MTH 786P Mini–Project: Disease prediction: Diabetes

Dhruv Vudayagiri

220375472

1 Introduction

Diabetes is a chronic condition that results in elevated blood sugar levels. There are two primary forms of diabetes:
Type 1, in which the immune system destroys cells that produce insulin
Type 2, in which the body either produces insufficient insulin or cells do not respond to insulin effectively.
Type 2 diabetes is significantly more prevalent than Type 1, with 90% of all adult diabetes cases in the UK being of Type 2.

1.1 Problem Statement

The objective is to construct (binary) regression/classification models utilizing any of the given variables that describe patients' characteristics, to predict their diabetes status.

2 Analysis of the dataset

2.1 Features Information

1. **Pregnancies:** describes the number of times pregnant.

2. **Glucose:** describes the glucose concentration in the body.

3. **BloodPressure:** describes the measure of pressure the heart uses to pump the whole body.

4. **SkinThickness:** denotes Triceps skin fold thickness (mm).

5. **Insulin:** measure of insulin in blood.

6. **BMI:** Body mass index.

7. **DiabetesPedigreeFunction:** specify the likelihood of diabetes, measured between 0 and 1.

8. **Age:** Measured in years.

9. **Outcome:** Indicates whether having diabetes or not [1 (Yes) or 0 (No)].

2.2 Visualizations

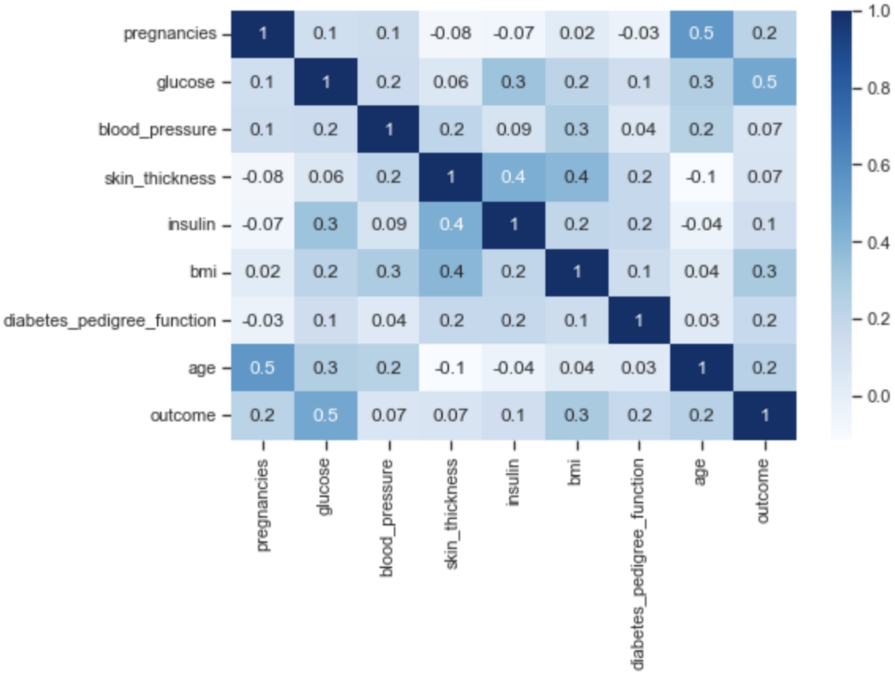

Figure 1: Correlation Plot

The above plot shows the correlation coefficients between all features in the dataset. It can help identify highly correlated features and also detect multicollinearity. Correlation coefficients range from -1 to 1, where -1 represents a perfect negative correlation, 0 represents no correlation, and 1 represents a perfect positive correlation. A correlation matrix is a symmetrical matrix, so the values in the upper triangle are the same as the values in the lower triangle. We are able to see that there is a strong correlation between glucose and outcome, along with pregnancies and age.

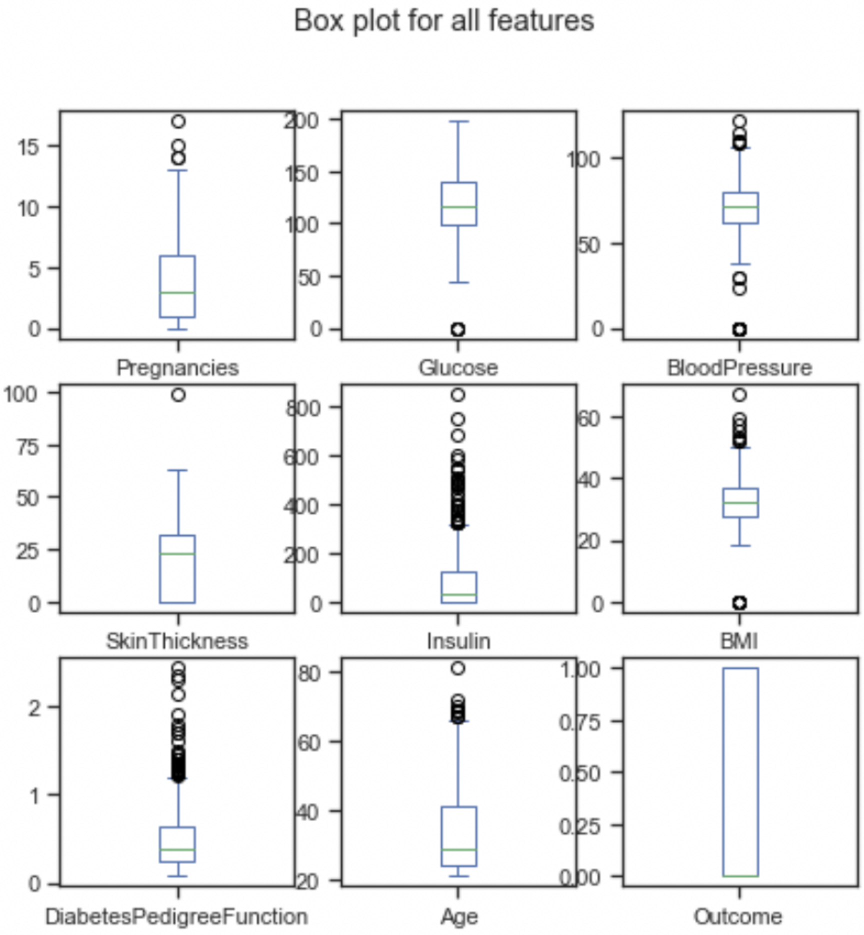

Figure 2: Box Plot for individual features

Box plots can be used to visualize the distribution of each feature in the dataset. It can help identify outliers and skewness in the data. A box plot shows the median, quartiles, and range of the data. The box represents the interquartile range (IQR) and the whiskers represent the minimum and maximum values that are within 1.5 times the IQR. Any data point outside of this range is considered an outlier. In figure 2 we can see that there many outliers for Insulin and DiabetesPedigreeFunction. On the other side, we had a negligible amount of outliers for Glucose.

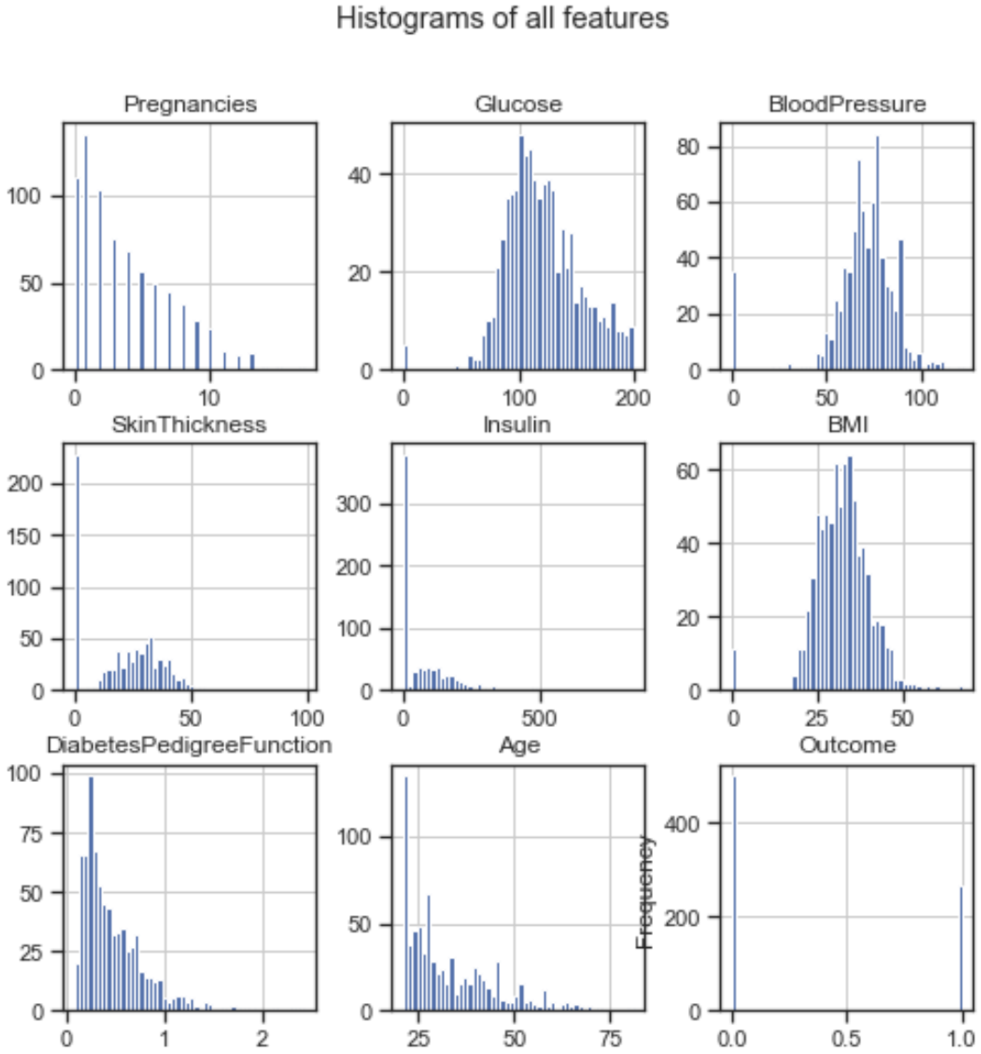

Figure 3: Box Plot for individual features

Histograms can be used to visualize the distribution of each feature in the dataset. It can help identify outliers and skewness in the data. A histogram shows the frequency of data points within a range of values. Each bar in the histogram represents a range of values and the height of the bar represents the number of data points in that range. In figure 3, we can observe that the majority of the data points are in between the range of 25 to 50 for the BMI feature. Similarly, the age feature is also in the same range.

3 Methods

3.1 Classification

Classification is the task of assigning a class label to input from a set of predefined classes. It is similar to supervised learning, but the output of the function is discrete instead of continuous. Binary classification deals with only two classes and multiclass classification deals with more than two classes. The goal is to find a function f that accurately maps inputs to their corresponding class labels.

3.1.1 Nearest neighbour classification

A simple method of classification is to assign a new data sample x to a class based on the class labels of its K-nearest neighbors in the training set. The class label of the unknown sample is determined by assigning a probability to it, based on the labels of the K nearest neighbors. The probability calculation is of the form:

$$\rho(y = c|x, K) := \frac{1}{K} \sum_{l \in N_K(x)} l(y_l = c) \tag{1}$$

where

$$l(z) := \begin{cases} 1 & \text{if z is true} \\ 0 & \text{if z is false} \end{cases} \tag{2}$$

In the method for classification called k-nearest neighbors, where a new data sample is classified based on the labels of its k-nearest neighbors in the training set. The neighborhood is defined as the set of k samples closest to the new sample, and the distances between samples are typically measured using Euclidean distance. The probability of a new sample belonging to a particular class is computed by averaging the labels of the k nearest neighbors with that class label. The class label with the highest probability is then assigned as the output of the classifier.

$$f(x) := \arg \max_{c \in C_0, C_1, \ldots, C_n} \rho(y = c|x, K) \tag{3}$$

The K-nearest neighbours classification method uses the labels of the K nearest neighbours of a new data sample x to assign a probability to the unknown output label. The number of neighbors, K, is a hyperparameter that is chosen, using techniques such as cross-validation.

3.1.2 Logistic Regression

Using mean-squared-error regression to solve classification problems may not be effective in practice, as it does not relate to minimizing misclassification.

Transforming the prediction $f(x,w)$ into a probability using the logistic function $\sigma(f(x,w))$ is a more appropriate approach for binary classification. Where $\sigma : (-\infty, \infty) \to [0,1]$ is the so-called logistic function

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{4}$$

We can then assign probabilities to the events of the output $f(x,w)$ belonging to either class with class label zero or class label one.

$$\rho(1|x) := \sigma(f(x,w)), \ \rho(0|x) := 1 - \sigma(f(x,w)). \tag{5}$$

From the definition of σ, we can see that $\rho(1|x) \geq 0$, $\rho(0|x) \geq 0$ and $\rho(1|x) + \rho(0|x) = 1$ which means that we can consider ρ as a probability. Now assuming we have a set of s pairs of samples (x_i, y_i) with y_i being either 0 or 1, that are independent and identically distributed and follow the probability density function defined in (4). The likelihood function can be determined from this.

$$\rho(y|X,w) = \prod_{i=1}^{s} \rho(y_i|x_i), \tag{6}$$

where X and y are shorthand for the matrix and vector that are obtained from the samples $\{y_i\}_{i=1}^{s}$ and $\{x_i\}_{i=1}^{s}$. It is worth noting that equation (5) can also be rewritten as

$$\rho(y|X,w) = \prod_{i=1}^{s} \sigma(f(x_i,w))^{y_i}(1 - \sigma(f(x_i,w)))^{1-y_i} \tag{7}$$

In logistic regression, the likelihood is calculated as the product of the individual probabilities of each sample belonging to a certain class. The parameters that maximize the likelihood can be obtained by minimizing the negative log likelihood. This can be represented as a logistic function of the model function f(x, w) and is known as logistic regression. Any model function can be used as long as it maps to real numbers, linear models concerning weights w are preferred for unique minimization.

Binary logistic regression is a method that focuses on polynomial data models, $f(x,w) = \langle \phi(x), w \rangle$. It uses the logistic function and gradient descent to find the optimal solution. The objective function is differentiable and the gradient can be computed. The gradient descent iterate is updated as,

$$w^{k+1} = w^k - \tau \Phi(X)^T (\sigma(\Phi(X)w^k) - y)$$

The logistic regression function is convex and the gradient descent will converge to a solution.

4 Results

4.1 Binary Logistic regression

Features	Training Set	Validation Set
Glucose	74.27%	75.32%
Glucose, BMI	76.06%	79.87%
Glucose, BMI, Age	77.36%	74.68%
Glucose, BMI, Age, Pregnancies	76.55%	77.92%
Glucose, BMI, Age, Pregnancies, DiabetesPedigreeFunction	76.06%	83.12%
Glucose, BMI, Age, Pregnancies, DiabetesPedigreeFunction, SkinThickness	77.85%	75.97%
Glucose, BMI, Age, Pregnancies, DiabetesPedigreeFunction, SkinThickness, Insulin	77.85%	75.32%
Glucose, BMI, Age, Pregnancies, DiabetesPedigreeFunction, SkinThickness, Insulin, BloodPressure	78.50%	75.32%

Table 1: Binary Logistic Regression Classification Accuracy

The important features can be determined by looking at the classification accuracy for each set of features. In general, the features that result in a higher classification accuracy are considered more important.

In the above table, we can see that the classification accuracy increases as more features are added to the model.When the feature set includes ”Glucose, BMI, Age, Pregnancies, DiabetesPedigreeFunction”, the accuracy of the validation set is 83.12%, but, when we add ”SkinThickness” the accuracy drops to 75.97%. The accuracy remains the same, 75.32%, when adding ”Insulin” and ”BloodPressure”.

In this case, it can be inferred that the most important features are ”Glucose”, ”BMI”, ”Age”, ”Pregnancies” and ”DiabetesPedigreeFunction” as they result in the highest validation set accuracy.

It’s important to note that, the accuracy figures may not be directly comparable, as the training and validation sets may have different sizes, distributions, etc.

4.2 K-nearest neighbours

Features	Accuracy	Optimal neighbours
Glucose	73.58%	20
Glucose, BMI	72.93%	15
Glucose, BMI, Age	75.14%	14
Glucose, BMI, Age, Pregnancies	75.79%	22
Glucose, BMI, Age, Pregnancies, DiabetesPedigreeFunction	75.91%	27
Glucose, BMI, Age, Pregnancies, DiabetesPedigreeFunction, SkinThickness	75.26%	15
Glucose, BMI, Age, Pregnancies, DiabetesPedigreeFunction, SkinThickness, Insulin	74.48%	15
Glucose, BMI, Age, Pregnancies, DiabetesPedigreeFunction, SkinThickness, Insulin, BloodPressure	73.44%	13

Table 2: KNN Classification Accuracy

The table 2 shows the classification accuracy for different sets of features. The first column lists the features being used, the second column shows the corresponding accuracy as a percentage and the third column shows the number of optimal neighbours to be considered. The table shows that as more features are added to the model, the accuracy generally increases, with the highest accuracy of 75.91% achieved when using the features ”Glucose, BMI, Age, Pregnancies, DiabetesPedigreeFunction". However, when adding the last two features "SkinThickness, Insulin, BloodPressure", the accuracy drops slightly to 73.44%. The table also highlights that the accuracy can fluctuate when adding different features, in particular, when adding "SkinThickness, Insulin, BloodPressure" the accuracy drops.

4.3 Comparision between model results

In both tables, the accuracy increases as more features are added to the model, with the highest accuracy achieved when using the features "Glucose, BMI, Age, Pregnancies, DiabetesPedigreeFunction". However, the highest accuracy in the first table is 78.50% while in the second table is 75.91%.

The second table has a more consistent performance across different sets of features, while the first table has a fluctuation in performance when adding different features. In the first table we can see that when adding the last two features "SkinThickness, Insulin, BloodPressure" the accuracy drops.

The first table shows that the accuracy for the training set is generally higher than that for the validation set, this is expected as the model can fit the training data better than the unseen validation data.

Both tables show that the accuracy fluctuates when adding different features, for example, in the second table when adding "SkinThickness, Insulin, BloodPressure" the accuracy drops from 75.91% to 73.44%.

In both tables, the accuracy fluctuates when adding different features, this may suggest that some of the features are more important or informative than others, or that some features may be redundant or even detrimental to the performance of the model.

5 Conclusion

The best performance is achieved when using the feature set "Glucose, BMI, Age, Pregnancies, DiabetesPedigreeFunction" with an accuracy of 75.91% in the second table. This feature set has the highest accuracy across both tables. However, it's important to note that the performance of a model can depend on multiple factors and accuracy fluctuates when adding different features. In summary, the problem tackled is the classification of diabetes in a dataset using different sets of features. The results found, show that the performance of the model is affected by the number and type of features.

6 References

@miscnhs_diabetes, title = *Diabetes*, url = `https://www.nhs.uk/conditions/diabetes/`, publisher = National Health Service (NHS), year = 2021, note = Accessed: January 24, 2023
@miscqmul_2047820, title = Machine Learning, author = "Martin Benning", url = `https://qmplus.qmul.ac.uk/mod/resource/view.php?id=2047820`, publisher = Queen Mary University of London, year = 2022, note = Accessed: January 24, 2023
@miscqmul_2126364, title = "Coursework9-Solutionsipynb", url = `https://qmplus.qmul.ac.uk/mod/resource/view.php?id=2126364`, publisher = Queen Mary University of London, year = 2022, note = Accessed: January 24, 2023
@inproceedingseverhart1988using, title=Using the ADAP learning algorithm to forecast the onset of diabetes mellitus, author=Everhart, JW and Dickson, JE and Knowler, WC and Johannes, RS, booktitle=Proceedings of the Symposium on Computer Applications and Medical Care, pages=261–265, year=1988, organization=IEEE Computer Society Press