear and structured representation of the model's prediction and can help in decision-making based on the specific requirements of the classification task. To find the accuracy of the correct percentage of predictions made by the model, we can use the following formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

In accuracy, a random classifier will get, on average, half of the classification correctly, which means that values above 0.5 indicate that the model has a higher accuracy as random guessing. A perfect prediction has an accuracy of 1.0.

2.2.3 Precision

The precision value is calculated when the positive predictive value of the diagnostic test or the positive predictions are actually positive. In other words, it is defined as the probability that an individual with a positive diagnostic test result will become ill(Torgo & Ribeiro [10] and Koenig & Fuchs[11]). It is measures with:

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

2.2.4 Recall

Recall estimates how many True Positives the model has captured . By the same logic, when False Negative has a huge cost, Recall is the model metric used to select the best model(Torgo & Ribeiro [10]). It is measured with:

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

2.2.5 F1-Score

The F1-score is a combined metric between precision and recall, making a balanced measure of a model's performance. The F1-score is equal to the harmonic mean of the two, providing an evaluation of the models ability to correctly predict positive instances while minimizing the false negatives and false positives (Lipton et al.

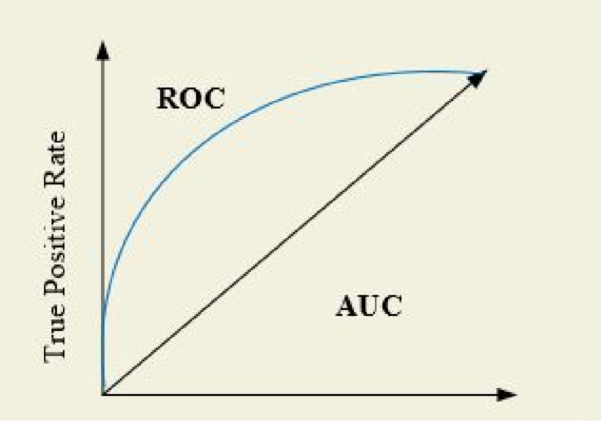


Figure 1: ROC-AUC curve.

[12]). Naturally, a negative with the F1-score is that it does not consider true negatives. A high f1-score indicates better overall model performance and is measured by:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{9}$$

2.2.6 Area Under the ROC Curve

The ROC (Receiver Operating Characteristic) curve is a commonly used tool to predict classifier performance. The ROC curve is plotted according to the True Positive Rate (TPR) to the False Positive Rate (FPR). The area under the ROC curve in performance evaluation is called AUC (Area Under Curve). ROC-AUC value will have been 1 for a perfect classifier. The ROC-AUC value approaching 1 indicates the successful separation of positives from negatives. The ROC-AUC curve is shown in Figure 1. The formulas for the True Positive Rate (TPR) and False Positive Rate (FPR) are

$$\text{TPR} = \frac{TP}{\text{Actual number of "1"}} \tag{10}$$

and

$$\text{FPR} = \frac{FP}{\text{Actual number of "0"}} \tag{11}$$

respectively.

3 Experimental Results

Diabetes is among the most prevalent chronic diseases in the United States, impacting millions of Americans each year and exerting a significant financial burden on the economy. Diabetes is a serious chronic disease in which individuals lose the ability to effectively regulate levels of glucose in the blood, and can lead to reduced quality of life and life expectancy. After different foods are broken down into sugars during digestion, the sugars are then released into the bloodstream. This signals the pancreas to release insulin. Insulin helps enable cells within the body to use those sugars in the bloodstream for energy. Diabetes is generally characterized by either the body not making enough insulin or being unable to use the insulin that is made as effectively as needed.

Complications like heart disease, vision loss, lower-limb amputation, and kidney disease are associated with chronically high levels of sugar remaining in the bloodstream for those with diabetes. While there is no cure for diabetes, strategies like losing weight, eating healthily, being active, and receiving medical treatments can mitigate the harms of this disease in many patients. Early diagnosis can lead to lifestyle changes and more effective treatment, making predictive models for diabetes risk important tools for public and public health officials.

In the study, the Diabetes Health Indicators Dataset, which is collected annually by the Centers for Disease Control and Prevention (CDC) in the form of a telephone questionnaire, was used. The survey collects responses from more than 400,000 individuals each year on health-related risk behaviors, chronic health conditions, and use of preventative services. This dataset has 21 feature variables. These characteristics are either questions asked directly to the participants or variables calculated based on individual participant responses. The dataset are obtained from the Kaggle website, which was available first from the UCI Machine Learning repository [13]. The features definitions of the variables are as in Table 4. The dataset includes 22 features and 253680 samples. Due to the large volume of data, for convenience, we select a sample of size n=3000 from the data set and perform the analysis on this sample.

Table 4: Features definitions.

Features	Label	Values
Diabetes_binary	Presence of diabetes	0. no diabetes, 1. diabetes
HighBP	High Blood Pressure	0. no high BP, 1. high BP
HighChol	High Cholesterol	0. no high cholesterol, 1. high cholesterol
CholCheck	Cholesterol Check	0. no cholesterol check in 5 years, 1. yes cholesterol check in 5 years
BMI	Body Mass Index	
Smoker	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]	0. no, 1. yes
Stroke	(Ever told) you had a stroke.	0. no, 1.yes
HeartDiseaseorAttack	Coronary heart disease (CHD) or myocardial infarction (MI)	0. no, 1. yes
PhysActivity	Physical activity in past 30 days - not including job	0. no, 1. yes
Fruits	Consume Fruit 1 or more times per day	0. no, 1. yes
Veggies	Consume Vegetables 1 or more times per day	0. no, 1. yes
HvyAlcoholConsum	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)	0. no, 1. yes
AnyHealthcare	Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc	0. no, 1. yes
NoDocbcCost	Was there a time in the past 12 months when you needed to see a doctor but could not because of cost?	0. no, 1. yes
GenHlth	Would you say that in general your health is: scale 1-5	1. excellent, 2. very good, 3. good, 4. fair, 5. poor
MentHlth	Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how	
PhysHlth	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30	
DiffWalk	Do you have serious difficulty walking or climbing stairs?	0. no, 1. yes
Sex		0. female, 1.male
Age	13-level age category	1. 18-24, 9. 60-64, 13. 80 or older
Education	Education level scale 1-6	1. Never attended school or only kindergarten, 2. Grades 1 through 8 (Elementary), 3. Grades 9 through 11 (Some high school), 4. Grade 12 or GED (High school graduate), 5. College 1 year to 3 years (Some college or technical school), 6. College 4 years or more (College graduate)

3.1 Data Visualization

Exploratory Data Analysis (EDA) is an approach for data analysis that often uses statistical graphics and other data visualization methods to maximize insight into a data set. Most EDA techniques are graphical in nature with a few quantitative techniques. The reason for the heavy reliance on graphics is that by its very nature the main role of EDA is to open-mindedly explore, and graphics gives the analysts unparalleled power to do so, enticing the data to reveal its structural secrets, and being always ready to gain some new, often unsuspected, insight into the data.

Count Plot

The Outcome column, as seen in Figure 2, can be considered as unbalanced because its two values are not nearly equivalent. We can also observe from the Figure 2 that the number of instances of patients with Diabetes disease is much less than those without Diabetes disease.

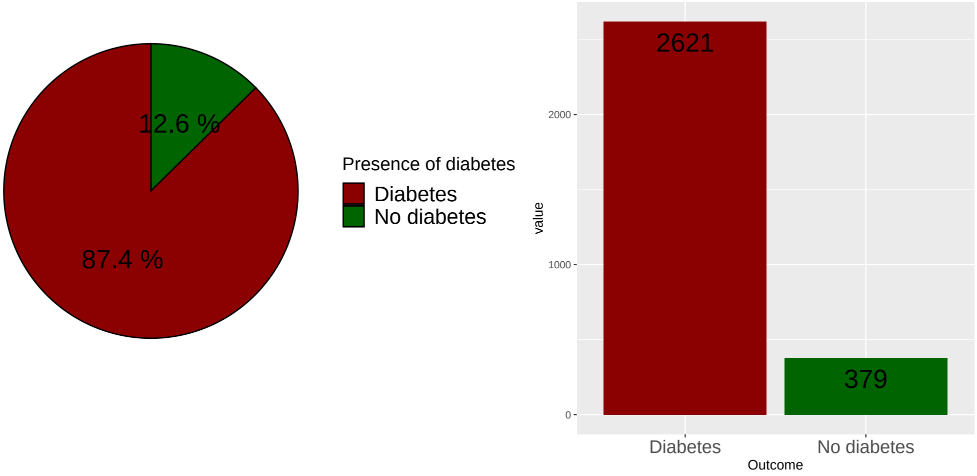


Figure 2: Plotting of Outcome variable.

Analyzing Relationships Between Variables

The relationship between independent variables is analyzed using the Correlation Matrix in one process. The results of this analysis show which independent variables might or might not affect your target variable. Through the Correlation Analysis, you may measure the correlation coefficient, which shows how much one variable changes when the other one does. You can get a linear relationship between two variables through Correlation Analysis. Higher positive values indicate a potential increase in the positive correlation, and higher negative values indicate a potential

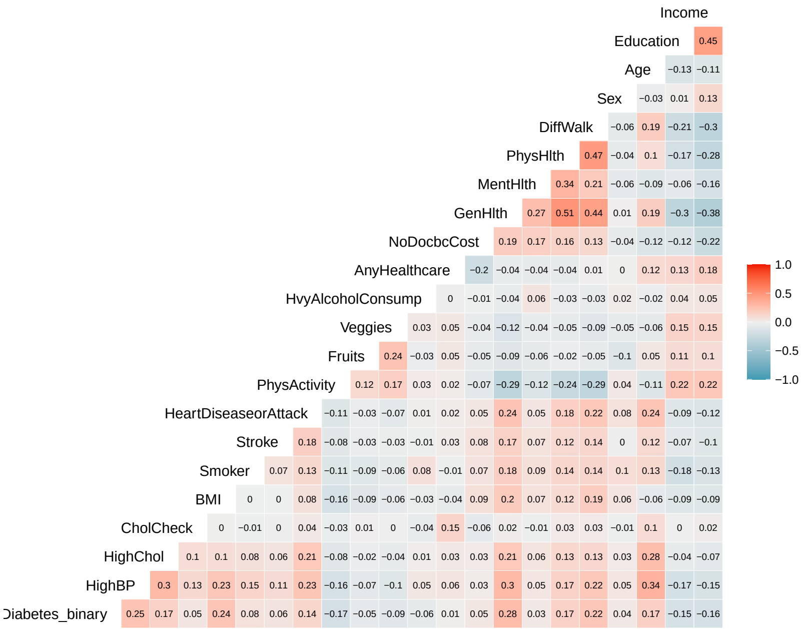


Figure 3: Correlation Analysis.

drop in the negative correlation (decrease). In this project, the range of positive correlation between independent variables and the target variable is highest for "GenHlth", "HighBP", "BMI", and "DiffWalk".

3.2 Dimensionality reduction

Dimensionality reduction is the process of reducing the number of variables considered (Domingos [14]). It can be used to extract latent features from raw datasets or to reduce the data while maintaining the structure. In this project three different dimensionality reduction methods have been used for feature selection: the Chi-Squared, the Pearson Correlation and the principal component analysis (PCA).

3.2.1 Chi-square Test for Feature Selection

The chi-square test is a statistical test used to determine if there is a significant association between two categorical variables. We calculate Chi-square between each feature and the target and select the desired number of features with best Chi-square scores. Features that show significant dependencies with the target variable are considered important for prediction and can be selected for further analysis. The statistic and p-value related to the chi-square test between each feature and the target variable are given in Table 5. For this project, we select the top 8 features using chi-square test for machine learning classification algorithms.

Table 5: The statistic and P-value based on the chi-square test for important features selection.

Id	**Features**	**Statistic**	**P-alue**
1	GenHlth	263.48938	0.00010
2	BMI	259.65160	0.00010
3	HighBP	183.90206	0.00010
4	DiffWalk	150.97974	0.00010
5	PhysHlth	109.93467	0.00010
6	Age	104.42274	0.00010
7	Income	100.17470	0.00010
8	PhysActivity	86.65190	0.00010
9	HighChol	82.78842	0.00010
10	Education	72.71065	0.00010
11	HeartDiseaseorAttack	58.25210	0.00010
12	MentHlth	28.17892	0.38236
13	Veggies	23.07176	0.00010
14	Smoker	17.59384	0.00010
15	Stroke	11.59583	0.00140
16	HvyAlcoholConsump	11.41797	0.00100
17	Fruits	8.92004	0.00270
18	CholCheck	8.37423	0.00490
19	NoDocbcCost	7.11436	0.01100
20	Sex	3.69710	0.05969
21	AnyHealthcare	0.30463	0.61504

3.2.2 Pearson Correlation for Feature Selection

Feature Selection is one of the prominent preprocessing steps in many of the machine learning applications. It is the process of reducing the feature set by choosing the relevant features from the original feature set according to an evaluation criterion and also removing the redundant features from the entire feature set. Correlation is a well-known similarity measures between two features. If two features are linearly dependent, then their correlation coefficient is ±1. If the features are uncorrelated, the correlation coefficient is 0. In this section, we use the feature selection algorithm based on correlation method for select the important features. The correlation between each feature and the target variable and the statistic and p-value related to the correlation test are given in Table 6. Therefore, we select the top 8 features using correlation test for machine learning classification algorithms.

Table 6: The correlation between each feature and the target variable and the statistic and p-value related to the correlation test for important features selection.

Id	Features	Correlation	|Statistic|	P-alue
1	GenHlth	0.284	16.19224	0.00000
2	HighBP	0.248	13.99217	0.00000
3	BMI	0.236	13.29997	0.00000
4	DiffWalk	0.224	12.60455	0.00000
5	Age	0.171	9.52747	0.00000
6	PhysActivity	−0.170	9.44297	0.00000
7	PhysHlth	0.167	9.25423	0.00000
8	HighChol	0.166	9.22394	0.00000
9	Income	−0.163	9.06128	0.00000
10	Education	−0.148	8.22106	0.00000
11	HeartDiseaseorAttack	0.139	7.70493	0.00000
12	Veggies	−0.088	4.82028	0.00000
13	Smoker	0.077	4.20545	0.00003
14	Stroke	0.062	3.41073	0.00066
15	HvyAlcoholConsump	−0.062	3.38437	0.00072
16	Fruits	−0.055	2.99010	0.00281
17	CholCheck	0.053	2.89691	0.00380
18	NoDocbcCost	0.049	2.66955	0.00764
19	Sex	0.035	1.92333	0.05453
20	MentHlth	0.033	1.79693	0.07245
21	AnyHealthcare	0.010	0.55178	0.58114

3.2.3 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is an unsupervised, non-parametric statistical technique primarily used for dimension reduction in machine learning. It considers the total variance in the data and transforms the original variables into a smaller set of linear combinations.In this section, we perform dimension reduction using Principal component analysis (PCA). To determine the number of meaningful components to be retained, we select the eigenvalue-one criterion for the analysis. With this, we kept all the components with an eigenvalue greater than 1.00. Therefore, components with an eigenvalue greater than 1.00 stood for a higher variance than their contribution as individual variables. In contrast, components with eigenvalues less than 1.00 contributed less than their individual value and were removed from analysis.

Table 7: Proportion of variance explained.

	Eigenvalue	**Variance percent(%)**	**Cumulative variance percent(%)**
Dim.1	89.621	47.992	47.992
Dim.2	42.013	22.498	70.490
Dim.3	38.713	20.731	91.221
Dim.4	8.991	4.815	96.035
Dim.5	4.115	2.204	98.239
Dim.6	0.823	0.441	98.680
Dim.7	0.651	0.348	99.028
Dim.8	0.291	0.156	99.184
Dim.9	0.234	0.125	99.309
Dim.10	0.224	0.120	99.429
Dim.11	0.209	0.112	99.541
Dim.12	0.166	0.089	99.630
Dim.13	0.159	0.085	99.715
Dim.14	0.126	0.068	99.783
Dim.15	0.092	0.049	99.832
Dim.16	0.077	0.041	99.873
Dim.17	0.073	0.039	99.912
Dim.18	0.051	0.027	99.940
Dim.19	0.044	0.024	99.963
Dim.20	0.036	0.020	99.983
Dim.21	0.032	0.017	100.000

From Table 7 and Figure 4, the first 5 components had a variance greater than 1.00 and an accumulated proportion of 0.9824. Thus, we choose 5 PCs. From Figure 5, **PhysHlth**, **MentHlth**, **BMI**, **Age** and **Income** are important variables.

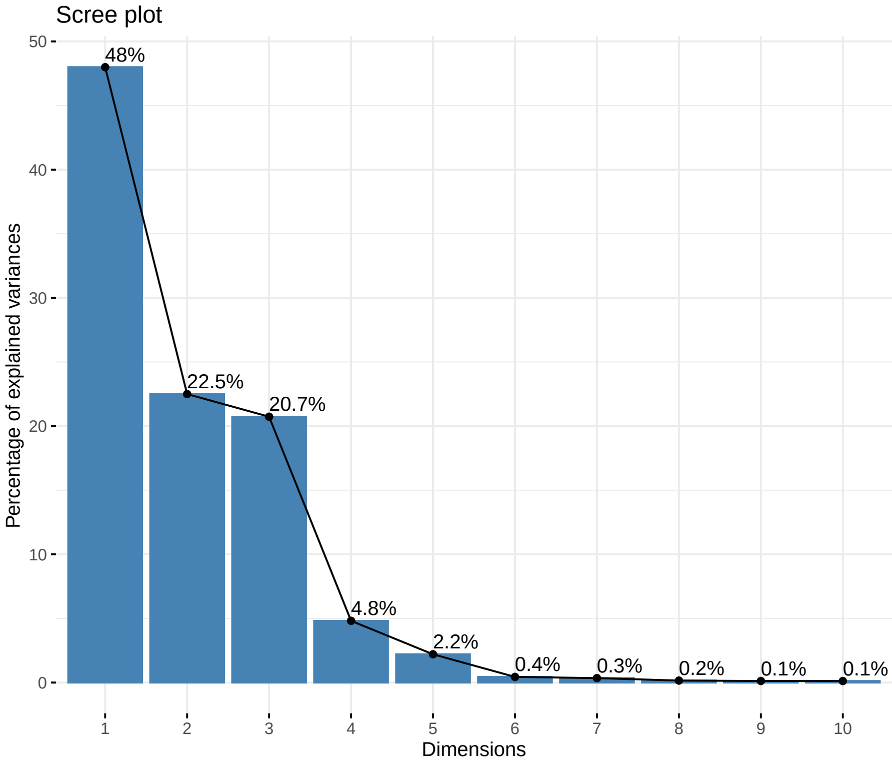


Figure 4: Scree plot.

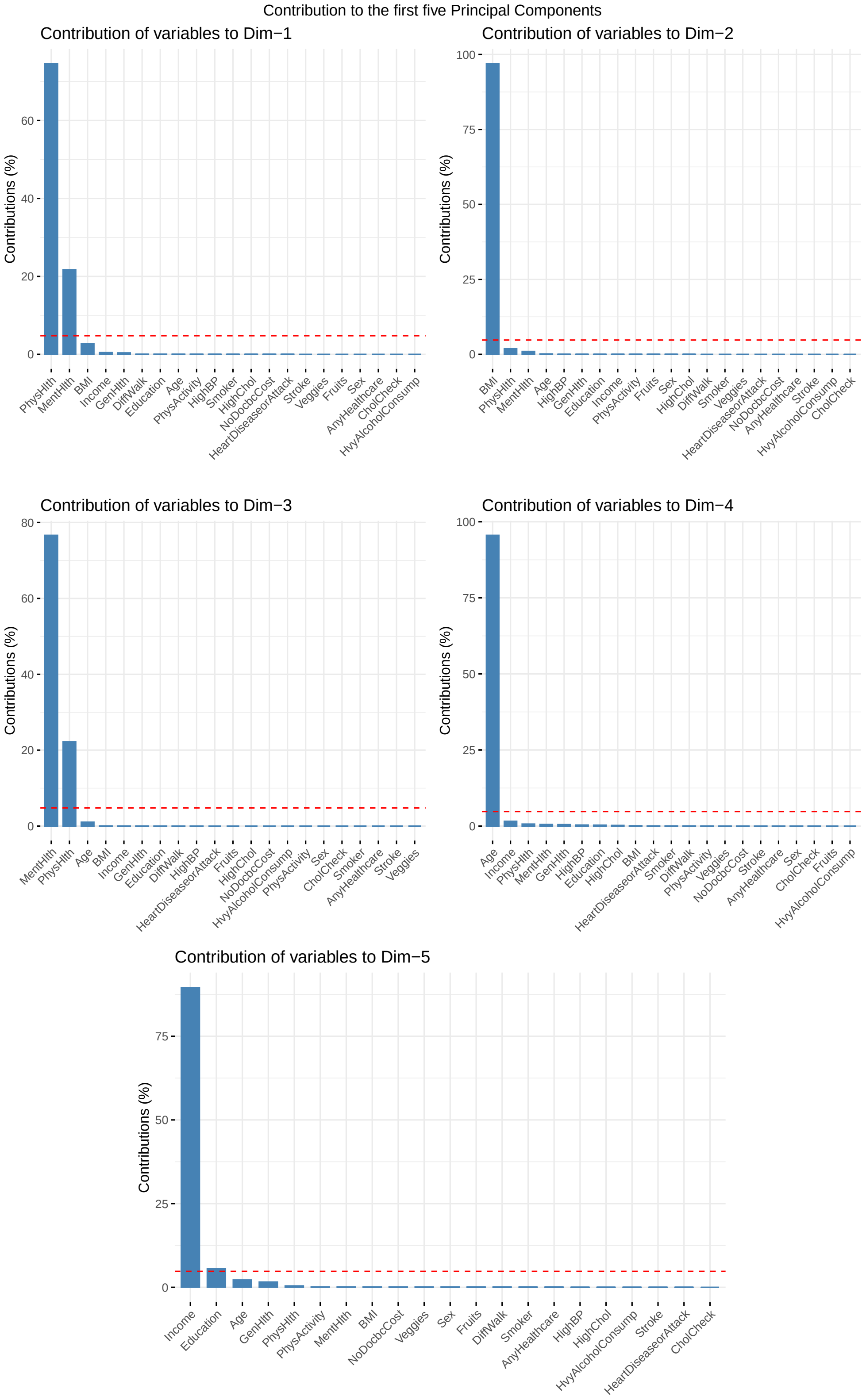


Figure 5: Contribution to the first five Principal Components.

3.3 Model Training and Testing

The next stage in the approach to reach the objectives of the experiment is the training and testing phase. The dataset that has been conducted from the previous stage is now used to train and test different algorithms. This phase is roughly split into two separate processes, which naturally are training and testing. Before either of the processes of training or testing can start, the dataset has to be split into training and testing sets. The reasoning for this cross-validating is to test the final algorithm settings on an unseen dataset. In this research experiment, a split of 20:80 is used, with 80% being the training set.

3.4 Results and Evaluation

Once the data is ready for modeling, the next step is to select an appropriate Machine Learning algorithm or model to explore patterns in your data. As we mentioned in section two, the classification algorithms used to predict diabetes include the following: Logistic regression, Random forest, Decision tree, K-Nearest Neighbor (KNN) and Support Vector Machines(SVM). This is the final step of prediction model. In addition, we evaluate the prediction results using classification algorithms based on various metrics like accuracy, precision, recall, f1-score and area under curve (AUC). Accuracy, precision, recall, f1-score and area under curve (AUC) values were calculated for each model and those are shown in Table 8.

Table 8: Comparative performance of ML models.

Feature Selection Method	**Models**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC**
Chi-square Test	**LR**	0.8633	0.9807	0.8756	0.9252	0.5566
	DT	0.8617	1.0000	0.8617	0.9257	0.5000
	RF	0.8617	0.9923	0.8666	0.9252	0.5202
	KNN	0.8533	0.9671	0.8757	0.9191	0.5558
	SVM	0.6767	0.7021	0.9007	0.7891	0.6101
Pearson Correlation	**LR**	0.8650	0.9826	0.8759	0.9262	0.5576
	DT	0.8617	1.0000	0.8617	0.9257	0.5000
	RF	0.8650	0.9923	0.8695	0.9268	0.5323
	KNN	0.8633	0.9768	0.8783	0.9249	0.5667
	SVM	0.5583	0.5455	0.9038	0.6803	0.5920
PCA	**LR**	0.8667	0.9923	0.8710	0.9277	0.5383
	DT	0.8617	1.0000	0.8617	0.9257	0.5000
	RF	0.8533	0.9652	0.8770	0.9190	0.5609
	KNN	0.8567	0.9671	0.8787	0.9208	0.5679
	SVM	0.6933	0.6905	0.9370	0.7951	0.7007

The confusion matrix in classification algorithms used to predict diabetes based on different dimensionality reduction methods , are given in Figures 9-11.

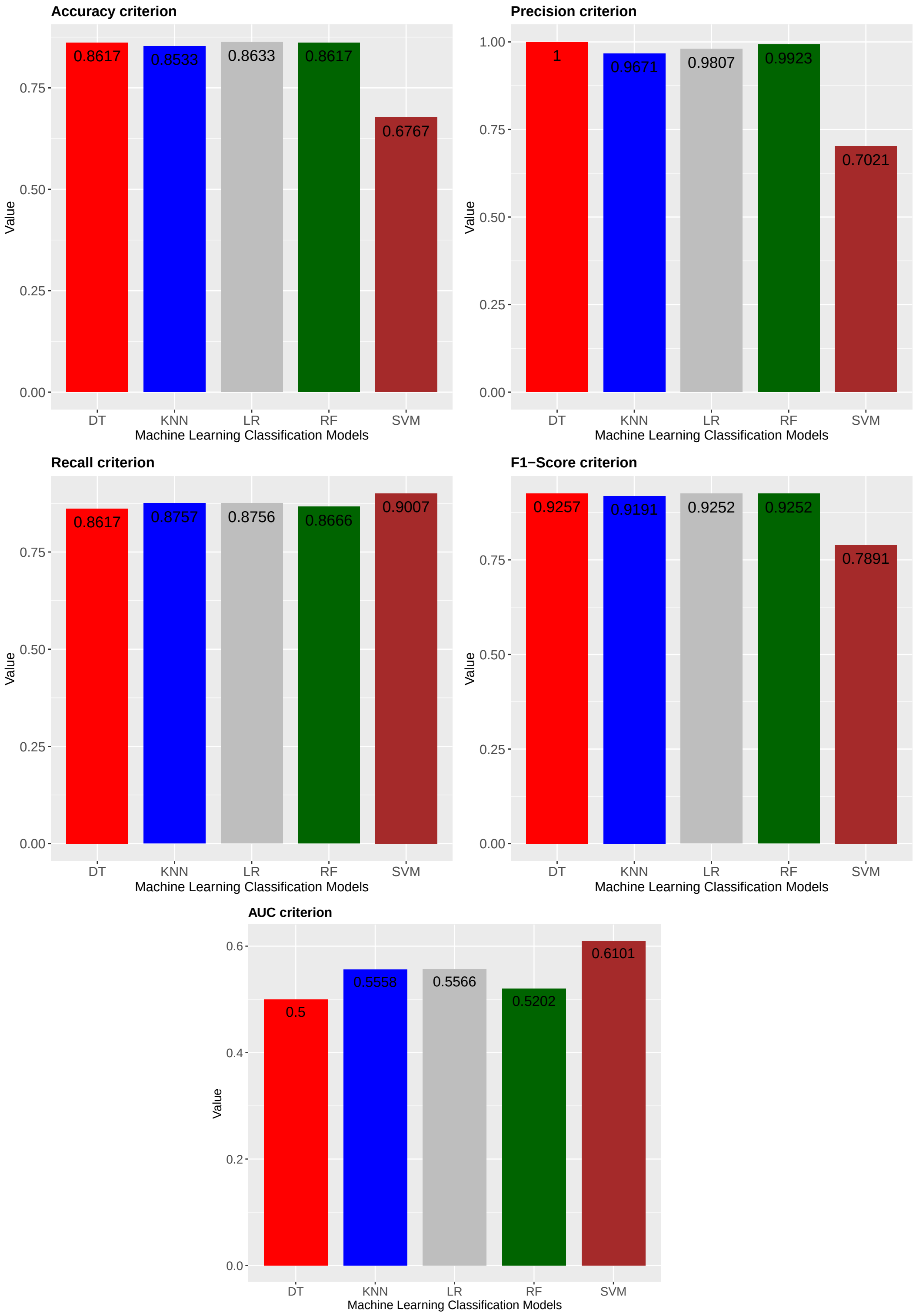


Figure 6: Comparison of models based on chi-square method.

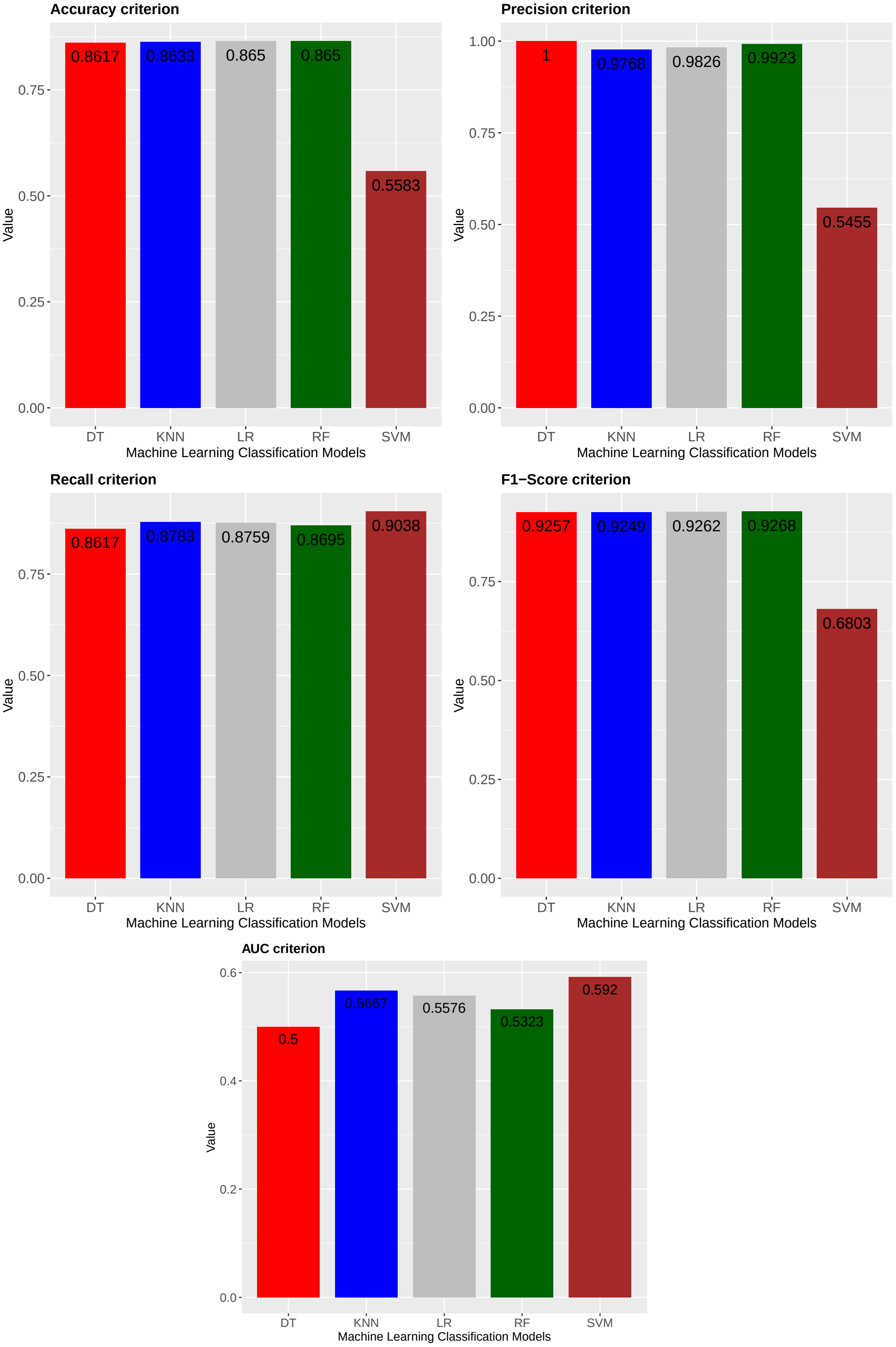
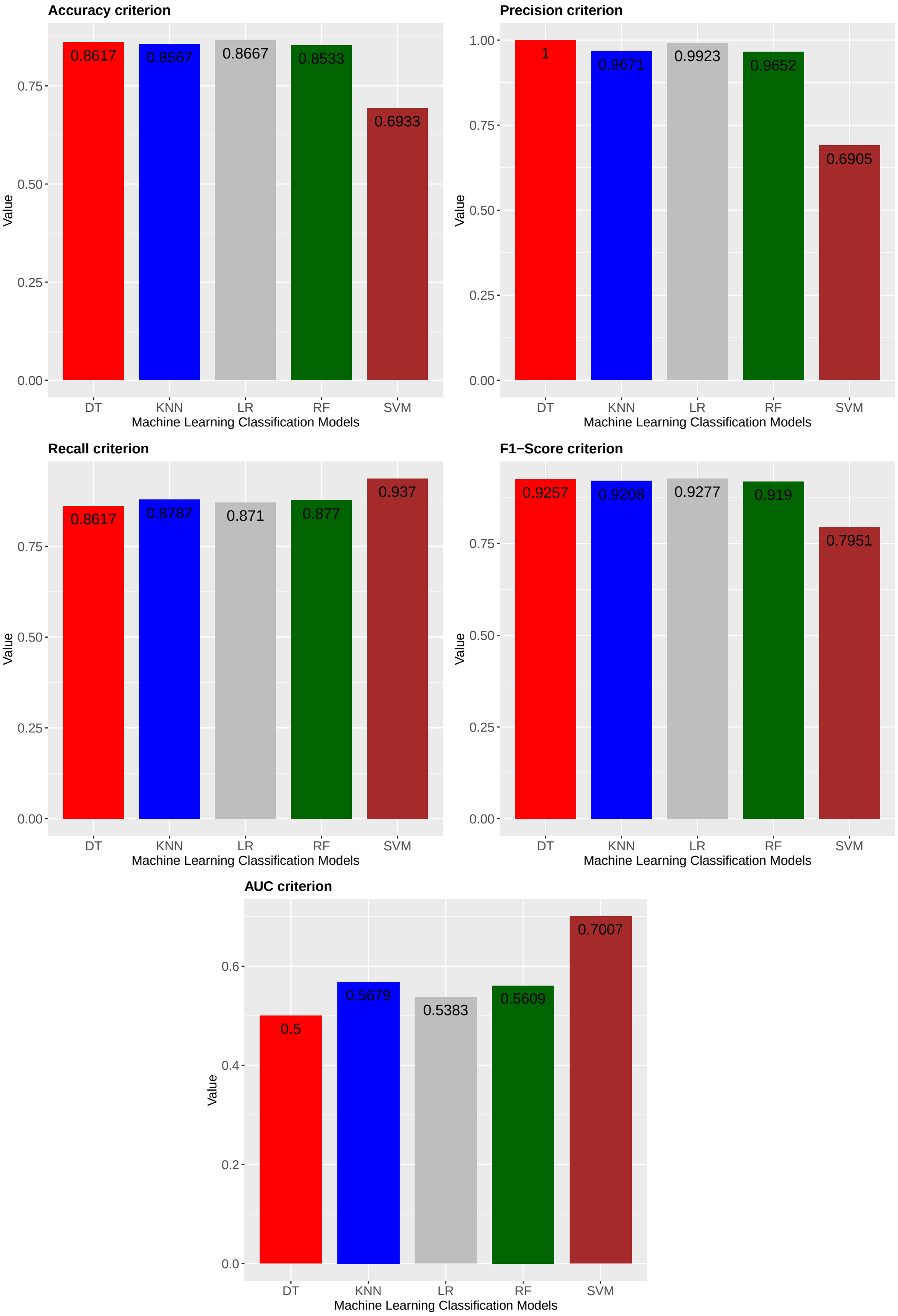


Figure 7: Comparison of models based on pearson correlation method.



Accuracy criterion

Precision criterion

Recall criterion

F1−Score criterion

AUC criterion

Figure 8: Comparison of models based on PCA method.

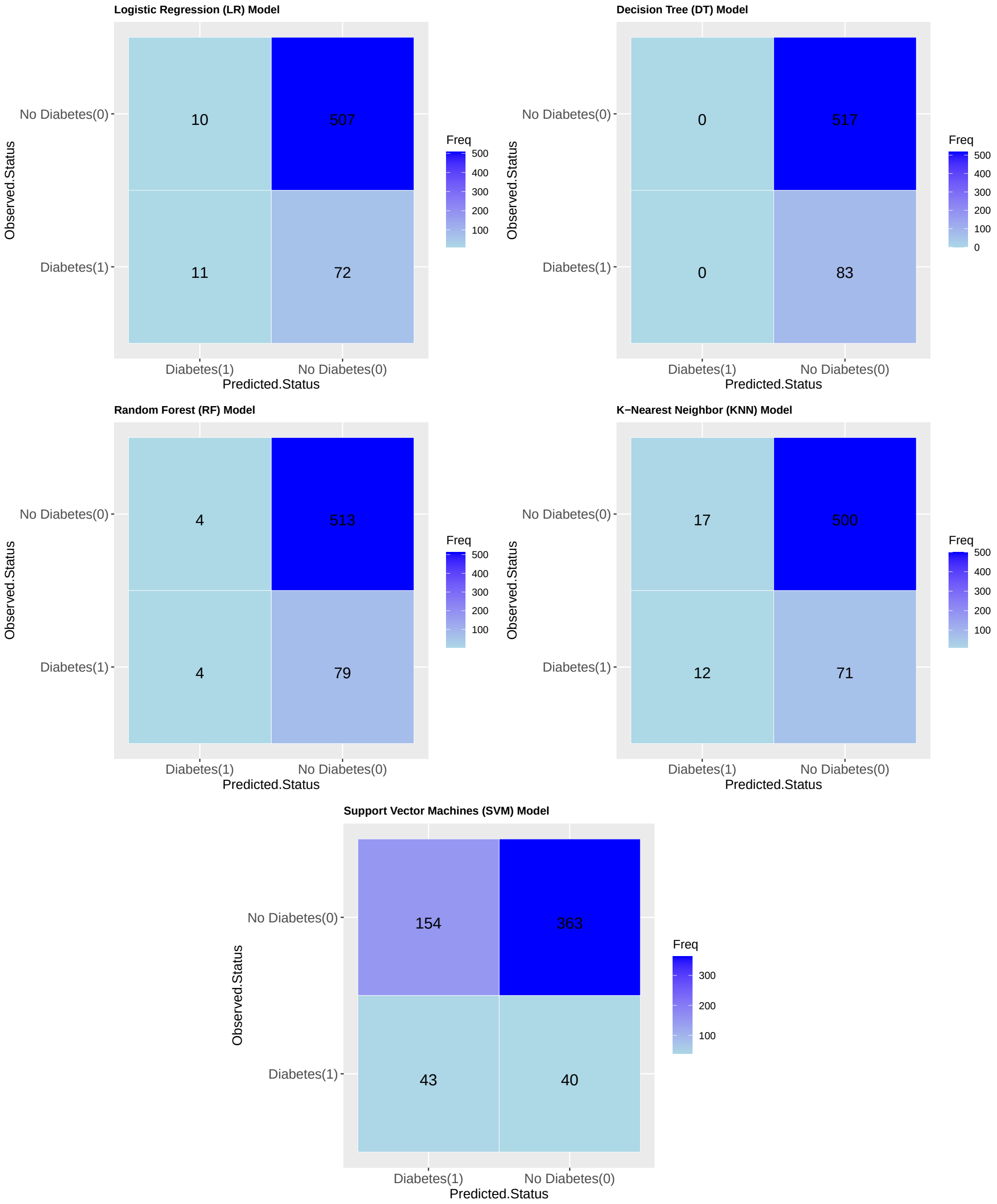


Figure 9: Confusion matrix of models based on chi-square method.

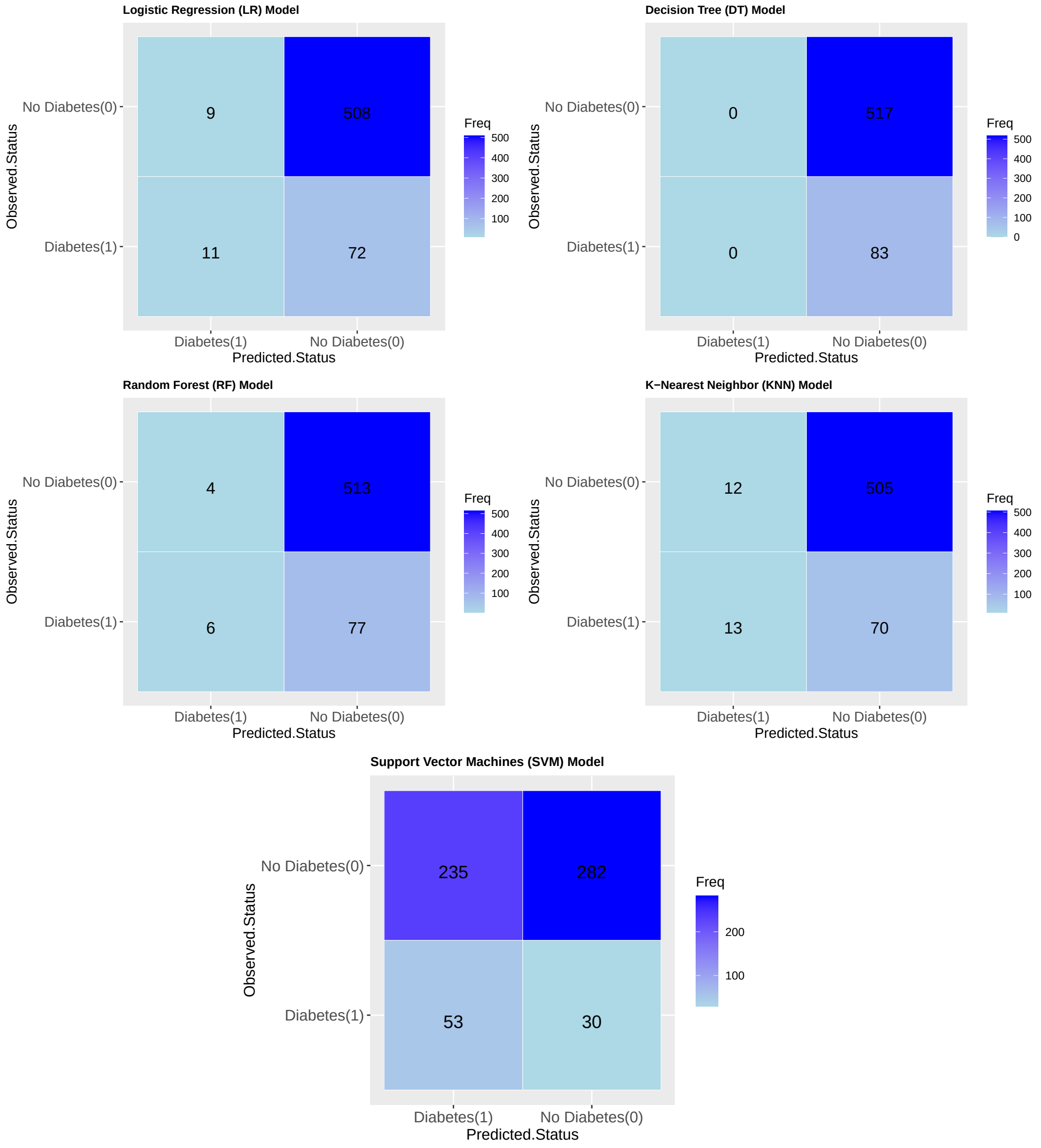


Figure 10: Confusion matrix of models based on pearson correlation method.

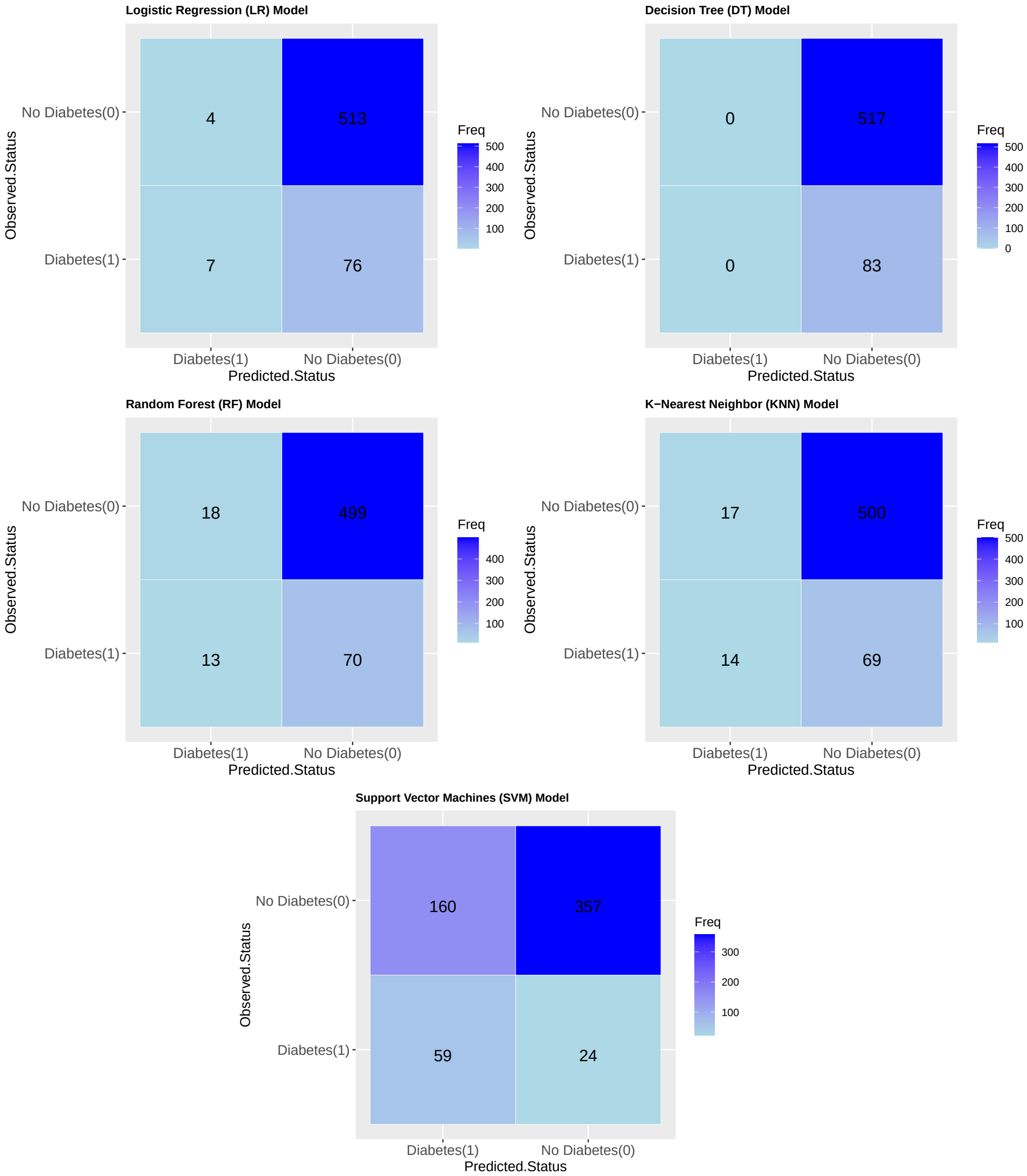


Figure 11: Confusion matrix of models based on PCA method.

4 Conclusion

Machine Learning (ML) methods have been used a lot lately for early diagnosis and planning in the field of health. ML models are very useful, especially in costly chronic diseases such as diabetes. While diabetes is a major public health problem, early detection is important. Individuals from different age groups are at risk of DM. In this project, we analyzed the prediction of Diabetes using Machine Learning algorithms on Diabetes Health Indicators Dataset, which is collected annually by the Centers for Disease Control and Prevention (CDC) to determine the most efficient classification model. At first, due to the large amount of data, we selected a sample of size n=3000 from data sets. In the following, we used the chi-squared, the pearson correlation and the principal component analysis (PCA) for important features selection. Then, we divided the data into training and testing sets. we investigated five Machine Learning (ML) algorithms: Logistic regression, Random forest, Decision tree, K-Nearest Neighbor (KNN) and Support Vector Machines (SVM) classification models. Accuracy, Precision, Recall, F1-Score and AUC metrics were used to compare the performance of the models. All algorithms have been successfully trained and yield high accuracy results and are very close to each other (except for SVM). we observed that the Logistic regression, Random forest, Decision tree, K-Nearest Neighbor (KNN) classification models gave the highest results in Accuracy, Precision and F1-score values than the Support Vector Machines (SVM) classification model. Moreover, the Support Vector Machines (SVM) classification model has highest results in Recall and AUC than the other classification models.