Selection of Machine Learning Algorithms for the Prediction of Diabetes with EDA

Module Code: UFCFMJ-15-M

Module Title : Machine Learning and Predictive Analytics

Student ID: 21069826

I. Introduction

Diabetes is one of the most common chronic diseases in the US, affecting millions of people annually and placing a heavy financial strain on the economy. According to the Centers for Disease Control and Prevention (CDC), 34.2 million Americans have diabetes and 88 million have prediabetes as of 2018. Kind II diabetes is the most prevalent type of the disease; its prevalence varies depending on factors such as age, education, income, place of residence, race, and other socioeconomic determinants of health. Further, it has been found that those with poor mental health have a 60% chance of developing diabetes as a result of biochemical alterations. So, this study mainly intends to find the patterns which have high impact on diabetes, including various types of uncommon attributes and to develop a prediction model. This study would be beneficial for the general population, researchers, policy makers and health care industries (CONDITIONS, n.d.).

II. Data set overview

The CDC conducts an annual telephone survey on health-related topics called the behavioral Risk Factor Surveillance System (BRFSS). Over 400,000 Americans participate in the survey each year, providing information on risky behaviors, chronic health issues, and usage of preventative treatments. Since 1984, it has been conducted each year. The Kaggle dataset for the year 2015 was used in this research as a csv file. This original dataset has 330 features and 441,455 responses from respondents (Data, n.d.).These characteristics are either direct participant inquiries or derived variables based on specific participant responses.

III. Methodology

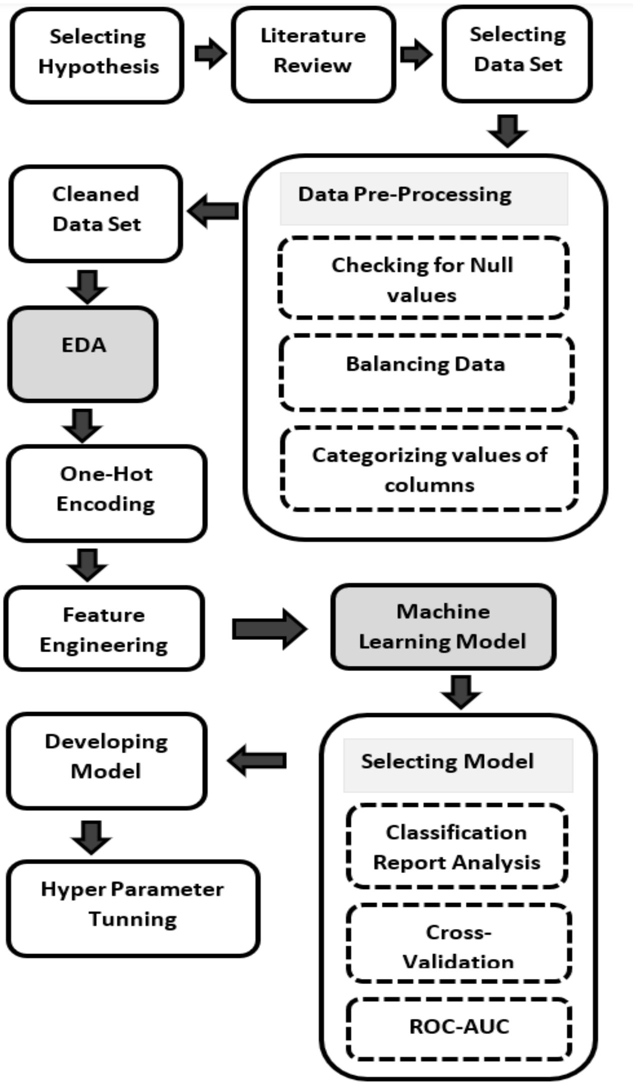


Figure-1

IV. EDA Analysis

Exploratory Analysis was able to find that those who are having cholesterol, high blood pressure, stroke issues, heart attacks and serious difficulty in walking have high impact of having diabetes than general population. Further it has been found that, people who are suffering bad in mentally and physically more than 25 days per month have high probability to have diabetes. In addition, those who take fruits and vegetables regularly has low possibility to occur diabetes.

Further, it can be seen than males are more likely to undergo with diabetes than females while the age above 60 people are more likely to have diabetes than other age groups. When the income levels of individuals are less than $ 25000 and those who do not visit a doctor because of financial issues, they have high tendency to suffer diabetes. This may happen because having low-income level can lead for poor mental health issues. It was confirmed that mental health has an impact on the presence of diabetes from both given results and literature reviews. Therefore, indirectly those factors can lead to have diabetes. Body Mass Index (BMI) is another risk factor for diabetes which means high BMI are more likely to occur diabetes. The following are some examples of pie charts that were taken from EDA.

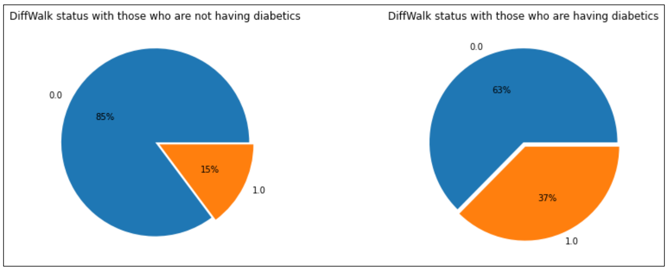


Figure 2-Serious walking difficulty vs diabetes

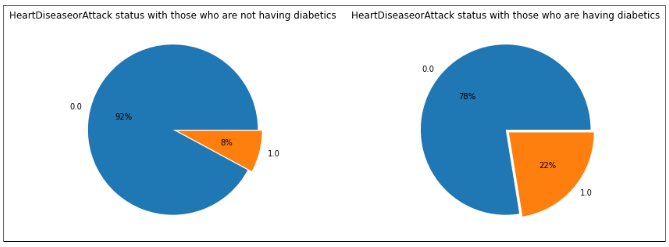


Figure 3-Heart Attack vs diabetes

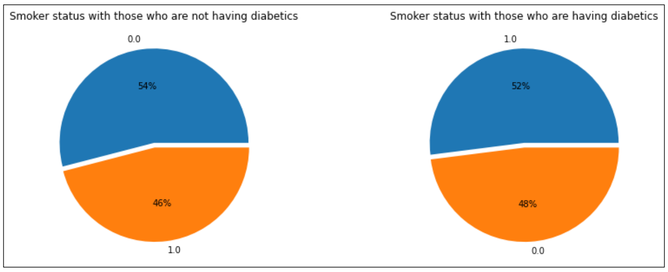


Figure 4-Smoking vs Diabetes

After analyzing the results of exploratory data analysis, features that have high impact on the presence of diabetes, have been selected using chi-square method. Here, all the column types of data frame were converted into object type in order to do one-hot encoding. Then the most important features to develop machine learning model were filtered using chi-square method at the feature engineering stage as follow.

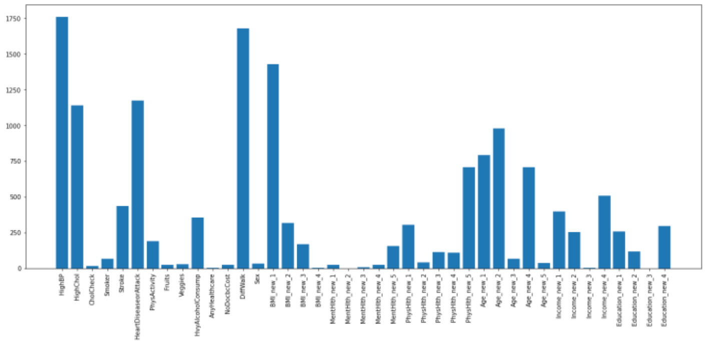


Figure 5-Feature Selection using chi-square

- High Blood Pressure
- Serious Difficulty in Walking
- Heart Attack
- High Cholesterol
- Having physical health issues more than 20 days per month
- Income level is between $ 25000 and $ 50000
- Having Stroke issues
- Age group between 18-30 or between 45-60

These attributes were taken to develop several machine learning models. Here, several supervised machine learning algorithms have been considered, including Logistic Regression, Random Forest, Decision Tree, Naïve bayes, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Xtreme Gradient Boosting (XGB) and two neural network models in order to select the best model.

1. K-Nearest Neighbour (KNN)

K-Nearest Neighbor is one of the machine learning algorithms which can be used for both classification and regression problems. First number of nearest neighbors will be defined and Euclidean distances between data points will be calculated to find nearest neighbors (Algorithm, n.d.). This does not give high accuracy when there is large amount of data in the training data set and high dimensions. It is also sensitive with outliers of the data set as it takes all details of inputs rather than generalizing data.

2. *Random Forest*

The ensemble learning technique known as random forest can be applied to both classification and regression applications. It is constituted with numerous decision trees, and they will get trained linked to each random sample data. Row and feature sampling techniques are used by each decision tree in the random forest to get sample data sets. Here, multiple decision trees will be executed as weak learners to develop a strong learner (Forest, n.d.).

3. *Logistic Regression*

Logistic regression can be applied for classification problems when the target value is categorical in nature. It takes the values between 0 to 1 and gives 0 or 1 outputs by using sigmoid function. Normally threshold value is taken as 0.5 to determine whether the output is 0 or 1. It can perform well when the dataset is linearly separated and has good accuracy for a variety of simple data sets. It has a linear decision surface; hence it cannot address non-linear issues (Regression, n.d.).

4. *Decision Tree*

Decision trees are a method used in supervised machine learning, a method that trains models using labelled input and output datasets. The method is mostly used to address classification issues, which include using a model to categorize or classify an object. Here data does not need to be scaled and normalized (Trees, n.d.). But a minor change in the data can result in a big change in the decision tree's structure, which can lead to instability.

5. *Support Vector Machine(SVM)*

support vector machine (SVM) is a supervised machine learning model that applies classification problems. It operates more efficiently in high dimensional spaces and uses relatively little memory. When there is more noise in the data set and there are more training data samples than features for each data point, SVM does not perform very well (Algorithm, n.d.).

6. *Bernoulli Naïve Bayes*

Naive Bayes classifier estimates the probability that a given input will fall into each of the classes as a probabilistic classifier. Bernoulli Naive Bayes is used for discrete data when there are binary values of columns. It is quick, able to make predictions in real time, and handles irrelevant features well. But if the features occasionally depend on one another, naïve bayes assumption can alter the accuracy of the model (Algorithm, n.d.).

7. *XGBoost*

Extreme Gradient Boosting, known as XGBoost, is a commonly used supervised-learning technique for regression and classification on huge datasets. Decision trees are used by XGBoost as basic learners, combining several weak learners to create a strong learner. As a result, it is known as an ensemble learning approach because the final prediction combines the output of numerous models. It encourages regularization and is very adaptable (XGBoost, n.d.).

8. *Keras Sequential Model*

The Sequential model API allows for the creation of deep learning models by creating objects of the Sequential class and adding model layers to them. It deploys quickly and is user-friendly. Models built with Keras rather than TensorFlow are more likely to be correct since Keras is less error prone. This is so that Keras can function under the framework's limitations, which include low-level errors and fast computation speed (Analysis of Preprocessing Techniques, n.d.), (Prediction, n.d.) (When to Use MLP, n.d.).

9. *Multilayer Perceptron(MLP)*

MLP, or multilayer perceptron, is the standard kind of neural network. They are made up of a single layer of neurons or several layers. It can be used both classification and regression problems as well as tabular data sets. In addition, it can be used to solve difficult non-linear issues and utilizes substantial data sets. The quality of the data set determines how well the system works, and calculations are challenging and time-consuming since there are too many parameters (When to Use MLP, n.d.), (Networks, n.d.)..

Selection of model have been considered in three parts as following.

1. Cross- Validation
2. Classification Report Analysis
3. Receiver Operating Characteristics Curve (ROC) and Area under the ROC curve (AUC)

1. Cross Validation

Model selection was carried out using k-fold cross validation among the supervised machine learning models. Here, data set was simply divided into 10 partitions and accuracy of the model was obtained by getting the average value of each accuracy related to the divided data set. As this is repeated only 10 times, computation time is low when calculating scores. Following are the results of each model that were obtained from k-fold cross validation.

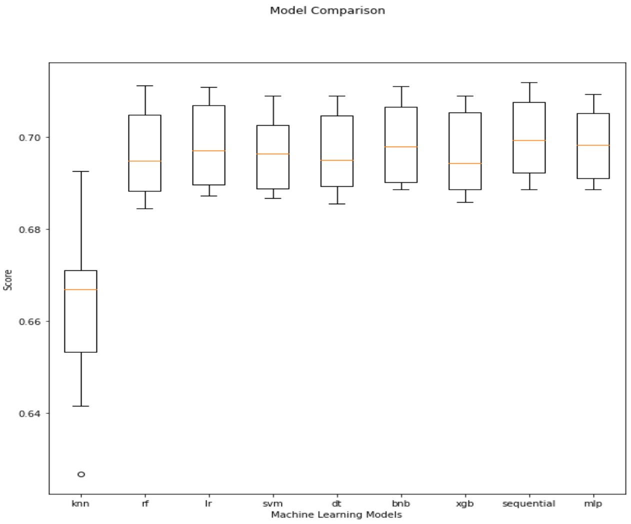


Figure 6-Cross Validation Score for each model

From the above cross validation diagram, sequential model, Multilayer perceptron, binomial naïve bayes, logistic regression and support vector machine look like same accuracies while K-Nearest Neighbor has the lowest average accuracy. Further it shows that, keras sequential model has the highest average accuracy than other models. One of the main disadvantages of using k-fold cross validation is that it can lead to overfitting when doing multiple times. For further confirmation, all the classification reports were analyzed in order to find the best model.

2. Classification Report Analysis with ROC & AUC

Confusion matrices and classification reports can be used easily to determine the model performance. As the data set was balanced at the data processing stage, it is reasonable to compare the accuracy of each model. But data set had not been exactly balanced after splitting into training and testing data sets. However, as the prediction model is related to diabetes prediction model, following states had to be considered when selecting the best model.

- Predicting someone is having diabetes when someone is not having diabetes. (Also called as type 01 error – False Positive Rate)
- Predicting someone is not having diabetes when someone is having diabetes. (Also called as type 02 error – False negative Rate)

In this scenario, both these cases have to be reduced as it gives offence results for users. In order to reduce both these two parameters, F_β score was considered as following.

$$F_{\beta\ score} = (1+\beta^2) \times \frac{Precision\ \times Recall}{\beta^2\ (Precision + Recall\ )}$$

As above two concerns needed to be reduced β was selected as one. Then equation can be derived as following.

$$F_{1\ score} = 2 \times \frac{Precision\ \times Recall}{Precision + Recall}$$

F1 score was selected related to the positive class of the output. That means that the best model was selected based on following two scenarios.

1. To what extent out of the patients who were predicted as diabetes patients by the model are matched with actual diabetes patients. (Precision)
2. To what extent out of all the patients that suffered from diabetes are matched with the diabetes patients who were predicted by the model. (Recall)

F1-scores were obtained for each model by getting confusion matrix and classification reports.

Table 1-F1 Scores for each machine learning model

Model	**F1-Score**
KNN	63.05110%
Random Forest	69.38230%
Logistic Regression	68.67860%
Decision Tree	69.37030%
Support Vector Machine	69.64020%
Binomial Naïve Bayes	69.03750%
XGBoost	69.44770%
Keras Sequential Model	71.00240%
Multilayer Perceptron	69.73330%

Above scores taken from classification reports, Keras sequential model has high F1 score than other models. Confusion matrix and classification report for keras sequential model have been obtained as follows.

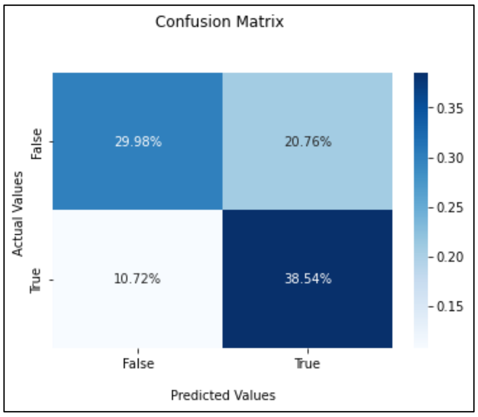
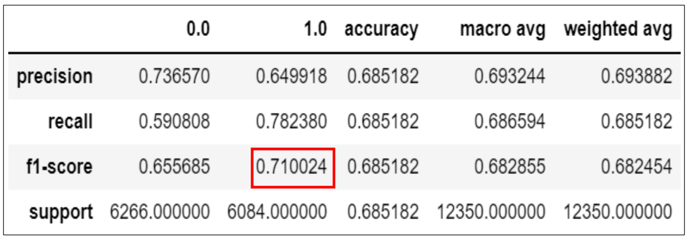


Figure 7- Confusion Matrix for keras sequential model



	0.0	1.0	accuracy	macro avg	weighted avg
precision	0.736570	0.649918	0.685182	0.693244	0.693882
recall	0.590808	0.782380	0.685182	0.686594	0.685182
f1-score	0.655685	0.710024	0.685182	0.682855	0.682454
support	6266.000000	6084.000000	0.685182	12350.000000	12350.000000

Figure 8- Classification Report for keras sequential model

In order to plot ROC and AUC, True positive rate (TPR) and False positive rates were calculated. Then AUC scores were obtained as follows.

Table 2- AUC score for each machine learning model

Model	AUC score
KNN	64.007%
Random Forest	68.749%
Logistic Regression	68.911%
Decision Tree	68.736%
Support Vector Machine	68.965%
Binomial Naïve Bayes	68.797%
XGBoost	68.740%
Keras Sequential Model	68.659%
Multilayer Perceptron	68.974%

According to the AUC values taken from ROC, every model has been taken almost same scores except KNN. Further the highest AUC score goes to Multilayer Perceptron

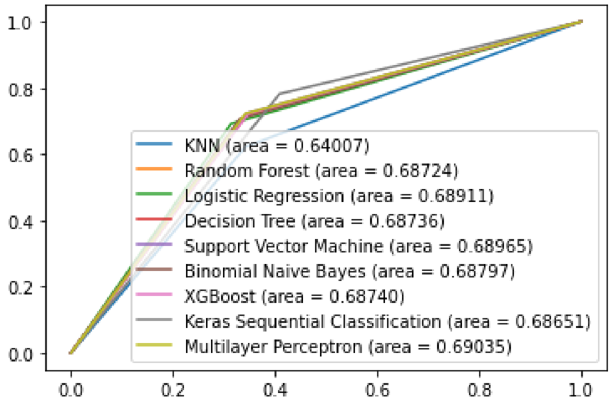


Figure 9- AUC scores for each model

V. Results Analysis

According to the cross-validation scores taken for each model, it was found that the best accuracy shows keras sequential model. After analyzing the classification reports related to the diabetes use case, the highest F1-score was taken by keras sequential model. Finally, the highest AUC was obtained for Multiplayer Perceptron. Therefore, comparison between AUC and F1 score had to be done in order to select the best method. Even though the data set was balanced at the initial stage related to the target feature, it was found that data set was not balanced after splitting in to training and testing data. Normally, F1-score is considered more significant than AUC when there is an imbalanced data set, or it is used for communicating results to end user. Therefore, Keras sequential model was selected as the best model to predict the presence of diabetes based on the results of cross validation and classification reports.

VI. Conclusion

This study produced several intriguing findings, most notably that diabetes can occur in people who also had high blood pressure, serious walking difficulty, heart attacks, high cholesterol, physical health concerns more than 20 days a month, and stroke issues. Different algorithms were taken into consideration regarding the diabetes use case. The ideal machine learning algorithm was ultimately decided upon as the Keras Sequential Model.

VII. Limitations And Further Research

The results may have limited applicability to other nations because the survey was conducted in the US using information gathered from respondents of a survey conducted there. The qualities might be correlated in both directions. Determining what occurred at the beginning of an illness is therefore quite difficult. It is better to do face-to-face interviews as they are more trustworthy and informative for surveys than telephone interviews as the interviewer can observe key body signs that occur during a talk about issues relating to health diseases.

VIII. References

Algorithm, K.-N. N., n.d. *IBM.* [Online] Available at: https://www.ibm.com/topics/knn

Algorithm, N. B., n.d. *KDnuggets.* [Online] Available at: https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html

Algorithm, S. V. M., n.d. *GeeksforGeeks.* [Online] Available at: https://www.geeksforgeeks.org/support-vector-machine-algorithm/

Analysis of Preprocessing Techniques, K. T. a. T. L. o. C. S. i. d., n.d. *IEEE Xplore.* [Online] Available at: https://ieeexplore.ieee.org/abstract/document/9671878?casa_token=OJx7liThauEAAAAA:a5rLgSPUyJRHLp3AQZ5klo2Vh-GdiwfbAjG-DznOgEEmzyTmDLqIER5VL_2cXAHNX-_5GNk7

CONDITIONS, D. R., n.d. *DIABETES UK.* [Online] Available at: https://www.diabetes.org.uk/diabetes-the-basics/related-conditions

Data, A. S., n.d. *Centers for Disease Control and Prevention.* [Online] Available at: https://www.cdc.gov/brfss/annual_data/annual_data.htm

Forest, R., n.d. *IBM.* [Online] Available at: https://www.ibm.com/cloud/learn/random-forest

Networks, C. C. o. M.-L. P. N., n.d. *Machine Learning Mastery.* [Online] Available at: https://machinelearningmastery.com/neural-networks-crash-course/

Prediction, K. -. M. E. a. M., n.d. *tutorialspoint.* [Online] Available at: https://www.tutorialspoint.com/keras/keras_model_evaluation_and_prediction.htm#:~:text=score%20%3D%20model.evaluate%28x_test%2C%20y_test%2C%20verbose%20%3D%200%29%20print%28%27Test,a%20best%20model%20to%20identify%20the%20handwriting%20digits.

Regression, L., n.d. [Online] Available at: https://www.sciencedirect.com/topics/computer-science/logistic-regression

Trees, D., n.d. *scikit learn.* [Online] Available at: https://scikit-learn.org/stable/modules/tree.html

When to Use MLP, C. a. R. N. N., n.d. [Online] Available at: https://machinelearningmastery.com/when-to-use-mlp-cnn-and-rnn-neural-networks/

XGBoost, n.d. *GeeksforGeeks.* [Online] Available at: https://www.geeksforgeeks.org/xgboost/

Words Count : 2188 (*without considering the words included to figures and tables)*