The Report on Clustering and Regression to Identify Health Risk Patterns

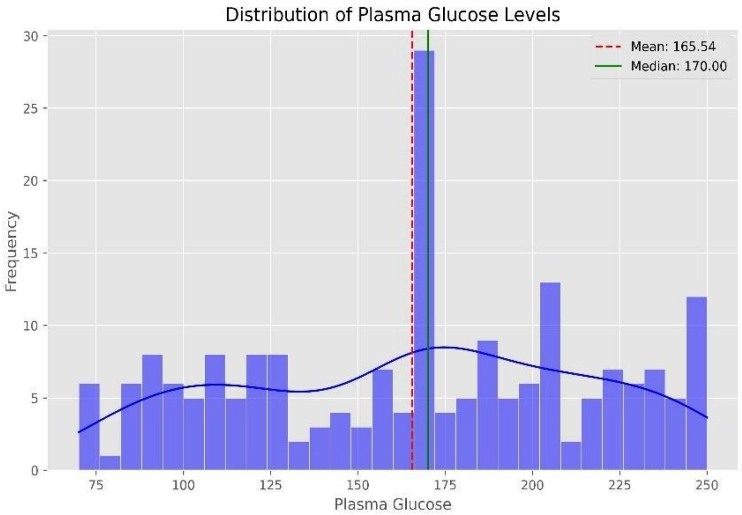
Name : Vaishnav Rajith Kalathil

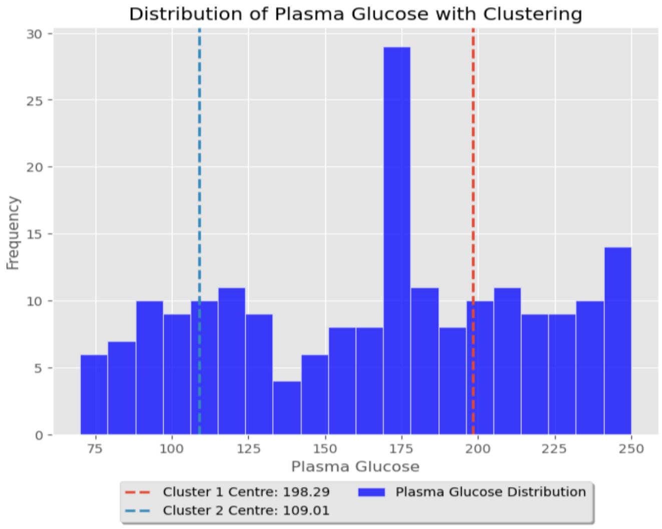
Repository Link: https://github.com/zk24aao/Clustering-and-Fitting---Applied-Data-Science.git

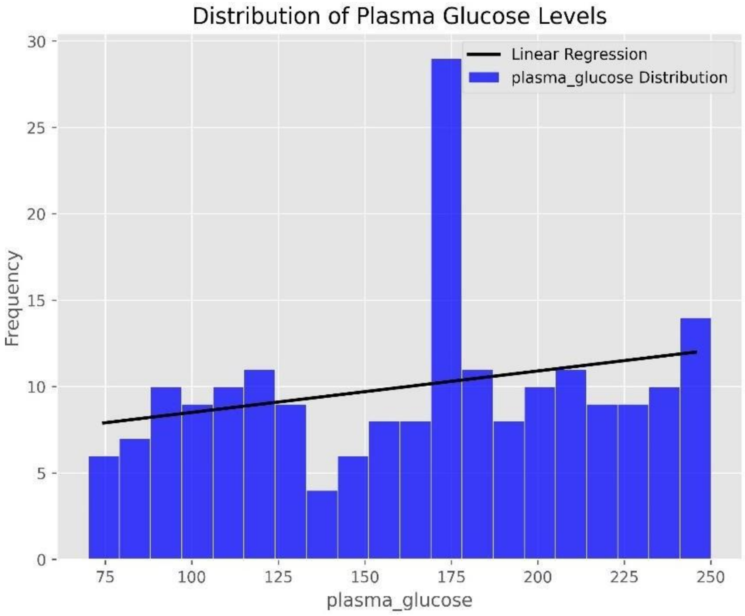
***Introduction*:**

This report consist of Clustering and Fitting function with the use of Kmeans and Linear Regression on the dataset contains information about patients such as Age, Blood pressure ,Cholesterol, Plasma Glucose , Body Mass Index etc. Through different visualization such as histogram, scatter plots, elbow plot. This report explores the application of clustering and regression techniques to identify health risk patterns among patients, with the aim of better understanding the relationships between various factors and health conditions.

Distribution Of Plasma Glucose level :







A Closer Look at **Plasma Glucose Levels** As we examine the data, we notice a concerning trend. The regression line shows a slight upward slope, indicating that higher plasma glucose levels are becoming more frequent. This could mean that a significant number of individuals are struggling with elevated glucose levels, which is alarming from a medical perspective. The High-Risk Zone

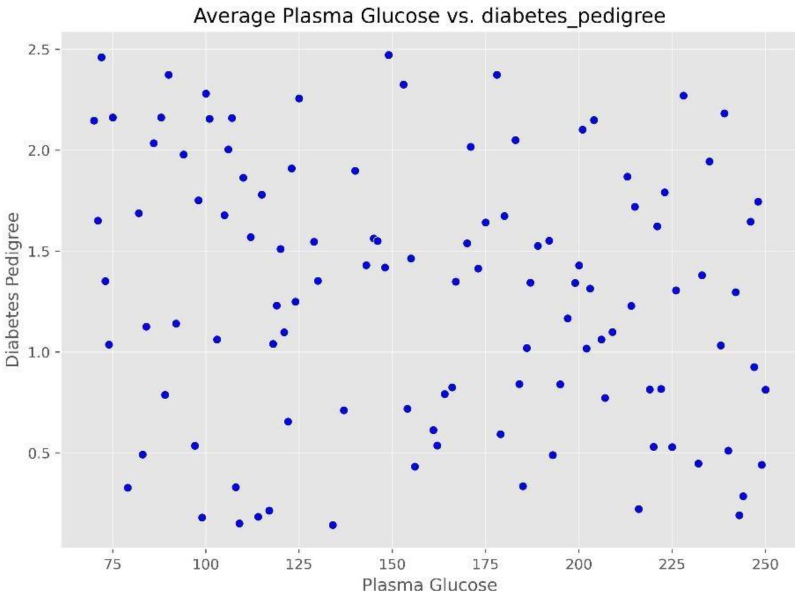
The right side of the histogram reveals a right-skewed pattern. The bars representing glucose levels between 200-250 indicate individuals with high blood sugar. High glucose levels can lead to serious health issues.

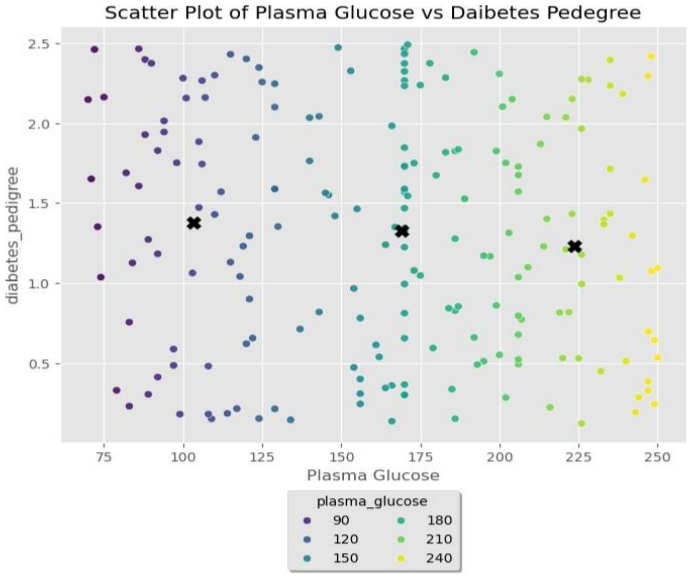
On the other hand, the left side of the graph shows individuals with normal plasma glucose levels, typically below 110. This reassuring sign indicates that these individuals have healthy control.

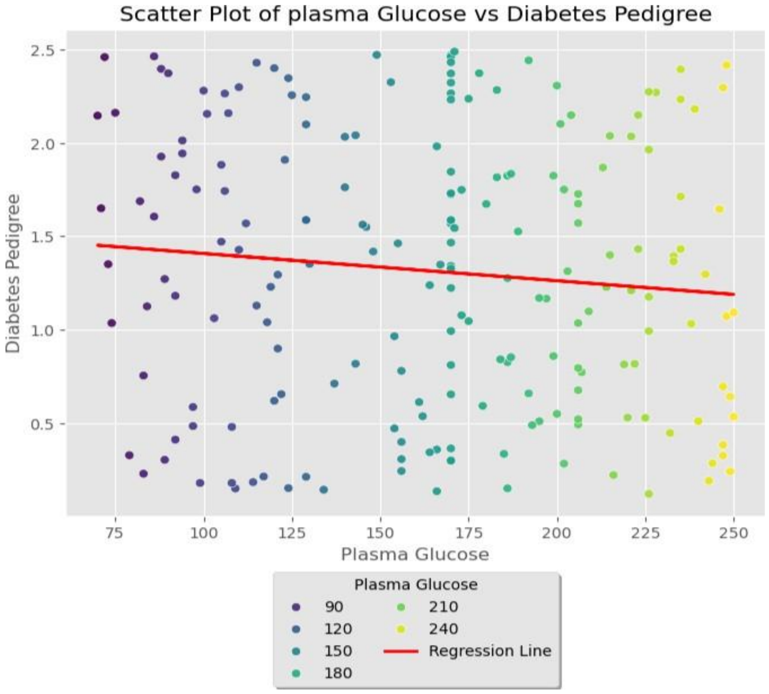
Predicting Future Health Risks:

If this trend continues, it may signal a growing risk of diabetes within the population. By incorporating additional factors like age, BMI, or insulin levels, we can refine our predictions and better estimate future health risks. This will enable us to take proactive steps to prevent or manage diabetes.

<u>Relation between Plasma Glucose and Diabetes Pedigree:</u>



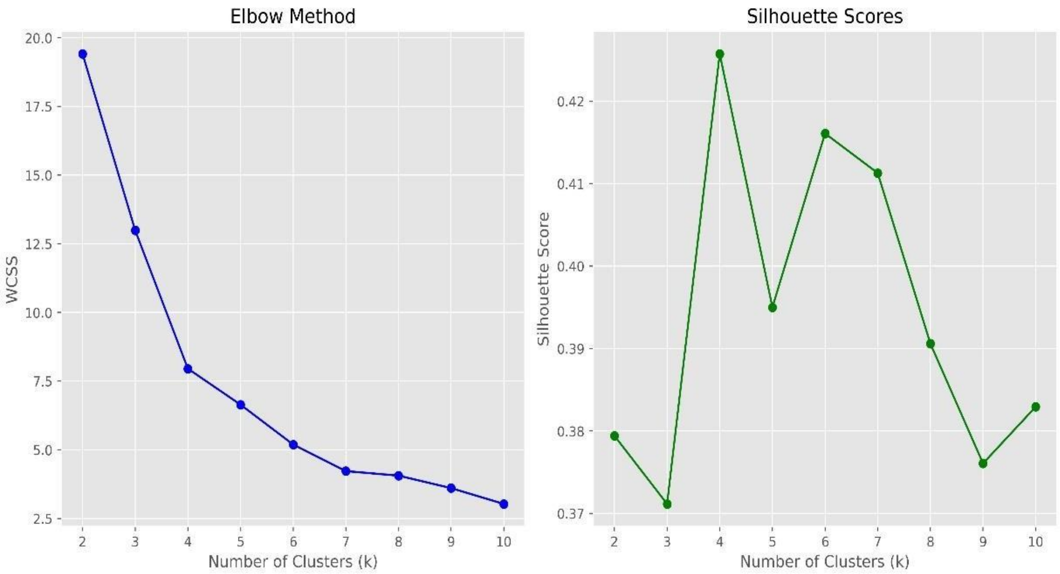




This relationship could help us understand how our biology and lifestyle choices interact in the context of diabetes risk

By examining this, we can gain insights into how genetics might influence our health and the importance of monitoring glucose levels for prevention. **<u>Cluster 1</u>**: Low plasma glucose levels with varying diabetes pedigree scores. **<u>Cluster 2</u>**: Moderate plasma glucose levels. **<u>Cluster 3</u>**: High plasma glucose levels. High blood sugar levels can be a sign of diabetes. Individuals in the high-glucose level should be monitored closely. The scatter points are colour-based on plasma glucose levels, with darker colours representing lower glucose levels and lighter colours representing higher levels. The diabetes pedigree function shows how likely a person is to develop diabetes based on their family history and genetics. In simple terms, linear regression helps us see if people with higher blood sugar levels (plasma glucose) are also more likely to have a higher genetic risk of diabetes.In simple terms, people with higher blood sugar levels might have a greater risk of developing diabetes, especially if they also have a strong family history of the condition.

Elbow Plot & Silhouette Score : (Plasma Glucose and BMI)



The code creates a powerful visualization tool with two side-by-side plots. These plots work together to help us find the spot – the optimal number of clusters – in our data.

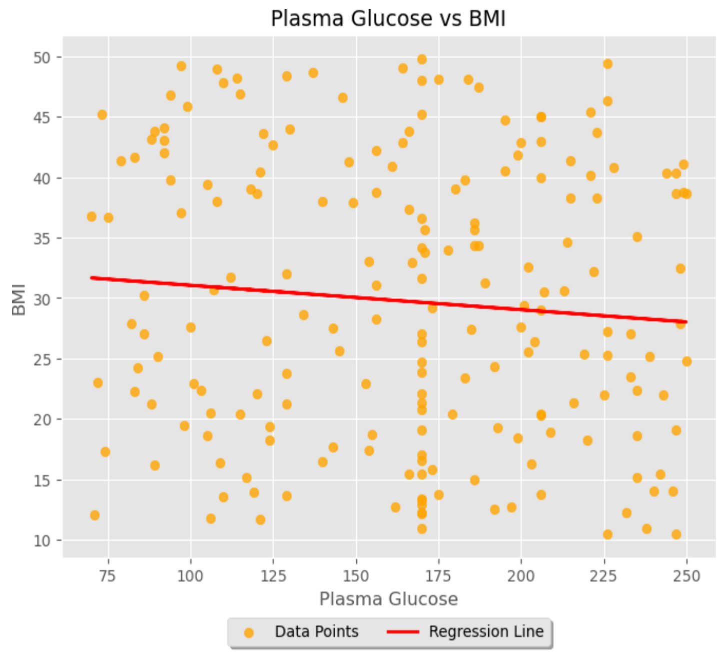
Plot 1: The Elbow Method

It gives us a visual to find the best number of clusters.

Plot 2: Silhouette Score

Which provides a more precise measure of how well our clusters are formed. A high score means our clusters are well-separated and cohesive. By examining both plots together, we can gain a deeper understanding of our data's clustering structure. There is a positive correlation between plasma glucose levels and BMI. In other words, as BMI increases, plasma glucose levels tend to increase as well.

Linear Regression (Plasma Glucose & BMI):



The red regression line represents the best-fit linear relationship between plasma glucose and BMI. This line indicates how BMI tends to change as plasma glucose levels increase. the line slopes downward in direction, it means that as plasma glucose levels decreases, BMI also tends to decrease as per the given data.

Conclusion:

Through this analysis we came to know about the behaviour of patient’s data. Firstly, from the distribution of plasma glucose level we know that few outliers is high and rest all have same frequency that is low. Secondly the relation between plasma glucose and diabetes pedigree ,the diabetes pedigree function shows how likely a person is to develop diabetes based on their family history and genetics. In these report we can see how plasma glucose has slightly negative relations with BMI as well diabetes pedigree. That means when plasma glucose increases diabetes pedigree and bmi decreases. Additionally histogram has an increasing trends which means the chance of having blood sugar level is higher.