# Improvising Risk Prediction of Early Stage Diabetes

Om Anish Dalal
Munster Technological University, Ireland

### *Abstract*

*Diabetes is a condition when blood glucose or sugar is too high in a person. Early stage detection of diabetes helps in treatment and better health for the individual who would be at risk for diabetes. The objective of this study is to predict the likelihood of having early stage diabetes with respect to the symptoms found in the patient developing over time. The dataset contains the data of newly diabetic or would be diabetic patient with the symptoms identified. Several classification algorithms were applied on the dataset to predict whether the patient would be at risk for diabetes or not. Before applying the machine learning algorithms for the prediction, steps such as data pre-processing, label encoding, data transformation, and feature selection were performed. In this study, feature selection was done using three methods which plotted out the importance of the features in the prediction of diabetes: Pearson's correlation coefficient, SelectKBest, and Random Forest Classifier. 8 best features were selected out of 16 features from the dataset. Logistic regression, K-nearest neighbors classifier, Decision tree classifier, Support vector classifier, Random forest classifier and Naive Bayes algorithms were applied for the prediction of diabetes. From the comparison of accuracy scores of the these algorithms once with the baseline model including all the16 features and secondly with the modified models including only the 8 features from the feature selection process, Decision tree classifier, Support vector Classifier, and Random forest classifier predicted better than the other models. Hyperparameter tuning was applied on these three algorithms to improve the performance of the models using various parameters. Highest accuracy score of 97.12%, F1 score of 0.978417 and the roc value of 0.964182 was obtained from the Support vector classifier algorithm after the tuning of hyperparameters.*

## 1. Introduction

IDF Diabetes Atlas provided the data of diabetes for the year 2019 where approximately 463 million adults (20-79 years) were living with diabetes and by 2045 it is expected to increase to 700 million adults. 374 million people are at increased risk of developing type 2 diabetes [1]. Pre-diabetes, a stage before diagnosis of Type 2 diabetes, is a state where the blood sugar levels are higher than normal, but not high enough for a person to have a diagnosis of Type 2 diabetes. Early treatment of pre-diabetes can help in preventing the Type 2 diabetes [2]. The health is likely to be deteriorated if the person has undiagnosed and untreated diabetes for a longer time [3]. Predicting likelihood of the early stage diabetes in the patients can be helpful in the clinical environment for treating and early diagnosis of diabetes based on the symptoms occurred.

The dataset used in this study is publicly available in the UCI Machine Learning Repository website [4]. The data has been collected using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh and approved by a doctor. The dataset having 520 observations contains 16 features and one target variable named class having values Positive and Negative which indicates the likelihood whether the patient would be at risk for diabetes or not. Out of the 16 features, one feature Age is quantitative and other 15 features are Categorical with Yes/No values and Gender with Male/Female values. Patients from the age group between 16 to 90 years are present in this dataset. The description of the dataset is given in Table 1.

**Table 1:** Variables in the dataset

| Type | Feature | Value | Description |
|---|---|---|---|
| *Quantitative* | Age | Age in years | Age |
| *Categorical* | Gender | Male / Female | Gender |
| | Polyuria | Yes / No | A condition where the body urinates more than usual |
| | Polydipsia | Yes / No | Excessive thirst or excess drinking |
| | Sudden weight loss | Yes / No | Sudden Weight Loss |
| | Weakness | Yes / No | Weakness |
| | Polyphagia | Yes / No | Excessive or extreme hunger |
| | Genital Thrush | Yes / No | Yeast infection |
| | Visual blurring | Yes / No | Blurred vision |
| | Itching | Yes / No | Itching |
| | Irritability | Yes / No | Irritability |
| | Delayed healing | Yes / No | Delayed healing |
| | Partial Paresis | Yes / No | Weakening of a muscle or group of muscles |
| | Muscle stiffness | Yes / No | Muscle stiffness |
| | Alopecia | Yes / No | Hair loss |
| | Obesity | Yes / No | Obesity |
| *Label* | Class | Positive / Negative | Diabetes present or not |

Existing research on the prediction of early stage diabetes has been carried out using the classification machine learning algorithms. Islam et al (2020) analyzed the dataset with Logistic Regression Algorithm, Naive Bayes Algorithm, Random Forest Algorithm. After applying ten-fold cross validation, Random Forest model had shown the best accuracy on this dataset [5]. Another work on predicting diabetes was done by Mirzajani et al (2018) but on a different dataset with some other features included, Neural network, Basin network, C5.0, and support vector machine models were compared for predicting diabetes and C5.0 model showed the highest accuracy. [6]

This study aimed to predict the early stage diabetes from the set of 16 features using various machine learning algorithms: Logistic regression, K-nearest neighbors classifier, Decision tree classifier, Support vector classifier, Random forest classifier and Naive Bayes algorithms. Following the pre-processing of the data, label encoding and standardization of features were carried out on the dataset. Feature selection being an important part while building a machine learning model was carried out using three techniques: Pearson’s correlation coefficient, SelectKBest, and Random Forest Classifier. Best features were extracted from the feature selection comparing all the three techniques. Comparing the scores with the baseline model with all the features and improved model with the best selected features, three algorithms (Decision tree classifier, Support vector Classifier, and Random forest classifier) were identified with higher accuracy and f1 scores and were chosen to do hyperparameter tuning to further improve the model performance. Performance metrics such as accuracy score, AUC value (Area under the ROC Curve), precision score, recall score, f1 score, confusion matrix and plotting the ROC curve (Receiver Operating Characteristic) were used for analyzing the model performance. Support vector classifier algorithm showed the highest performance after the tuning of hyperparameters.

The remainder of this paper contains following sections. Section 2 outlines the various techniques that were implemented for feature selection process. Section 3 describe the methodology which includes the data pre-processing techniques used, how the data was transformed, model selection and validation with hyperparameter tuning. Section 4 contain a comprehensive evaluation of the results obtained at every stage. In Section 5, we conclude and point out some possible future works.

## 2. Research

In this study, three different methods have been carried out for selecting the important features which can be used to improve the base model performance. Number of features play an important role in the model performance. Feature selection is identifying the features that contributes more to the target variable. Some benefits of feature selection include reduction of overfitting, improvement in accuracy and reduction in training time of the model. The following three methods were incorporated in this study for feature selection. Results of the three methods are discussed in the Evaluation section.

- Pearson Correlation Coefficient
- Univariate Selection: SelectKBest method using chi-squared statistical test
- Random Forest Classifier: Feature Importance

### *a. Pearson Correlation Coefficient*

Correlation coefficients are used to find the relationship among the features and the target variable. It gives the gist of how well the data fits a line. The strength of the linear relationship can be known between the variables along with the target variable. Pearson Correlation coefficient is a type of correlation coefficient denoted by r which takes the values from -1 to +1. One of the assumptions in calculating Pearson Correlation coefficient is that all the variables are treated the same. There is no differentiation between the independent and dependent variables. A value of 0 indicates there is no relationship among the variables. A value of +1 indicates a strong linear relationship while a value of -1 indicates a strong negative relationship among the variables. Strong, moderate or weak correlations can be determined on the range of values of r. If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation. It is said to be a medium correlation when the value lies between ± 0.30 and ± 0.49 and weak correlation when the value lies below ±0.29. The Pearson Product-Moment Correlation equation is used to calculate the correlation coefficients.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$r$ = correlation coefficient

$x_i$ = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

$y_i$ = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

### *b. Univariate Selection: SelectKBest method using chi-squared statistical test*

Univariate feature selection selects the best features based on the statistical tests. SelectKBest class scores the features using a statistical score function and select features according to the k highest scores. Various score functions such as Chi-square test, F-test and mutual_info_classif test can be used with the SelectKBest method. In this study, we have used Chi-Square test. Chi-Square test is a statistical test to determine the independence of two variables. Chi-Square is used for the categorical features. Chi-square is calculated between each feature and the target variable and dependence is obtained based on the values. If dependence is shown, then the feature variable is important, otherwise the feature can be discarded. The Chi-square formula is used to calculate the scores.

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

**where:**

$c =$ degrees of freedom

$O =$ observed value(s)

$E =$ expected value(s)

*c. Random Forest Classifier: Feature Importance*

Feature importance gives the relevant features and top features can be selected to improve the model performance. The higher the score more important or relevant is the feature with the target variable. Various Decision tree classifiers provide the *feature_importance_* property which stores the relative scores of each feature after fitting the model. Random Forest Classifier, Extra Tree Classifier and XGBoost Classifier are some of the decision tree algorithms which gives the *feature_importance_* property after fitting the model. In this study, Random Forest Classifier was implemented to get the scores of the features. Random forest iterates each feature and shuffles it in the dataset to make prediction. If a feature is important in making predictions, shuffling that feature will lead to an increase in error. The features that lead to maximum increase in error are found to be the most important.

## 3. Methodology

The dataset was pre-processed to find the association among the features and the target variable and evaluate some of the machine learning algorithms. The study consisted of following phases:

*a. Dealing with missing values and Outliers*

The dataset containing 520 observations and 17 variables was checked for the missing values. If missing values are present, the machine learning algorithm cannot process. Missing values must be imputed, or the observation must be removed having the missing value.

Outliers were checked for the respective variables as a part of pre-processing. Out of the 16 features, one feature Age is quantitative and other 15 features are Categorical. Outliers were checked for Age feature plotting the box plot.

*b. Encoding Data*

Categorical features in the dataset needs to be encoded for many of the machine learning algorithms to perform. Values of the categorical features are converted to integers. Label Encoding, Ordinal Encoding and One Hot Encoding are some of the techniques which encodes the categorical data. Label Encoding was used in this study to convert the categorical features. Label Encoding replaces the value with a number between 0 and the number of distinct classes of the feature minus 1. Label Encoding was considered since the number of features is quite large and one-hot encoding can lead to high memory consumption.

*c. Scaling Data*

Scaling of the data is important for a machine learning algorithm to behave better with the features being on the same scale. Normalization and Standardization are the data transformation techniques. Normalization of data would mean changing the values into the range between 0 and 1. Standardization scales each feature’s distribution to have a mean of 0 and standard deviation of 1. Standardization was performed on the dataset in this study. Standard Scaler, Robust Scaler and Min Max Scaler are some of the techniques available to standardize the data. Scaling of the 16 features was done using Standard Scaler technique.

Standardization:

$$z = \frac{x-\mu}{\sigma}$$

with mean:

$$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i)$$

and standard deviation:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}$$

*d. Data Splitting*

The dataset was split into training set and test set into the ratio of 80:20. The training dataset will contain 80% of the observations and the test set will contain 20% of the observations from the dataset.

*e. Feature Selection*

Feature selection is identifying the features that contributes more to the target variable. Three methods were incorporated in this study for feature selection: Pearson Correlation Coefficient, Univariate Selection: SelectKBest method using chi-squared statistical test and Random Forest Classifier: Feature Importance. Top features were identified from comparing all the techniques and 8 best features were selected from the 16 features to improve the base model evaluated with all the features.

*d. Model building*

The training set was used for training the models and tuning the parameters and the test set was used to evaluate the performance of the model. Six machine learning algorithms were implemented in this study to find the model with highest accuracy. Ten-fold cross validation was also performed on the training set where the data is segmented into parts and use all but one part for training and the remaining one for testing.

*Logistic Regression:* Logistic regression is a predictive analysis to conduct when the target feature is categorical. It models the probability with two possible outcomes for classification target feature.

*K-Nearest Neighbors:* This algorithm determines the similarity between the new observation and existing ones and groups the new observation into the category that is most similar to the available categories. It calculates the Euclidian distance between the existing data and the new data and assigns the new data to that category where the number of the neighbor is maximum.

*Decision Tree:* The decision tree algorithm uses the tree representation to classify the label. Each internal node of the tree corresponds to a feature, and each leaf node corresponds to a target label. It is used to predict the target by simple decision rules learned from the training data.

*Support Vector Machine:* Support Vector Classifier are classes that can perform both binary and multi-class classification. The objective is of this algorithm is find the hyperplane in an N-dimensional space where N are the number of features that classifies the data.

*Random Forest:* Random Forest is a tree-based algorithm which creates forest of many trees. It takes the ensemble approach and creates multiple decision trees from a random sample of the training data. Random forest is prone to overfit the data and hyperparameter tuning can help to prevent that.

*Naïve Bayes:* Naïve Bayes algorithms are based on Bayes theorem. This algorithm assumes that the features are conditionally independent. Gaussian Naive Bayes is a variant of Naive Bayes that follows Gaussian normal distribution and supports continuous data.

*e. Hyperparameter Tuning*

Hyperparameter Tuning is finding the right combination of their values from the set of parameters while building a model which can help to find either the loss or the accuracy of a function. Out of the six models implemented, three models were selected from the comparison between the accuracy score of the base models with all the features and improvised models with only the selected best features. Hyperparameter Tuning was performed on the Decision tree classifier, Support vector Classifier, and Random forest classifier algorithms to further improve the performance of the model.

GridSearchCV class was used to perform detailed search over the specified parameter values for an estimator and find the best set of parameter values. This function loop through predefined hyperparameters and fit the model on the training set. Various hyperparameters with the range of values that were examined for the three models:

*Random Forest:* Function to measure the quality of a split (criterion = [gini ,entropy]; The maximum depth of the tree (max_depth = [4, 6]); The minimum number of samples required to be at a leaf node (min_samples_leaf = [3, 4, 5]); The minimum number of samples required to split an internal node (min_samples_split = [8, 10, 12]); Number of trees in the forest (n_estimators = [10, 50, 100, 200])

*Support Vector Machine:* Regularization parameter (C = [0.1, 1, 10, 100, 1000]); Kernel coefficient (gamma = [1, 0.1, 0.01, 0.001, 0.0001]); Kernel type to be used (kernel = ['rbf','linear', 'poly'])

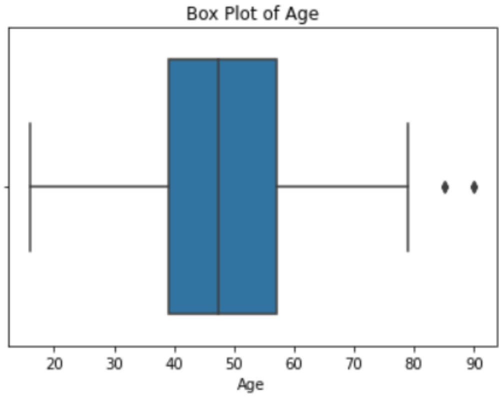
*Decision Tree:* Function to measure the quality of a split (criterion = [gini ,entropy]; Maximum depth of the tree (max_depth = [1,2,3,4,5,6,7,10,12]); Number of features to consider when looking for the best split (max_features = [5]); Minimum number of samples required to be at a leaf node.(min_samples_leaf = [1,2,3,4,5,6,7,8,9,10,11])

## 4. Evaluation

*a. Data Pre-processing and Exploratory Data Analysis*

There were no missing values found in this dataset of 520 observations. Out of the 16 features, one feature Age is quantitative and other 15 features are Categorical with Yes/No values and Gender with Male/Female values. Age feature was checked for any outliers. Figure 1 shows the box plot of Age feature where for 4 of the observations the patients were within the range of 80-90 years and were found to be the outliers. Outliers have not been imputed or dealt with in this study. From the data, it was found that patients above the age of 75 were more likely to be diagnosed with diabetes.

**Figure 1:** Box Plot of Age



In the dataset, there are 320 observations where the label which indicates the likelihood of the risk of Diabetes is Positive and 200 observations the risk of Diabetes is Negative. Females in the dataset turned out to be more positive than males (Figure 2). Diabetic positive patients are less likely to experience Alopecia than negative patients. Diabetes positive patients are likely to experience Visual Blurring, Polyphagia, Muscle Stiffness, and partial Paresis. Most of the negative patients showing delayed healing symptoms negative are above 45 years while those that showing no symptoms are below 50 years. Most of the patients with diabetes had Polyuria. Only few observations were positive with respect to Obesity (Figure 3).

**Figure 2:** Gender in relation to Diabetes

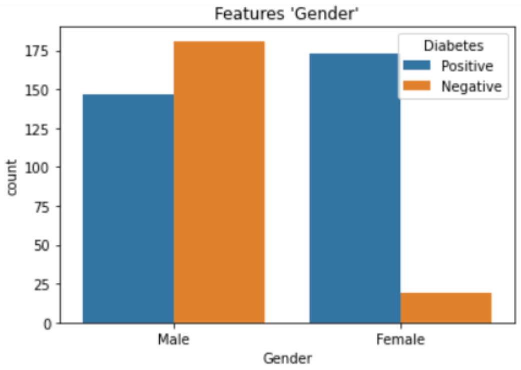
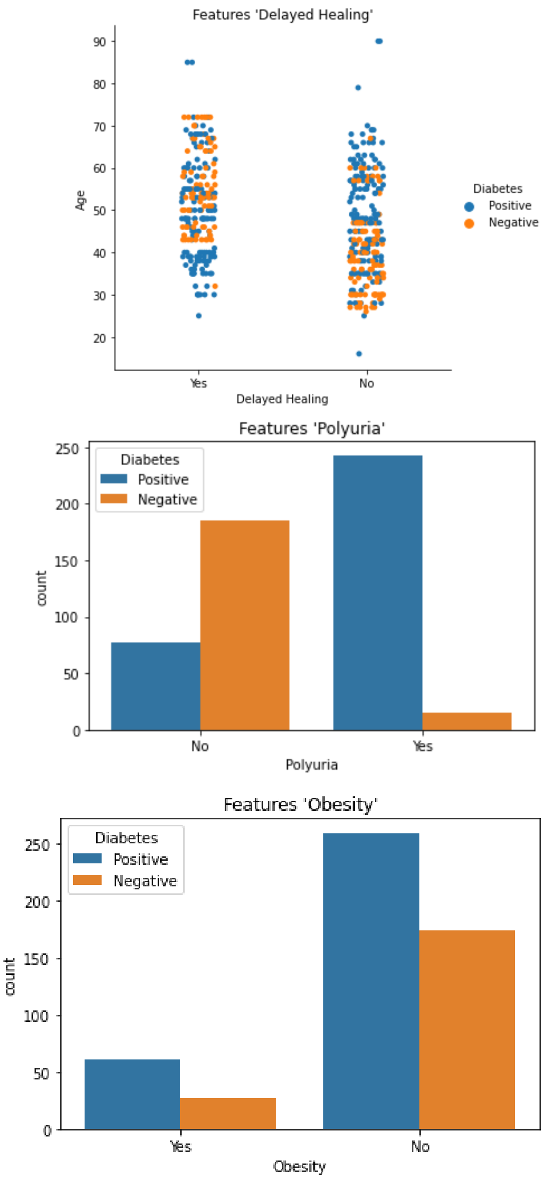


**Figure 3:** Features with respect to Diabetes



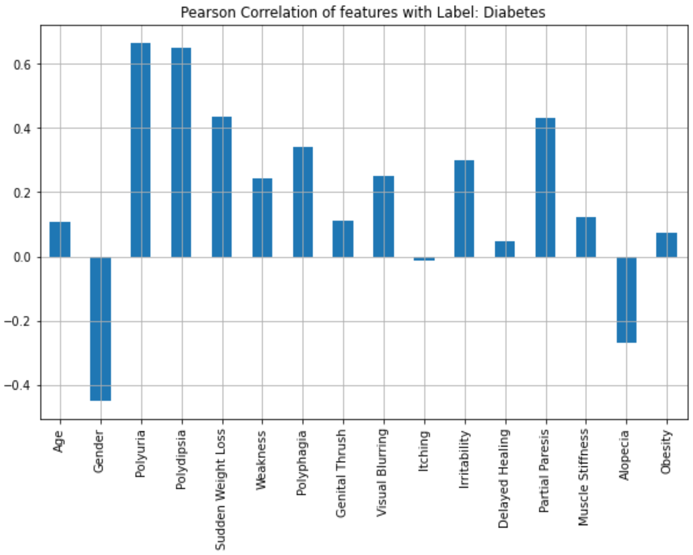
Every feature apart from Age contains categorical data. Label Encoder was used to convert these binary data to numeric data for the machine learning algorithm to evaluate. Age feature is of the different scale than the rest of the data as it is a quantitative variable. Standardization was performed using Standard Scalar on all the features for the machine learning algorithms to perform better. After splitting the data into training set and test set, 416 observations were in the training set and 104 observations were in the test set.

*b. Feature Selection*

The dataset was analyzed to check for the relations between the features and the target variable Diabetes. Three methods were incorporated in this study for feature selection: Pearson Correlation Coefficient, Univariate Selection: SelectKBest method using chi-squared statistical test and Random Forest Classifier: Feature Importance. Top features were identified from comparing all the techniques and 8 best features were selected from the 16 features to improve the base model evaluated with all the features.

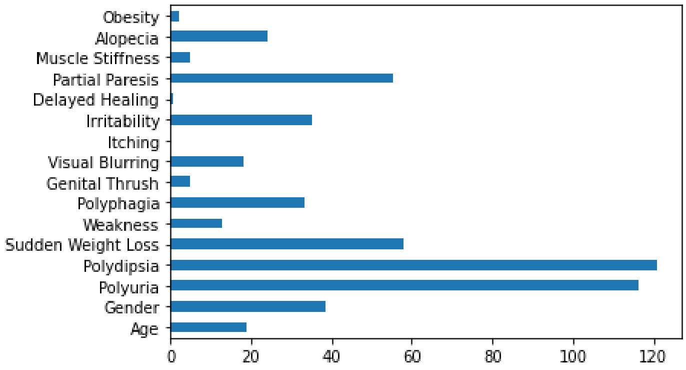
From the Pearson Correlation Coefficient scores, Polyuria (0.665922) and Polydipsia (0.648734) were found out to be strongly positively correlated with the target variable Diabetes and Gender (-0.449233) has the strongest negative correlation. Polyuria and Polydipsia symptoms strongly determine whether a patient has Early Stage Diabetes or not. The top 10 features selected using this technique are Polyuria, Polydipsia, Gender, Sudden Weight Loss, Partial Paresis, Polyphagia, Irritability, Alopecia, Visual Blurring and Weakness. Figure 4 shows the score of the features with respect to the target Diabetes.

**Figure 4:** Pearson Correlation Coefficient scores



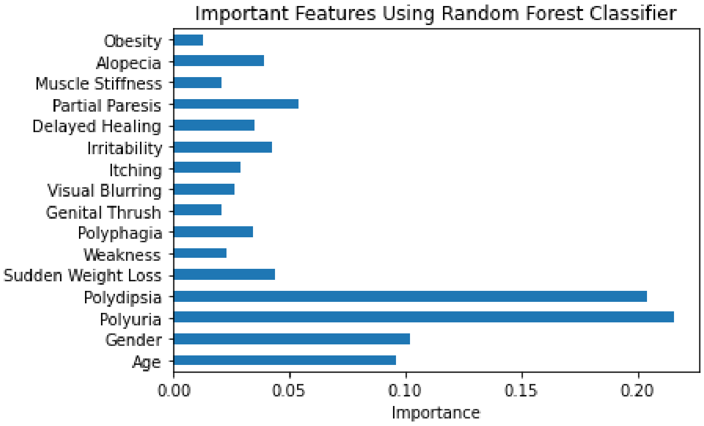
Top 10 features were identified using the SelectKBest method using chi-squared statistical test. Polyuria, Polydipsia, Gender, Sudden Weight Loss, Partial Paresis, Polyphagia, Irritability, Alopecia, Visual Blurring and Age were identified with the top scores. Age was identified as one of top 10 feature which was not the case in Pearson Correlation Coefficient. Polydipsia (120.785515) and Polyuria (116.184593) and) were found out to be the top features. Figure 5 shows the feature importance scores using the SelectKBest method using chi-squared statistical test.

**Figure 5:** SelectKBest method scores



Important features that were identified using the Random Forest Classifier: Feature Importance method were Polyuria, Polydipsia, Gender, Sudden Weight Loss, Partial Paresis, Polyphagia, Irritability, Alopecia, Delayed Healing and Age. Delayed Healing was not in the top 10 features for the other two methods. Polyuria, Polydipsia, Sudden Weight Loss, and Partial Paresis were the top 4 features identified with the Random Forest Classifier. Figure 6 shows the feature importance scores using the Random Forest Classifier: Feature Importance.

**Figure 6:** Random Forest Classifier: Feature Importance



From comparing the top features from all the three methods, 8 features were chosen out of 16 to check and improve the performance of the base models. Polyuria, Polydipsia, Gender, Sudden Weight Loss, Partial Paresis, Polyphagia, Irritability, and Age were selected in the feature selection process.

*c. Model performance*

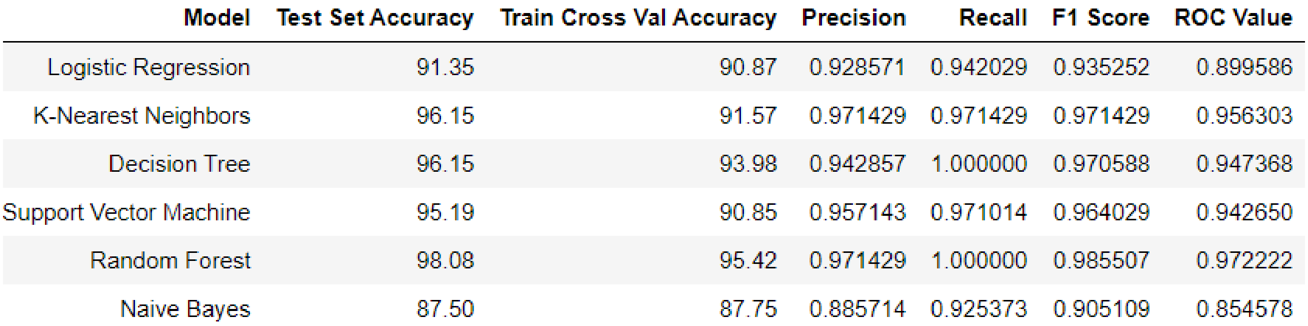
Performance metrics such as accuracy score, AUC value (Area under the ROC Curve), precision score, recall score, f1 score, confusion matrix and plotting the ROC curve (Receiver Operating Characteristic) were used for analyzing the model performance.

*Accuracy Score:* Accuracy score is the fraction of predictions the model correctly identified the target.
*Precision Score:* This score refers to the ability of the classifier not to label as positive when it is negative.
*Recall:* This score refers to the ability of the classifier to find all the positive samples.
*F1 Score:* F1 score is the harmonic mean of precision and recall. The classifier will only get a high F1 score if both recall and precision are high.
*Receiver operating characteristic (ROC) curve*: The ROC curve plots the true positive rate against the false positive rate.
*Confusion Matrix:* The first row and column of the matrix denotes the True Positives and the second row, and the second column denotes the True Negatives.

Logistic regression, K-nearest neighbors classifier, Decision tree classifier, Support vector classifier, Random forest classifier and Naive Bayes algorithms were applied for the prediction of diabetes. These models were first trained with the default parameters and baseline performance metrics were obtained. All the features were included while building the model. Ten-fold cross validation was also performed on the training set where the data is segmented into parts and use all but one part for training and the remaining one for testing. Table 2 shows the baseline scores for the models. Random Forest, Support Vector Machine, and Decision Tree models showed higher accuracy scores among other models with the accuracy of 97.12%, 95.19% and 97.12% respectively.

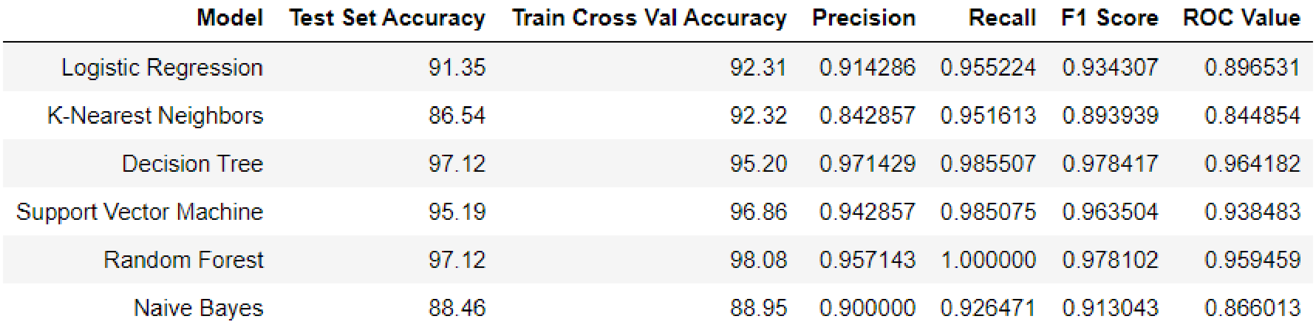
From the feature selection, 8 features were identified which showed higher importance with respect to the target variable. With only the 8 features, these models were trained with the default parameters. After using the important features, K-Nearest Neighbors showed the increase in the accuracy by 9.61%. Accuracy of Random Forest Classifier also increased from 97.12% to 98.08%. Among the 6 models, Random Forest, Support Vector Machine, and Decision Tree models showed higher accuracy scores among other models with the accuracy of 98.08%, 95.19% and 96.15% respectively. Table 3 shows the accuracy score of the baseline models with default parameters and all the features and improvised models with default parameters and only the best selected features.

**Table 2:** Baseline Model Performance Metrics – All Features



| Model | Test Set Accuracy | Train Cross Val Accuracy | Precision | Recall | F1 Score | ROC Value |
|---|---|---|---|---|---|---|
| Logistic Regression | 91.35 | 90.87 | 0.928571 | 0.942029 | 0.935252 | 0.899586 |
| K-Nearest Neighbors | 96.15 | 91.57 | 0.971429 | 0.971429 | 0.971429 | 0.956303 |
| Decision Tree | 96.15 | 93.98 | 0.942857 | 1.000000 | 0.970588 | 0.947368 |
| Support Vector Machine | 95.19 | 90.85 | 0.957143 | 0.971014 | 0.964029 | 0.942650 |
| Random Forest | 98.08 | 95.42 | 0.971429 | 1.000000 | 0.985507 | 0.972222 |
| Naive Bayes | 87.50 | 87.75 | 0.885714 | 0.925373 | 0.905109 | 0.854578 |

**Table 3:** Improvised Model Performance Metrics – Best Features



| Model | Test Set Accuracy | Train Cross Val Accuracy | Precision | Recall | F1 Score | ROC Value |
|---|---|---|---|---|---|---|
| Logistic Regression | 91.35 | 92.31 | 0.914286 | 0.955224 | 0.934307 | 0.896531 |
| K-Nearest Neighbors | 86.54 | 92.32 | 0.842857 | 0.951613 | 0.893939 | 0.844854 |
| Decision Tree | 97.12 | 95.20 | 0.971429 | 0.985507 | 0.978417 | 0.964182 |
| Support Vector Machine | 95.19 | 96.86 | 0.942857 | 0.985075 | 0.963504 | 0.938483 |
| Random Forest | 97.12 | 98.08 | 0.957143 | 1.000000 | 0.978102 | 0.959459 |
| Naive Bayes | 88.46 | 88.95 | 0.900000 | 0.926471 | 0.913043 | 0.866013 |

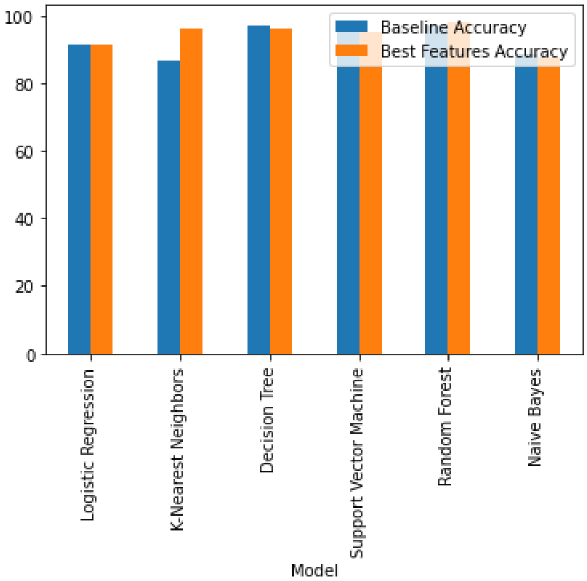
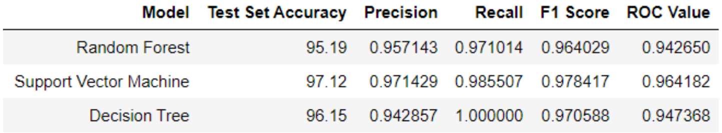


**Figure 7:** Baseline vs Improvised models comparison

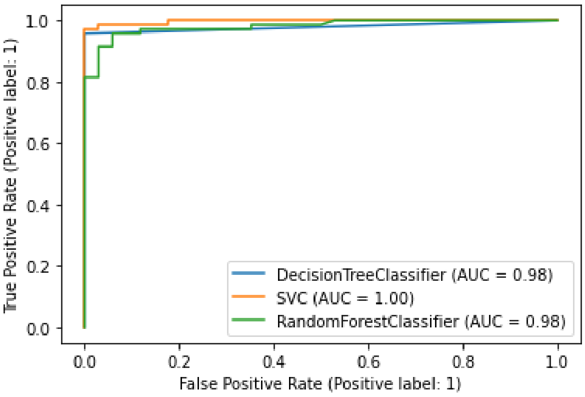
Figure 7 shows the comparison graphically of the baseline vs the improvised models. Three models: Random Forest, Support Vector Machine, and Decision Tree models were selected for further hyperparameter tuning to improve the performance of the models. With GridSearchCV class and running different set of parameters described in the Section 3, best parameters were found out. For Random Forest Classifier, best parameters were {'criterion': 'entropy', 'max_depth': 6, 'min_samples_leaf': 3, 'min_samples_split': 10, 'n_estimators': 50}. For Support Vector Classifier, best parameters were {'C': 100, 'gamma': 1, 'kernel': 'rbf'}. For Decision Tree Classifier, best parameters were {'criterion': 'entropy', 'max_depth': 10, 'max_features': 5, 'min_samples_leaf': 1}. After running the models with the hyperparameter tuning, the performance of Decision Tree classifier remained the same. Support Vector Classifier showed improvement in the accuracy score by 1.93%. The F1 Score is now 0.978417 which was 0.964029 earlier for the model which was not tuned (Table 4). From the ROC curve shown in the figure, Support Vector Classifier had an almost perfect classifier with an AUC close to 1 (Figure 8).

**Table 4:** Performance Metrics after Hyperparameter Tuning



| Model | Test Set Accuracy | Precision | Recall | F1 Score | ROC Value |
|---|---|---|---|---|---|
| Random Forest | 95.19 | 0.957143 | 0.971014 | 0.964029 | 0.942650 |
| Support Vector Machine | 97.12 | 0.971429 | 0.985507 | 0.978417 | 0.964182 |
| Decision Tree | 96.15 | 0.942857 | 1.000000 | 0.970588 | 0.947368 |

**Figure 8:** Receiver Operating Characteristic (ROC) Curve



## 5. Conclusions and Future Work

The objective of this study is to predict the likelihood of having early stage diabetes with respect to the symptoms found in the patient developing over time. Early detection and treatment of diabetes is a significant step toward keeping people with diabetes healthy. From the initial data analysis, it was found that the risk of diabetes is extremely high for patient above 70 years. Before applying the machine learning algorithms for the prediction, steps such as data pre-processing, label encoding, data transformation, and feature selection were performed. 8 best features were selected out of 16 features from the dataset. Decision tree classifier, Support vector Classifier, and Random forest classifier predicted better than the other models. Hyperparameter tuning was applied on these three algorithms to improve the performance of the models using various parameters. Highest accuracy score of 97.12%, F1 score of 0.978417 and the roc value of 0.964182 was obtained from the Support vector classifier algorithm after the tuning of hyperparameters.

Future work includes obtaining more training data for better performance of the machine learning algorithms. Neural Networks and ensemble techniques like Majority Voting can be performed on this data.

## References

[1] ‘Facts & figures’. https://idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html.

[2] ‘Pre-diabetes’. https://www2.hse.ie/conditions/pre-diabetes.html.

[3] ‘WHO | 10 facts on diabetes’, *WHO*. http://www.who.int/features/factfiles/diabetes/en/ .

[4] ‘Index of /ml/machine-learning-databases/00529’. https://archive.ics.uci.edu/ml/machine-learning-databases/00529/.

[5] M. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, ‘Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques’, in *Computer Vision and Machine Intelligence in Medical Image Analysis*, vol. 992, M. Gupta, D. Konar, S. Bhattacharyya, and S. Biswas, Eds. Singapore: Springer Singapore, 2020, pp. 113–125.

[6] S. Mirzajani and siamak salimi, ‘Prediction and Diagnosis of Diabetes by Using Data Mining Techniques’, *Avicenna Journal of Medical Biochemistry*, vol. 6, pp. 3–7, Jun. 2018, doi: 10.15171/ajmb.2018.02.