Diabetes Prediction Using Machine Learning

Represented By:

Manav Pandya - 30169952
Hetalben Virani - 30183515
Pujan Patel - 30169747

ENEN 682 – Applied Machine

Learning and Predictive Analytics

Professor Leanne Dawson

April 12th, 2023

Abstract

Diabetes is a chronic condition that could lead to a global health care disaster.382 million people worldwide have diabetes, according to the International Diabetes Federation. This will double to 592 million by 2035. Diabetes, often known as diabetes mellitus, is a condition brought on by elevated blood glucose levels. For diagnosing diabetes, several conventional techniques based on physical and chemical examinations are available. Although diabetes affects human organs including the kidney, eye, heart, nerves, foot, and others, early diabetes prediction is a difficult task for medical professionals due to intricate dependencies on many elements. Data science techniques can advance other scientific disciplines by providing fresh perspectives on old problems. Making predictions using medical data is just one of many tasks. The fundamental aim of this project is to design a diabetes prediction model using four different machine learning methods including Logistic Regression, Support Vector Machine (SVM), Random Forest Classifier as well as Neural Network.

Table of Content

List of Figures

Introduction

In today's fast paced world, diabetes is unarguably the fastest growing disease amongst the people, including youngsters. To elaborate, Sugar gets injected into our body through the food we consume, especially carbohydrate food. However, one can't stop consuming carbohydrates as it is the primary source of energy, even those people with diabetes need carbohydrates. Carbohydrates foods include bread, cereal, pasta, rice, fruit, dairy products, and vegetables. When we consume food, the body breaks it down into glucose. In the bloodstream, glucose circulates throughout the body. For us to think effectively and function, some of the glucose is transported to our brain. The remaining glucose is transferred to our body's cells for usage as fuel, and it is also stored as energy in our liver for later use by the body. Insulin is necessary for the body to use glucose as fuel. The beta cells in the pancreas create the hormone insulin. Insulin functions as a door's key. To allow glucose to enter the cell from the blood stream, insulin attaches to the cell's doors, opening them. Glucose builds up in the bloodstream (hyperglycemia) and diabetes occurs if the pancreas is unable to generate enough insulin (insulin deficit) or if the body is unable to utilize the insulin it produces (insulin resistance). Diabetes Mellitus is characterized by elevated glucose (sugar) levels in the blood and urine.

1 Data Set

The data set which we have used is collected by conducting large scale surveys of Pima's (North American Indians who traditionally lived along the Gila and Salt rivers in Arizona United States). Data set has been collected from Kaggle. The fundamental objective of the dataset is to predict whether the patient has diabetes or not. Moreover, there are several independent variables and one dependent variable. The independent variables include the number of previous pregnancies, body mass index, the insulin level, age etc. On the other hand, the dependent variable is the "Outcome", which is a feature we are going to predict, 0 means no diabetes and 1 means diabetes.

In detail description of the data set can be found in the below mentioned table:

Serial No	Assigned Names	Description
1	Pregnancies	Number of times pregnant
2	Glucose	Plasma-Glucose Concentration
3	Blood Pressure	Diastolic blood pressure
4	Skin Thickness	Triceps skin fold thickness in mm
5	Insulin	2-h serum insulin
6	BMI	Body mass index
7	Diabetes-Pedigree Function	Diabetes pedigree function

8	Outcome	Class variable (0 or 1)
9	Age	Age of Patient

Figure 1. Dataset Parameters

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Figure 2. Dataset Values

- The dataset consists of 786 data points, with 9 features each.
- "Outcome" is the feature we are going to predict, 0 means no diabetes and 1 means diabetes.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Figure 3. No-null data values.

From the above results, it can be stated that there are no null values in the dataset.

1.1. Flowchart of Diabetes Prediction Model

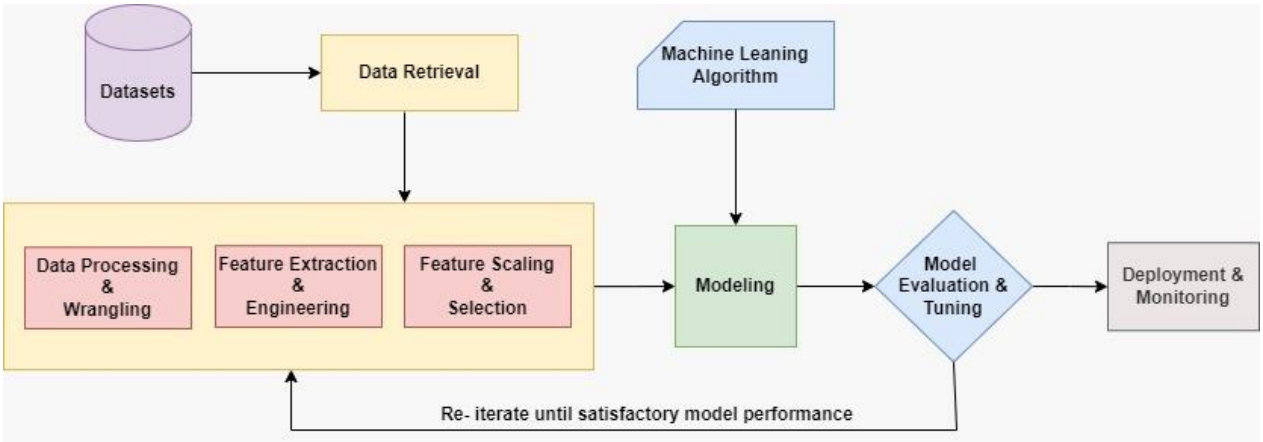


Figure 4. Flowchart of the Diabetes Prediction Model.

1.2. Correlation Matrix

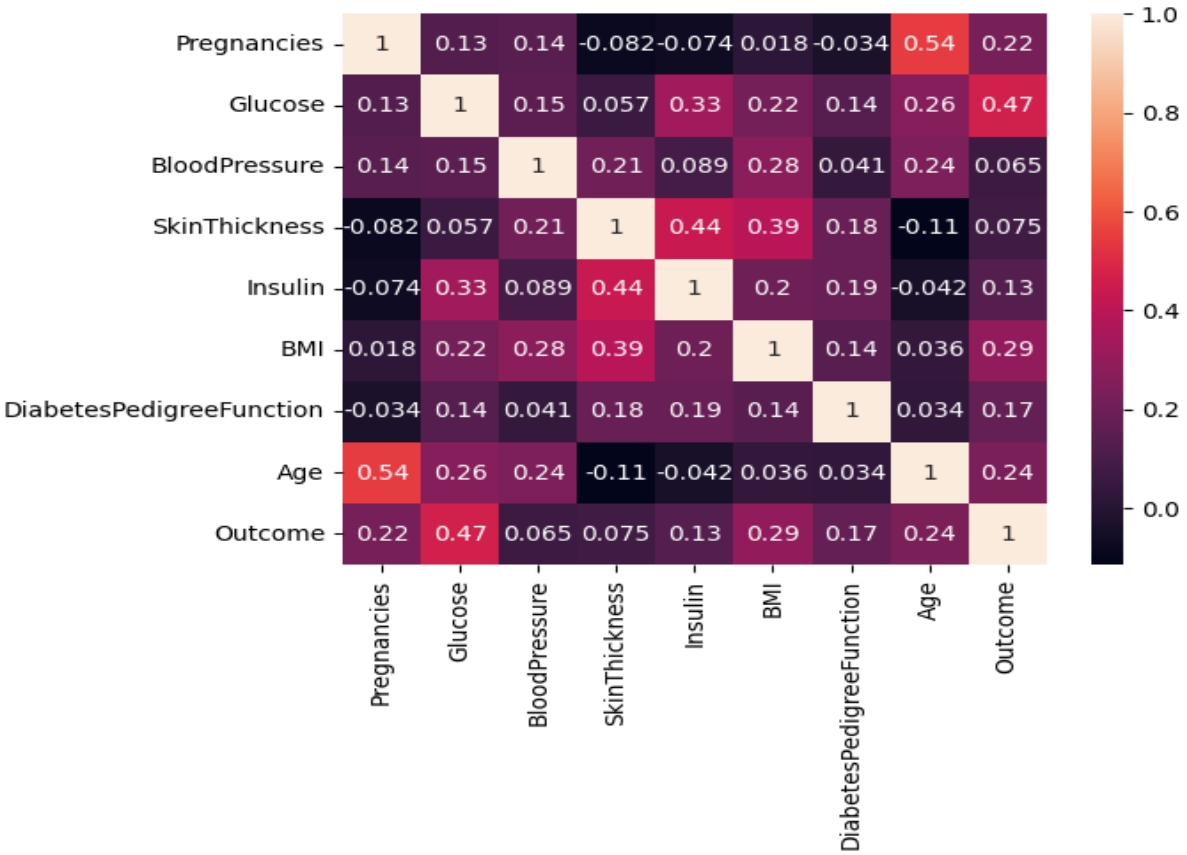
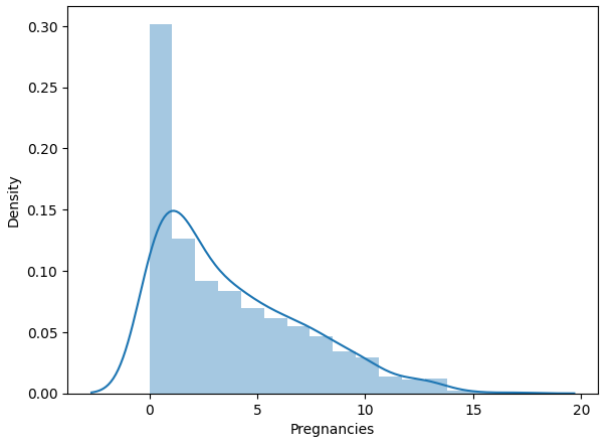
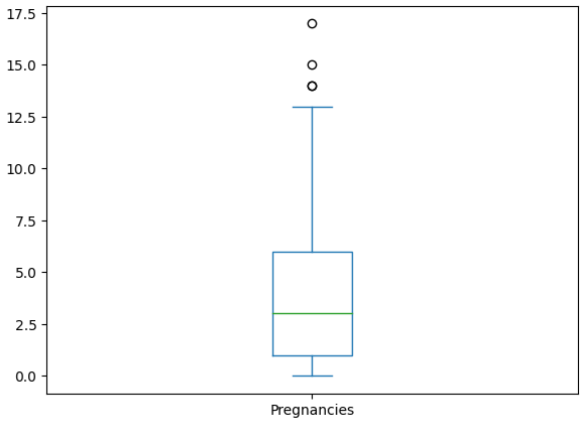


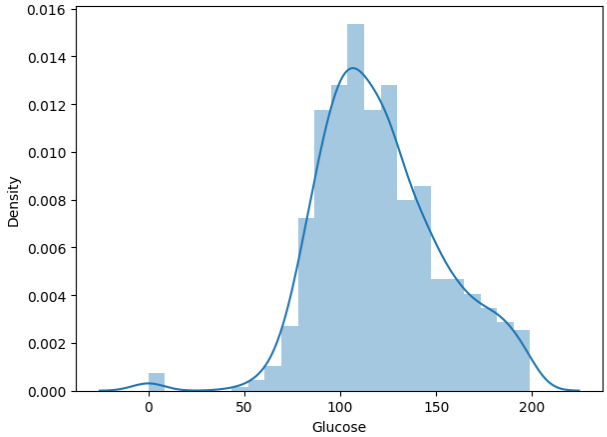
Figure 5. Correlation values and heat map.

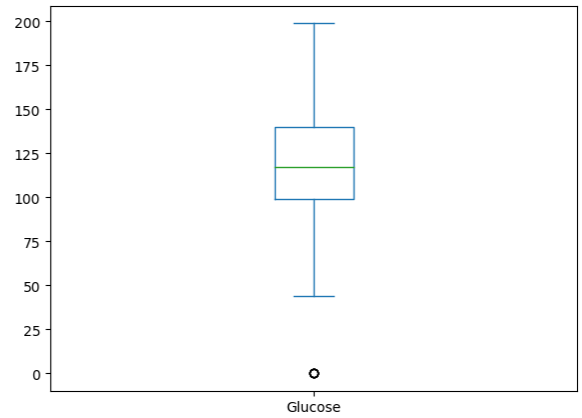
No particular property has a strong link with our result value. Some attributes and the outcome value correlate negatively whereas others do not.

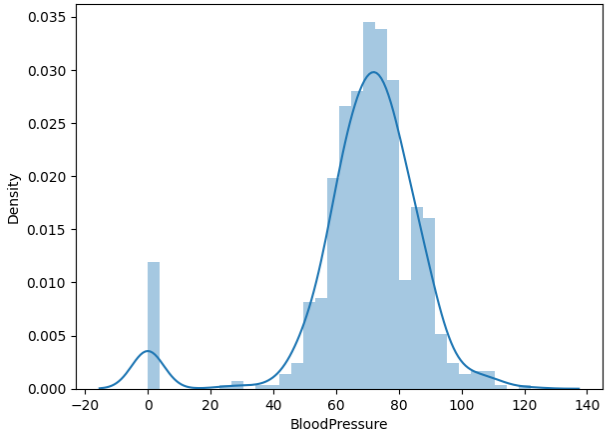
1.3. Skewing the Data

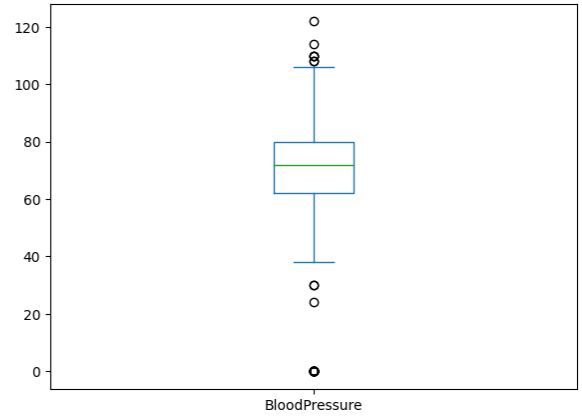


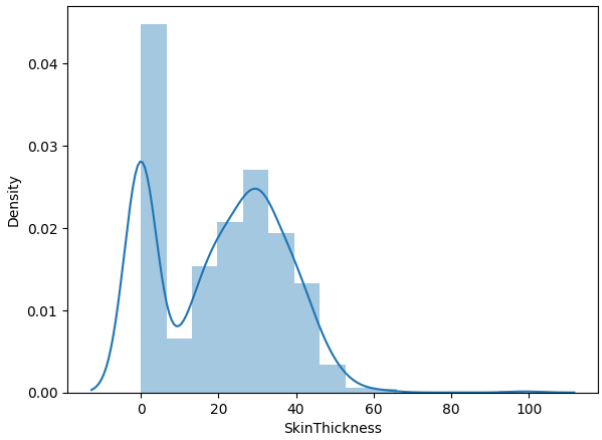




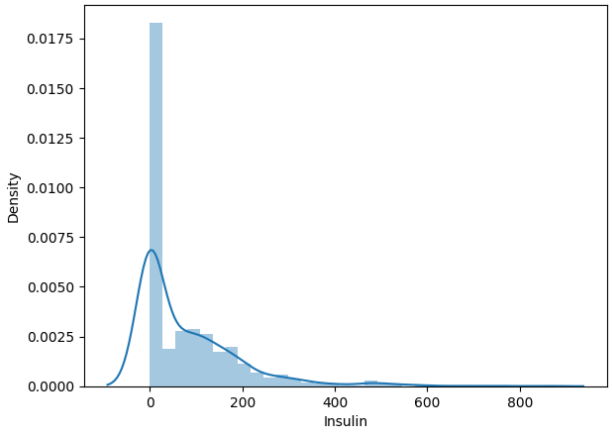


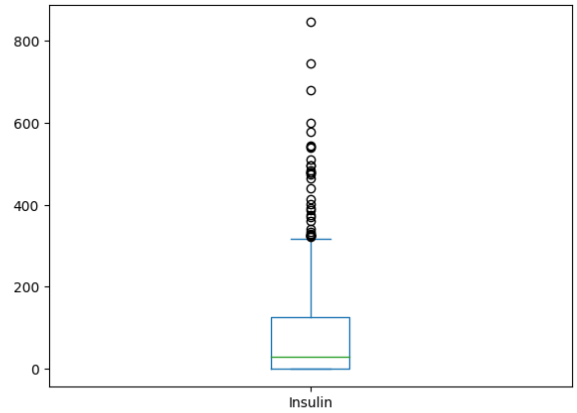


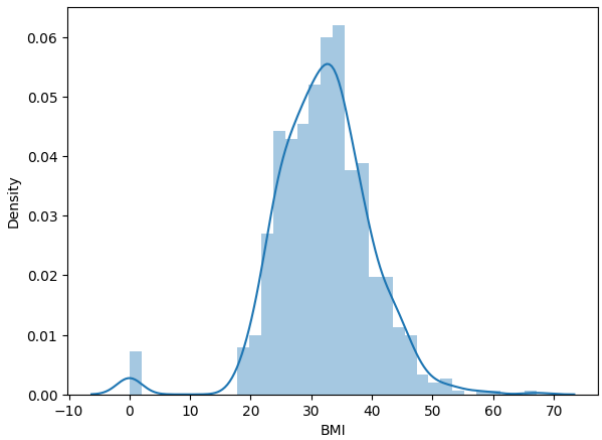


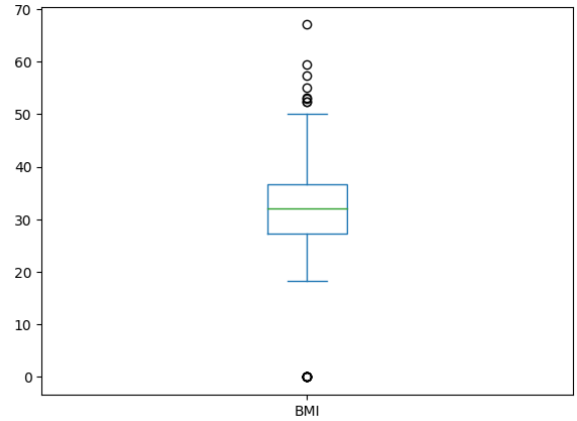

Density
SkinThickness


SkinThickness


Density
Insulin


Insulin


Density
BMI


BMI

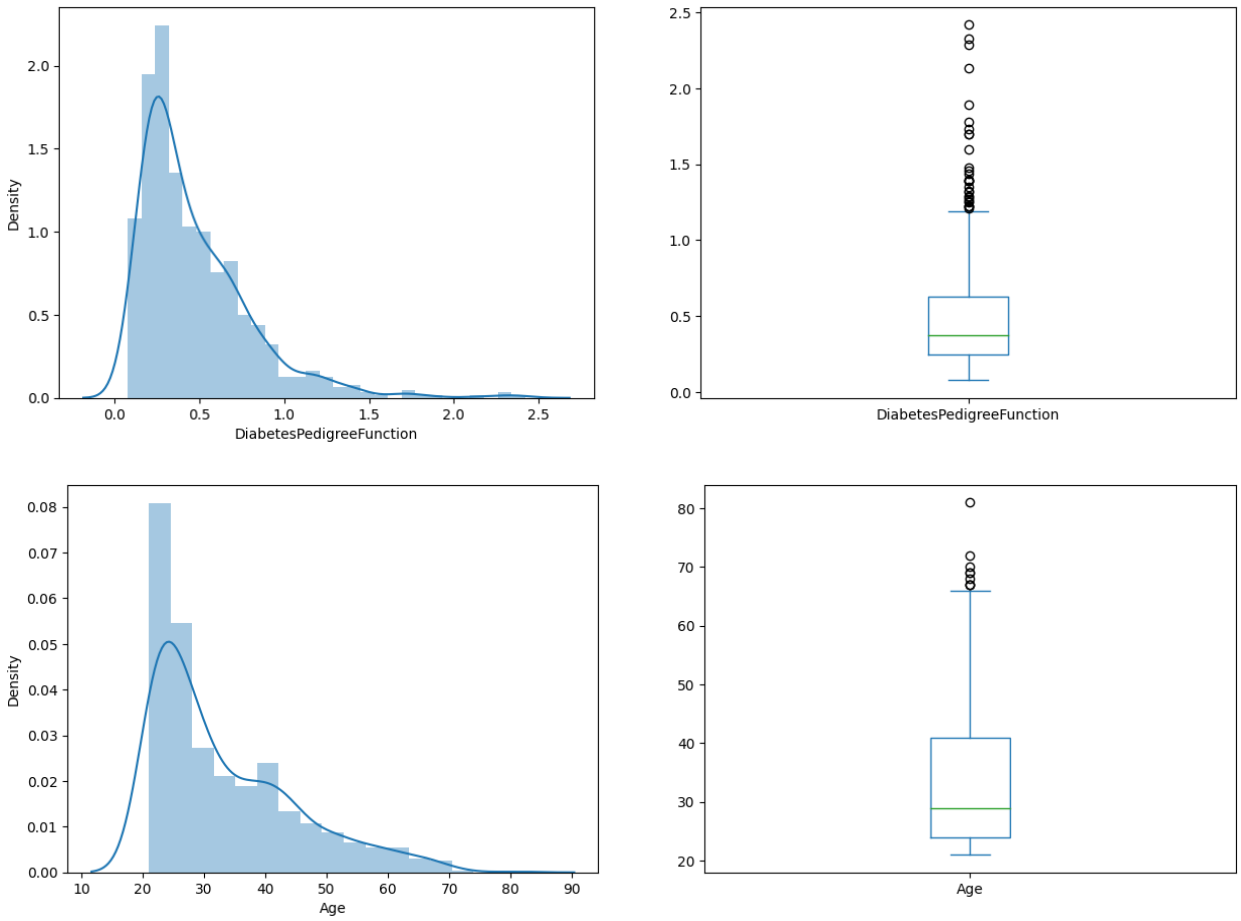


Figure 6. Skew of data

Let's look at the plots. It demonstrates how each feature and label is dispersed across many ranges, further demonstrating the necessity of scaling. Next, discrete bars basically indicate that each of them is a categorical variable anywhere they are present. Before using machine learning, we need to handle these categorical variables. We have two categories for our outcome labels: 0 for no disease and 1 for disease.

1.4. Visualizing the outcome values

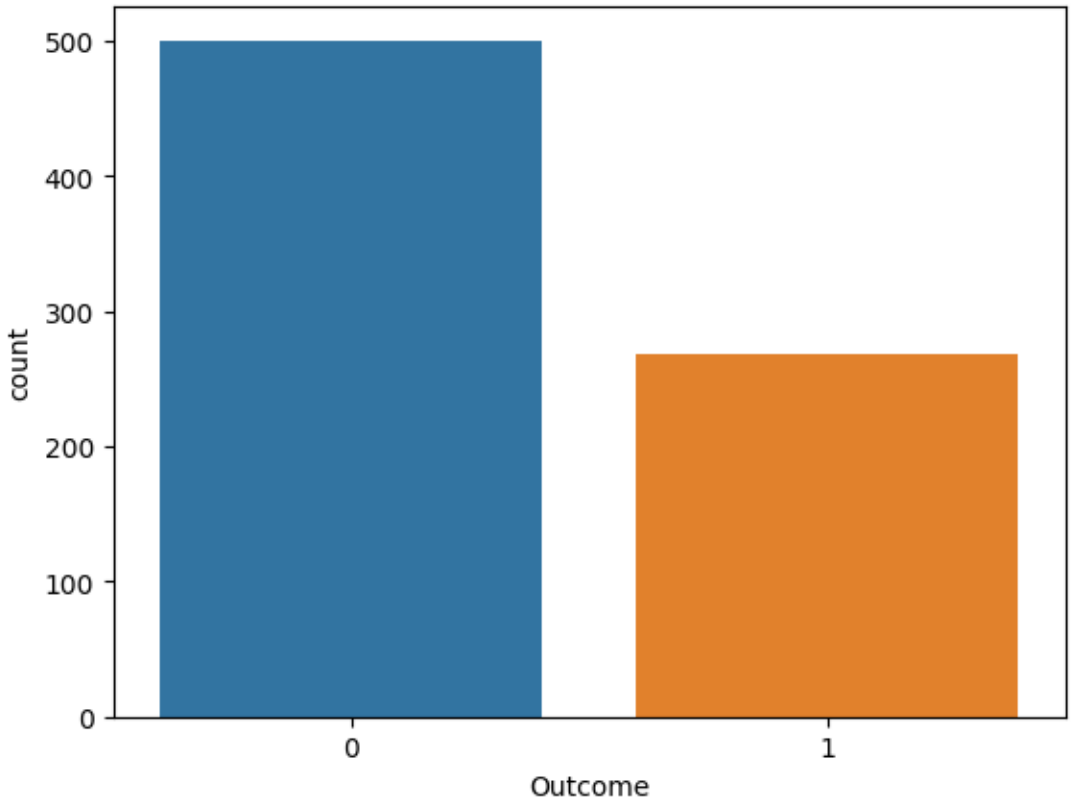


Figure 7. Bar plot for outcomes class

The graph above demonstrates how the data is skewed towards datapoints with result values of 0, which indicate that diabetes was not actually present. Diabetic patients are roughly two times as numerous as non-diabetics.

2 Proposed Methods

2.1 Dataset Collection

This involves collecting the data and making sense of data to explore hidden patterns and trends that help predict and evaluate results. The Dataset carries 768 rows, i.e., the total number of data and 10 columns i.e., total number of features. The features include vivid aspects such as Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, Body Mass Index (BMI), Diabetes Pedigree Function, Age as well as Outcome.

2.2 Missing Values Identification

The dataset has no missing values, hence there was no need for replacing the values with some other values.

2.3 Feature Selection

The Pearson Correlation method is a popular way to find the most important attributes. This method calculates a correlation coefficient that correlates the output and input attributes. The value of coefficient is between -1 and 1. A value above -0.5 means a significant correlation and zero means no correlation.

	Pregnancies	Glucose	BloodPressure	SkinThickness	\
Pregnancies	1.00	0.13	0.14	-0.08	
Glucose	0.13	1.00	0.15	0.06	
BloodPressure	0.14	0.15	1.00	0.21	
SkinThickness	-0.08	0.06	0.21	1.00	
Insulin	-0.07	0.33	0.09	0.44	
BMI	0.02	0.22	0.28	0.39	
DiabetesPedigreeFunction	-0.03	0.14	0.04	0.18	
Age	0.54	0.26	0.24	-0.11	
Outcome	0.22	0.47	0.07	0.07	

	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	-0.07	0.02	-0.03	0.54	0.22
Glucose	0.33	0.22	0.14	0.26	0.47
BloodPressure	0.09	0.28	0.04	0.24	0.07
SkinThickness	0.44	0.39	0.18	-0.11	0.07
Insulin	1.00	0.20	0.19	-0.04	0.13
BMI	0.20	1.00	0.14	0.04	0.29
DiabetesPedigreeFunction	0.19	0.14	1.00	0.03	0.17
Age	-0.04	0.04	0.03	1.00	0.24
Outcome	0.13	0.29	0.17	0.24	1.00

Figure 8. Feature selection

2.4 Splitting the Data

Once the data was cleaned, the data set was ready to train and test. Using the train/split method, we split the dataset randomly into training and testing set. For training we took 614 samples and 154 were selected for testing.

```
Shape of X_train:  (614, 8)
Shape of y_train:  (614,)
Shape of X_test:  (154, 8)
Shape of y_test:  (154,)
```

Figure 9. Splitting the data

2.5 Design and implementation of classification model

For this project, we will primarily use four machine learning algorithms namely, Logical Regression, Support Vector Machine, Random Forest Classifier as well as Neural Network.

2.6 Machine Learning Classifier

We developed the model using machine learning technology. Various classification and clustering techniques have been used to predict the diabetes dataset. We applied Support Vector Machine (SVM), Logical Regression, Neural Network and Random Forest machine learning classifier to analyze the performance by finding the accuracy of each classifier. All classifiers are implemented in Python using the scikit tutorials. The applied classification algorithms are described in the next section.

3 Modeling and Analysis:

To build a model to predict diabetes, we used machine learning techniques, including different classifier and ensemble approaches such as LR, SVM, NN, and RF. Machine Learning classifiers to evaluate the accuracy of each classifier and analyse the performance. Each classifier is implemented using the scikit-learn modules for Python.

3.1 Logistic Regression

Logistic regression is used to analyze the relationship between the categorial variable and one or more independent variable. This is mainly used for the classification problem, where the dependent variable takes only two values for e.g., 0 or 1. The logistic regression model find the best fit of the logistic function of the data, which is sigmoid function that maps the values of independent variable to the probability of dependent variable.

```
[134] from sklearn.metrics import accuracy_score
      accuracy_score(Y_test,logreg.pred)

      0.8246753246753247
```

Figure 10. Result of Logistic Regression model

As per observation, in Logistic Regression it returned best score 82.46%.

3.2 SVM (Support Vector Machine)

SVM is a type of supervised learning algo that can be used for both classification and regression problem. The goal of SVM is to find the hyperplane which separate the two clusters according to their class and to maximize the margin, which is distance btw the hyperplane and the closest data of each class. SVM takes the data from both class which are closest to the hyperplane and draw the margin and with the help of the margin it draws the hyperplane and the two data which we have taken to draw our margin are known as support vectors.

```
[142] accuracy_score(Y_test,svm_model.pred)

      0.8181818181818182
```

Figure 11. Result of SVM model

SVC Model gave 81.81% accuracy which is bit less than Logistic Regression.

3.3 Random Forest Classifier

RFC is the combination of the decision tree to make the more accurate prediction. RFC is also used for the classification problem. In RFC each decision tree is trained randomly with the random number of input or data and random features of the data, this helps in reducing the overfitting of the model. The prediction is given by considering all the output and taking a mean of all the output. RFC is versatile algo which can achieve high accuracy in a variety of task.

Figure 12. Result of RFC model

Random forest model gave max 80.51% accuracy which is almost same to SVM model.

3.4 Neural Network

we have used sequential which is a type of the neural network model that allow us to create a stack of layers in a sequence. Here we create multiple layers where each layer is related with the previous layer. We have used the dense layer of the layer of the sequential and we have applied the rely on activation function in our program and we have given the input dim according to our feature.

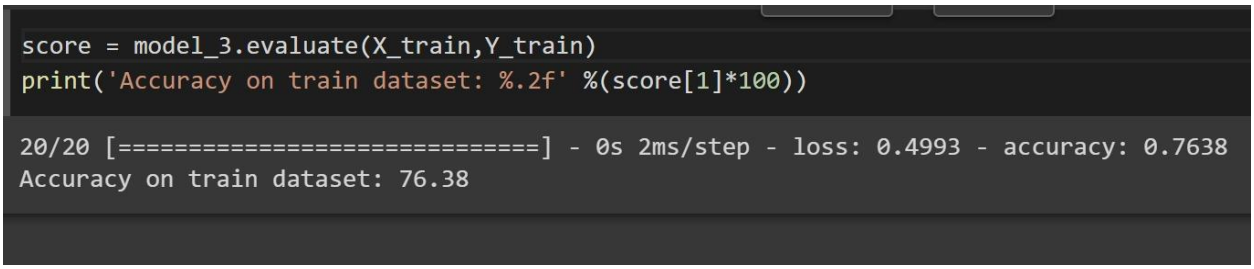


Figure 13. Result of Neural Network model

It is observed that, Neural network gave less the data accuracy of diabetic dataset which is 76.38% as compared to other models that we used in this project.

4. Results and Discussion

Machine Learning Algorithm	Result
Logistic Regression	**82.46%**
Support Vector Machine	**81.81%**
Random Forest Classifier	**80.51%**
Neural Network	**78.38%**

Figure 14. Tabular Comparison

Thus, a Machine Learning classification algorithm was developed which can predict diabetes in the initial stage. From the above table, it is quite evident that Logistic Regression predicted with the highest accuracy as high as 82.46%, whereas Neural Network was the lowest giving only 78.38 percent.

Conclusion

The main objective of this projects was to create a model that could recognise diabetic individuals who are most likely to be admitted to the hospital. The task of predicting the likelihood of hospital admission is very challenging. This procedure and its result are influenced by numerous factors. Methods to improve healthcare institutions' understanding of what matters in predicting the probability of hospital admission are urgently needed right now. By suggesting a system that might be utilised as an aid in identifying the patients who are most at risk of developing diabetes, this effort makes a tiny contribution to the currently employed methods of diabetes detection. This study accomplishes this by employing a variety of machine learning models to analyse numerous important aspects, including the patient's blood glucose level, body mass index, etc., as well as through retrospective examination of the patients' medical data. The model predicts with a 82.46% accuracy rate, which is respectable and trustworthy.

References

1) World Health Organization, 2021.
https://www.who.int/news-room/fact-sheets/detail/diabetes.
2) National Institute of Diabetes and Kidney Diseases, 2021,
https://www.niddk.nih.gov/health-information/diabetes.
3) A. Mujumdar, V. Vaidehi
Diabetes prediction using machine learning algorithms.
International Conference on Recent Trends in Advanced Computing", 2019, ICRTAC (2019).
4) PIMA Indian Dataset, 2021.
https://www.kaggle.com/uciml/pima-indians-diabetes-database.
5) Dataset from Vanderbilt, 2021.
https://data.world/informatics-edu/diabetes-prediction .
6) American Diabetes Association, Diabetes Care,
https://care.diabetesjournals.org/content/30/6/1562#:~:text=Individuals%20with%20BMI%20%E2%89%A530,life-years%20lost%20to%20diabe.
7) "Feature selection", 2021 Scikit learn.
https://scikit-learn.org/stable/modules/feature_selection.html.
8) "Chi-Square Test for Feature Selection – Mathematical Explanation", GeeksforGeeks 2021.
https://www.geeksforgeeks.org/chi-square-test-for-feature-selection-mathematical-explanation/?ref=lbp.
9) M.R.H. Subho, M.R. Chowdhury, D. Chaki, S. Islam, M.M. Rahman, "A univariate feature selection approach for finding key factors of restaurant business", Conference Paper, June 2019.
10) Mayo Clinic, 2021.
https://www.mayoclinic.org/diseases-conditions/diabetes/diagnosis-treatment/drc-20371451#:~:text=A%20blood%20sugar%20level%20less,mmol%2FL)%20indicates%20prediabetes.