# THE REPORT IS ON DIABETES IN INDIVIDUALS

# Machine Learning

# Final Exam – Saipriya Gourineni

# INTRODUCTION:

This study is based on information on Diabetic Individuals. For the National Institute of Diabetes and Digestive and Kidney Diseases, Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. gathered this data using Kaggle and made it public. Based on specific diagnostic metrics present in the dataset, the dataset's goal is to diagnostically forecast whether a patient has diabetes or not. These instances were chosen from a bigger database under several restrictions. Particularly, all patients at this facility are Pima Indian women who are at least 21 years old.

The datasets consist of several medical predictor (independent) variables and one target (dependent) variable, Outcome. Independent variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on. Here the features of the dataset are being used to predict the outcome which is whether the patient is diabetic or not. Since I am using K Nearest Neighbor algorithm to predict the outcome, it is essential to find an ideal K value in order to improve the performance of the model

A value of 0 in the columns like Glucose, Blood pressure, Skin thickness, Insulin and BMI indicates a missing value since it is illogical. Since counting them thereafter would be simpler and zeros need to be replaced with appropriate values, so Initially, the 0’s were replaced with Nan in the dataset.

The following image displays the histograms of the aforementioned columns with Null values. In order to replace these values with those that are more appropriate for the dataset, it is necessary to understand the data. Based on the histogram's skewness, Null values can be replaced with the median or mean.

Diagram

Description automatically generated

Because of the right skewness in the graph, I have replaced the column’s skin thickness and Insulin with the median based on the histogram data above. Since the remaining columns' data are roughly normally distributed, the Null values were replaced with their corresponding means.

Chart

Description automatically generated

The dataset's outcome field is examined and represented using a bar graph; 500 records have 1 as the outcome, whereas 268 records have 0 as the outcome.

Table

Description automatically generated with low confidence

To determine the relationship between the dataset's properties, I drew a correlation matrix of the cleaned data. Using this correlation table, feature selection can be improved to improve model’s performance.

Applying distance-based algorithms like KNN requires that all characteristics be scaled to the same value. It could be inferred that the larger feature will totally eclipse or dim the smaller feature, which will affect the performance of any distance-based model because it will assign more weight to variables with larger magnitudes. Consequently, I have scaled all the features using the r scale function.

To create a model that performs better, the best K value for KNN must be found. The count of the nearest neighbors is shown by the k value. Distances between test points and training label points must be calculated. Because it is computationally expensive to update distance measures after each iteration, KNN is a slow learning algorithm. The square root of the total number of samples is the optimal K value that is typically discovered. But in this case, I performed a trial-and-error approach for k=7,9,11,13, etc. values and discovered that KNN performs better at k=9.

Chart

Description automatically generated

A classification problem's prediction outcomes are compiled in a confusion matrix. Count values are used to describe the number of accurate and inaccurate predictions for each class. This is the confusion matrix's secret. The confusion matrix demonstrates the way your classification model makes predictions while being confused. It provides you with information on the types of errors being produced, which is more essential than just the faults your classifier is making. This breakdown gets over the drawback of relying solely on categorization accuracy. The dataset's confusion matrix is shown above, along with the corresponding TP, FP, FN, and FP values.

Chart, scatter chart, bubble chart

Description automatically generated

The KNN model was tested for the different number of K neighbors which are to be checked for a given point. At K = 9 neighbors, the model is performing better with an accuracy of 81% on the test data. The confusion matrix of the model illustrates that the false positives, as well as false negatives, are relatively low. Hence, it is accurate.