Diabetes Disease Prediction

ARJUN BAYADEGRE PRABHANNA (230850895)

January 2024

1 Introduction

The prevalence of diabetes is rising, which emphasizes the need for quick and precise diagnostic methods to find those at risk or in the early stages of the condition. Large-scale screens may not always be feasible or accessible using traditional diagnostic techniques, despite their effectiveness. In order to determine an individual's risk of developing diabetes based on their health characteristics, it is necessary to construct trustworthy predictive models that make use of developments in machine learning and data analytics. Numerous consequences, such as kidney problems, nerve damage, and cardiovascular ailments, are linked to the disease. Diabetes must be identified early and effectively managed in order to lessen its detrimental effects on people's health and well-being.

1.1 Problem Statement

The issue statement in this case focuses on developing a reliable model for predicting diabetes using mostly logistic regression and pertinent parameters including age, blood pressure, body mass index (BMI), skin thickness, insulin levels, pregnancies, and Diabetes Pedigree Function. The objective is to create a high-accuracy model that can accurately predict if a patient has diabetes or not. By aiding in the early diagnosis and treatment of diabetes, this predictive model may enhance general health outcomes and lessen the strain on healthcare systems.

2 Domain Analysis

1. ***Pregnancies***: Some women get diabetes before they get pregnant. This condition is known as pregestational diabetes. Some women may get a particular type of diabetes that only manifests in pregnancy. This is known as gestational diabetes.During pregnancy, a woman's body may use glucose differently. This could lead to gestational diabetes or worsen pre-existing diabetes. If you had gestational diabetes throughout your pregnancy, your blood sugar returns to normal after giving delivery.

2. ***Glucose***: Your body uses glucose as an energy source. To transfer glucose from your bloodstream into the cells of your muscles, fat, and liver, where it is converted into energy by your body, your pancreas produces insulin. Diabetes patients have excessive blood sugar levels because their bodies are unable to convert glucose into fat, liver, or muscle cells for storage as energy.

3. ***Blood Pressure***: The risk of high blood pressure is twice as great in individuals with diabetes as in those without the disease. High blood sugar levels can harm your blood vessels and the nerves that support your heart's pumping action if you have diabetes. In a similar vein, elevated blood pressure can put more stress on your heart and blood arteries. The combined presence of these two disorders raises the risk of stroke and heart disease (cardiovascular disease). For those who have diabetes, blood pressure should be less than 140/80 mmHg; if you have kidney or eye disease, or any other ailment that affects blood vessels and the blood supply to the brain, it should be less than 130/80 mmHg.

4. ***Skin Thickness***: Patients with diabetes are often noted to exhibit thicker skin. Skin in affected areas may look swelling, waxy, or thickened. Although these individuals frequently have no symptoms, they may experience less discomfort and feeling. The hands and feet are most commonly affected, while other body parts might sometimes be impacted.Small blood vessels can alter as a result of diabetes. Diabetic dermopathy is a term for the skin conditions caused by these alterations. Skin diseases frequently resemble scaly, light brown spots. These patches could be round or oval in shape.

5. ***Insulin***: Your pancreas produces the hormone insulin to reduce blood glucose, or sugar. Your body either doesn't react well to insulin or your pancreas doesn't produce enough of it if you have diabetes. To maintain a healthy blood sugar range, your body needs insulin.The insulin-producing beta cells in your pancreas are harmed by diabetes. Your body is unable to create enough of this hormone as a result.

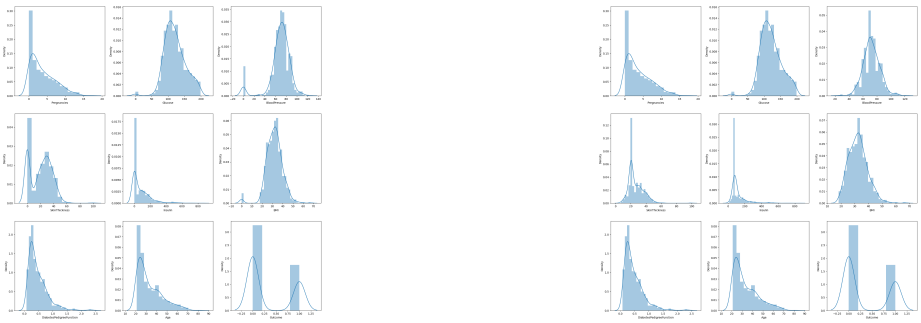
6. ***BMI***: There is a substantial correlation between insulin resistance and diabetic mass index. The development of insulin resistance is associated with elevated levels of inflammatory indicators, glycerol, hormones, cytokines, nonesterified fatty acids, and other chemicals in obese persons. The pathophysiology of diabetes is rooted in the impairment of the pancreatic β-islet cells, which results in an inability to regulate blood glucose levels. If insulin resistance coexists with the pancreatic β-islet cell failure, the development of diabetes becomes increasingly likely.

7. ***Diabetes Pedigree Function***: It provides information about a family member's history of diabetes as well as their genetic relationship to the patient. A patient with a higher pedigree function is more likely to have diabetes.

8. ***Age***: The two main causes of hyperglycemia are aging-related insulin secretion deficiencies and increasing insulin resistance brought on by sarcopaenia and changes in body composition. The aging process in humans results in disruptions to energy homeostasis and aberrant metabolism of carbohydrates. It is believed that increasing insulin resistance and a deficit in insulin secretion that develops with age are the main causes of hyperglycemia.

3 Exploratory Data Analysis

A few visual aids were created to help comprehend how different features related to the target variable. In order to detect skewness, the dataset for each column had to be shown first. Examine the distributions for indications of skewness (asymmetry) and kurtosis (tailedness). A longer right tail is indicated by positive skewness and a longer left tail by negative skewness. Kurtosis is a measure of how sharp the peak of the distribution is. The dataset has some skewness, as the graphic illustrates.



(a) Imbalanced Dataset

(b) Balanced Dataset

Figure 1: Skewness Distribution of Data for every columns

3.1 Univariate Analysis

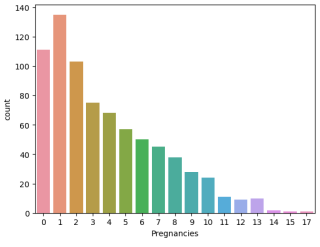
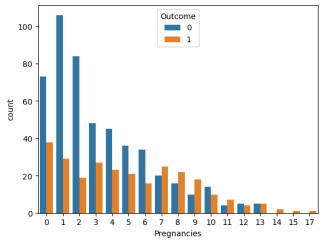


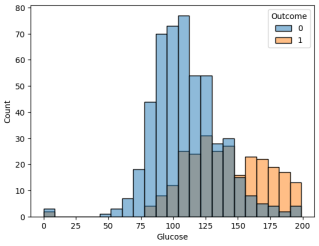
Figure 2: Visualising the patients who have conceived 0 or more times

Observations: Patients who are conceived 0 or 1 time are the highest compared to the other patients

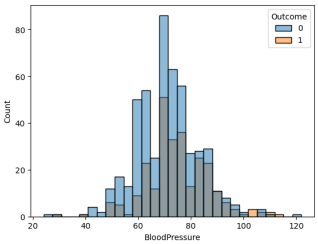
3.2 Bivariate Analysis



(a) Pregnancies impacting patients with diabetes



(b) Glucose vs Outcome



(c) BloodPressure vs Outcome

Figure 3: Visualising bivariate relationship between features and outcome

Observations:

- In figure 3(a), the patient will get diabetes if she anticipates one. However, upon examination of this data, we discovered that the likelihood of developing diabetes also rises with the number of pregnancies. All pregnant women are diabetic from the ages of 14 to 17.
- In figure 3(b), based on the available data, an individual with a glucose range of up to 100 may be deemed to have a lower risk of developing diabetes. If a person's blood sugar is between 125 and 150, they are neither at risk nor not, and more patient features need to be examined. A person's risk of developing diabetes increases if their blood glucose level is greater than 150.
- In figure 3(c), more patients have BP between 60 to 80

3.3 Data Correlation and Visualisation

The direction and strength of a relationship between variables are studied and measured using correlation. There may or may not be a relationship between

the two variables, as indicated by a correlation coefficient that is positive or lower than zero. A positive association exists between the two variables when the correlation coefficient is near +1, indicating that increases in one measure are linked to rises in the other. A negative correlation coefficient, whereby an increase in one variable is linked to a drop in the other, denotes a relationship between two variables that is near to -1.

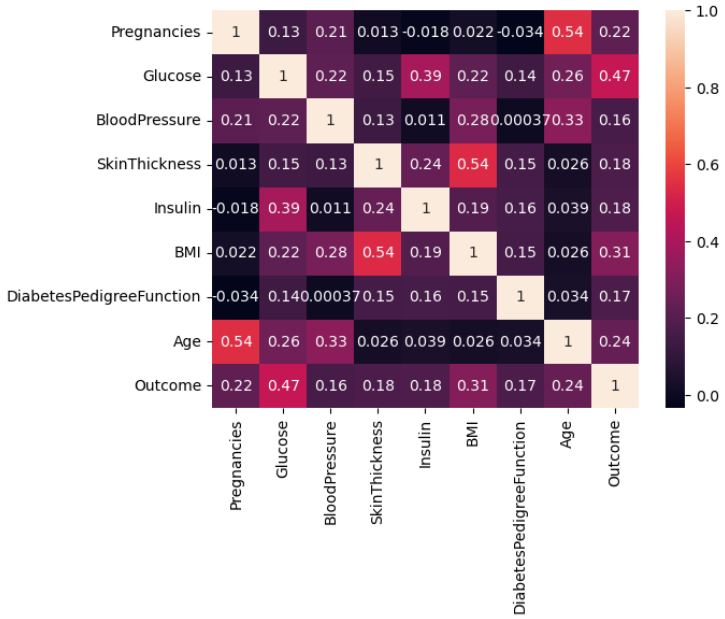


Figure 4: Visualising Data Correlation between various features and target variable

4 Methods

Algorithms Used:

1. Binary Logistic Regression Classification (Primary)
2. KNN Classification
3. Random Forest Classification

4.1 Binary Logistic Regression Classification (Primary)

The above method is typically for determining the likelihood that an instance belongs to a specific class. When the estimated probability exceeds 50%, the model indicates that the instance is either part of the positive class (designated

as '1', "Diabetic") or it is not (designated as '0', "Not Diabetic"), indicating that it belongs to the negative class.

Steps Involved:

1. ***Data Preparation:*** The target variable ('Outcome') and its features are extracted.

2. ***Data Splitting:*** The data is split into training and validation sets at random using the "data split function". In this instance, 20% of the data is used for validation and the remaining 80% is used for training.

3. ***Standardisation:*** The characteristics for the training and validation sets are standardized independently using the standardise function.

4. ***Input Matrices Preparation:*** The input matrices for the logistic regression model are created by adding a column of ones to the feature matrices using the "linear regression data function".

5. ***Gradient Descent:*** To determine the ideal weights that reduce the binary logistic regression cost function, we use the "gradient descent function". With the logistic regression gradient, the weights are updated periodically.

6. ***Model Training:*** Using the training set and an initial set of weights, the logistic regression model is trained. The procedure is done 1000 times with a learning rate of 0.01.

7. ***Model Prediction:*** The "binary prediction function" is subsequently utilized to forecast binary labels for the validation set based on the training model.

8. ***Model Evaluation:*** By contrasting the predicted labels in the validation set with the actual labels, the model's classification accuracy is calculated.

9. ***Output:*** On the validation set, the ideal weights, anticipated labels, and classification accuracy are printed.

4.2 KNN Classification

A supervised learning approach called K-nearest neighbors (KNN) is used mostly for classification but can also be utilized for regression. By calculating the distance between the test data and all of the training points, KNN attempts to predict the proper class of test data given a dataset with several classes. The k points that are closest to the test data are then chosen. Following point selection, the algorithm determines the likelihood (in the case of classification) that a test point will belong to each of the k training point classes, and the class with the highest likelihood is chosen.

Steps Involved:

1. ***Pairwise Distance Calculation:*** Euclidean pairwise distances between two sets of data points (from data and to data) are calculated using the pairwise distance calculation function (pairwise distances). It makes effective use of NumPy for efficient computation.

2. ***KNN Classification Function:*** This function conducts KNN classification given a set of testing inputs, training inputs, training outputs, and the number of neighbors (no of neighbors). Using the "pairwise distances" function that was previously defined, it determines the pairwise distances between the training and testing inputs. It returns the indices of the closest neighbors after sorting the distances. It finds the majority class among its k-nearest neighbors for each testing input, then assigns the projected label based on that determination.

3. ***Classification Accuracy Function:*** The classification accuracy is calculated using the "classification accuracy function" (classification accuracy), which takes the mean of the estimated and true labels.

4. ***K Fold Split Function:*** Divides indices into K subsets at random and uses this to do K-Fold cross-validation.

5. ***KFold Cross Validation:*** The "KFold Cross-Validation Function" (KFold cross validation knn) assesses KNN performance using K-Fold cross-validation. It repeats this process K times, training the model on K-1 folds and evaluating it on the remaining fold. It determines the overall average categorization accuracy for each fold.

6. ***Grid Search Function:*** Uses cross-validation to minimize error and does a grid search to get the ideal value of k (number of neighbors).

7. ***Model Evaluation:*** Extracts the DataFrame's target variable (y) and features (X) and divides the dataset in half, with 80% going toward training and 20% toward testing. The process also involves cross-validation and grid search to determine the ideal k value and finally the ideal k value is applied to the testing set in order to test the KNN model and assesses the model's accuracy in classification using the testing set.

8. ***Output:*** Shows the ideal k value that was found using grid search and displays the KNN model's classification accuracy on the test set.

4.3 Random Forest Classification (Extra)

A Random Forest Classifier function is initialized by Scikit library, which then trains it on the training set, makes predictions on the testing set, assesses and outputs the classification accuracy. The Random Forest model is well-known for its ensemble learning methodology, which combines several decision trees to enhance generalisation and prediction accuracy.

5 Results of the prediction tasks

The accuracy rates that various classification models predict are as follows:

1. Logistic Regression Classification - 75-80%
2. K Nearest Neighbors - 69.48%
3. Random Forest Classification - 70-75%

6 Inference

The Binary Logistic Regression model produced a higher accuracy of roughly (72-80)% on the same test set. Based on the chosen evaluation metric (accuracy), the logistic regression model performs well on the test set. It is trained using a gradient descent optimization technique to identify the optimal weights for the linear combination of features.

This shows that by using the ideal value of K, which in this case is 4, the KNN model was able to achieve an accuracy of around 69.48% on the test set. This accuracy represents the proportion of correctly predicted cases among all the instances in the test set. The ideal K, which denotes the number of closest neighbors taken into consideration when making predictions, is established by cross-validation.

Based on the same dataset, random forest comes in second best for diabetes prediction, spending more time for training data and having an accuracy range of 70–75

7 Conclusion

In conclusion, due to its higher classification accuracy, the Binary Logistic Regression model appears to perform better on this specific data set than the KNN model and Random Forest Classifier. When it comes to testing data, the Random Forest Classification model performs well, but it requires a little more time to train the dataset than the KNN model, which performs better on smaller datasets.

8 References

Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261–265). IEEE Computer Society Press.