**DIABETES DATASET**

Name: Nawfal Khan

Student Number: 23025098

GitHub Repository:

The Pima Indians Diabetes Database is extensively used for predicting diabetes via iagnostic measures, encompassing 768 records and 9 distinct columns. It contains various medical predictor variables alongside a single target variable named "Outcome." These predictors cover a broad spectrum of health metrics, including the number of pregnancies a patient has undergone, their Body Mass Index (BMI), insulin levels, and age, among others.

The database encompasses comprehensive records from 768 female individuals of Pima Indian descent, each 21 years of age or older. It documents vital health indicators through nine columns: the count of pregnancies, the concentration of plasma glucose after a two-hour oral glucose tolerance test, the diastolic blood pressure in millimetres of mercury, the triceps skin fold thickness in millimetres, the two-hour serum insulin in microunits per millilitre, the body mass index calculated from the weight in kilograms divided by the square of the height in meters, the Diabetes Pedigree Function which evaluates the probability of diabetes based on family history, the chronological age in years, and the Outcome—a dichotomous variable signifying whether a diabetes diagnosis was confirmed (1) or absent (0). This detailed assembly of health information is crucial for devising predictive models to gauge diabetes risk within this specific population groupTop of Form

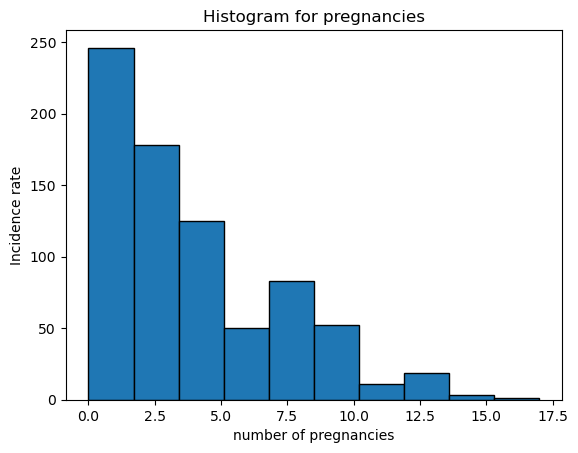


Figure 1. Histogram

The histogram offers a graphical depiction of the range of pregnancy frequencies among the participants in the dataset, aimed at examining the prevalence and variation of pregnancy instances recorded.

Displayed here, the histogram charts the distribution of pregnancy occurrences, with each bin marked and enumerated. It is important to note that the y-axis reflects the incidence rate, and the x-axis represents the tally of pregnancies.

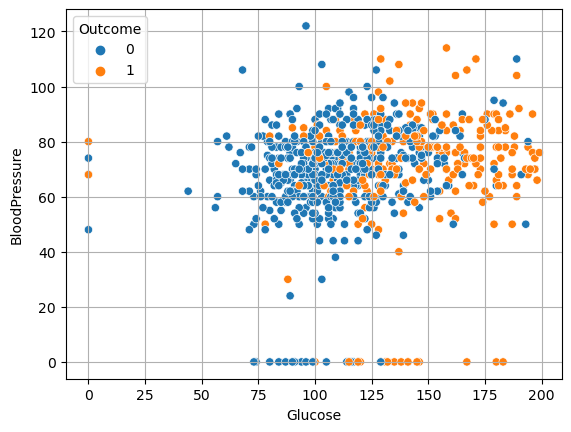


Figure 2. Scatter Plot

The scatter plot is designed to illustrate the relationship between blood glucose and blood pressure levels. On the x-axis, we have glucose values obtained from an oral glucose tolerance test, and on the y-axis, there are diastolic blood pressure readings, denoted in millimetres of mercury. An upward trend in glucose may be associated with an increase in blood pressure, indicating a potential risk marker for diabetes.

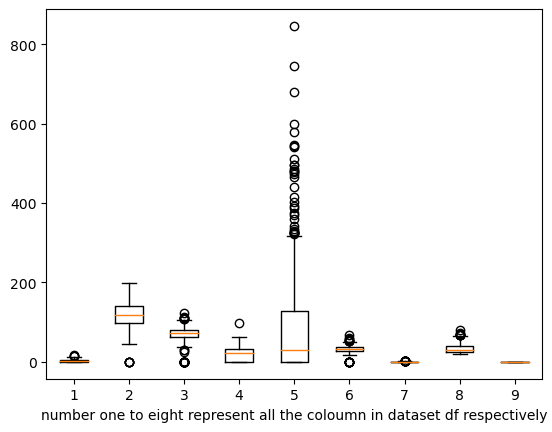


Figure 3. Box PLOT

The image presents a box plot which conveys essential statistical benchmarks. The lowest value in the data, barring any anomalies, is depicted at the left whisker's terminus. The first quartile (Q1) identifies the bottom 25 percent of the dataset. The median (Q2), situated in the plot's center, bisects the dataset, denoting the central value. The third quartile (Q3), signifying the 75th percentile, demarcates the top quarter of the data. At the extreme of the right whisker lies the maximum value, discounting any outliers. The whiskers extend from each quartile to the most extreme data points within a reasonable range, illustrating the data's spread. The minimum data points, discounting outliers, are located at the tips of the left whiskers, while the whiskers themselves span across values that fall outside the interquartile range, representing the central 50% of the dataset

Top of Form

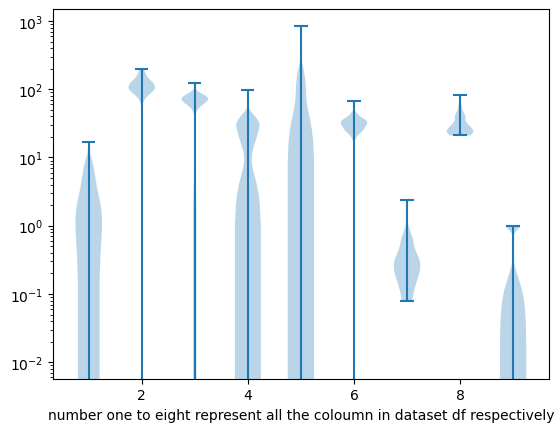


Figure 4. Violin plot

The violin plot visualization derived from the Pima Indians Diabetes Database yields critical insights into the spread and concentration of key health indicators throughout the dataset. Such insights are vital for researchers and healthcare professionals who specialize in predicting and managing diabetes within the Pima Indian population. Glucose and BMI are likely to display significant variation, underscoring their correlation with diabetes risk. By analyzing the contours of these distributions, we can identify common glucose and BMI ranges, establishing potential risk benchmarks. Similarly, the Blood Pressure and Insulin Levels plots might show a propensity for elevated readings, crucial for diabetes diagnostics. Furthermore, examining the distribution across different ages and the number of pregnancies can illuminate demographic patterns and their implications for diabetes prevalence. Overall, the violin plot visualization offers invaluable perspectives on how these health metrics vary among individuals, enhancing our comprehension of their roles in diabetes risk within this community.

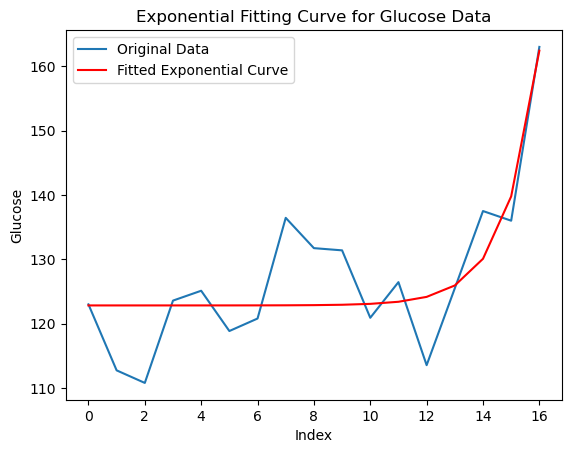


Figure 5. Exponential Fitting

This report details the findings from an exponential fitting curve analysis performed on the glucose data from the Pima Indians Diabetes Database, accessible on Kaggle. The primary aim of this study is to investigate the behavior of glucose concentrations within the dataset and to discern any patterns that might predict outcomes related to diabetes. The dataset encompasses health data collected from female members of the Pima Indian community, including critical diabetes diagnostic markers such as glucose levels, blood pressure, BMI, and insulin levels, with a specific focus on glucose as a key diagnostic indicator.

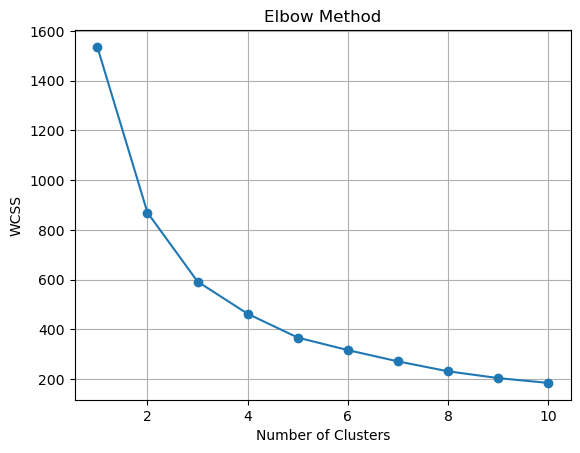


Figure 6. Elbow method

The image displays a graph illustrating the Elbow Method, a technique used to determine the optimal number of clusters in k-means clustering. This method involves plotting the within-cluster sum of squares (WCSS) against the number of clusters (k). The WCSS is a measure of variance within each cluster, and the goal is to minimize this value, which indicates a better-defined cluster

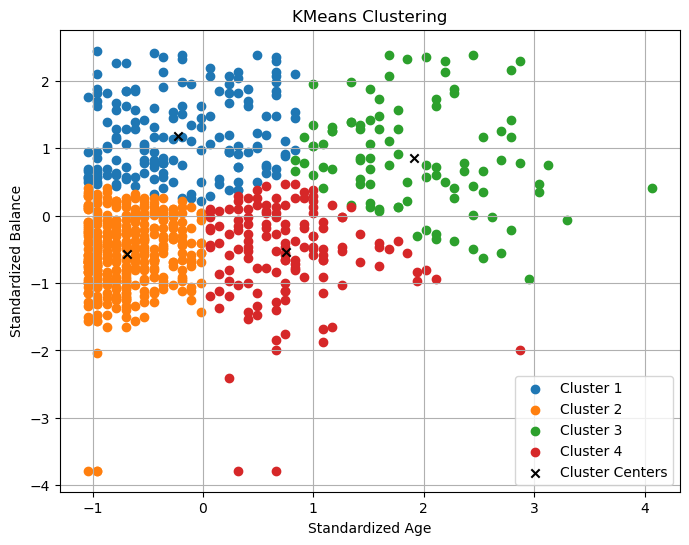


Figure 7. KMeans Clustering

The scatter plot demonstrates the categorization of the dataset into four distinct clusters executed by the KMeans clustering algorithm. It utilizes a pair of standardized variables, which possibly correspond to individual characteristics like 'Standardized Age' and 'Standardized Balance.' The plot reveals a conspicuous segregation of clusters, delineated along these two axes. Such a demarcation suggests effective clustering, as it implies that entities within each cluster share greater similarity with one another than with those in different clusters. The KMeans algorithm's proficient performance in dividing the data into four separate and well-defined clusters affirms the relevance and discriminatory power of the selected standardized features for the purpose of data segmentation.

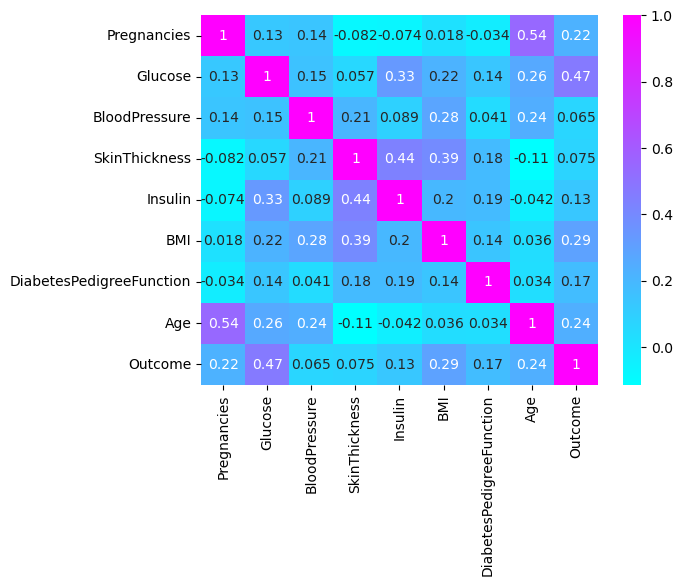


Figure 8: Heatmap

The heatmap is a visual representation of a correlation matrix, detailing the relationships between various diabetes-related variables in a dataset. Key variables include Pregnancy, Glucose, blood pressure, skin thickness, Insulin, BMI, DiabetesPedigreeFunction, Age, and the Outcome, which indicates the presence of diabetes. Notably, Glucose has a pronounced positive correlation with the Outcome (0.47), underscoring its importance as a diabetes indicator. There's a notable positive link between Age and Pregnancies (0.54), suggesting a trend of higher pregnancy counts in older participants. Conversely, Blood Pressure and Skin Thickness have minimal correlations with the Outcome, hinting at their limited individual predictive value

**Statistical Methods:**

**Mean() :**

• Mean is calculated the average value. The average number of pregnancies is 3.85.

• The average plasma glucose concentration 2 hours in an oral glucose tolerance test is approximately 120.89 mg/dL.

• The average diastolic blood pressure is 69.11 mm Hg.

**Median() :**

Half of the participants have been pregnant more than 3 times, while the other half less.

Median blood pressure is 72 mm Hg.

Median skin thickness is 23 mm

**Skewness:**

Age and Diabetes Pedigree Function are positively skewed, indicating a distribution with a longer tail to the right.

Pregnancies and Insulin levels are also positively skewed.

**Kurtosis:**

Pregnancies and Insulin levels show leptokurtic kurtosis, suggesting heavy tails and more outliers.

**Describe ():**

F function provides a comprehensive view of the dispersion and central tendency across all measured variables.

**Corr ():**

There is a noticeable correlation between age and the number of pregnancies.

BMI shows a mild positive correlation with skin thickness and insulin levels.