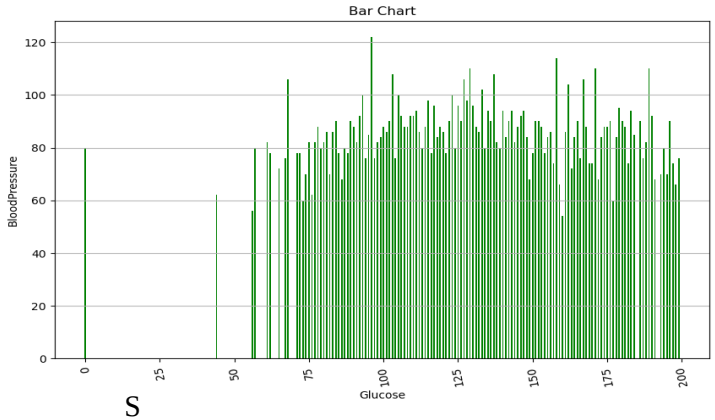
Report On Analysis of Clustering and Fitting

Author: Somraj Bharadwaj Cheppela
Student Id: 23032481
GitHub Repository:
https://github.com/SomrajBharadwaj/clustering-and-fitting-assignment.git

1. **Introduction**
   We will use the Diabetes dataset, which was obtained from Kaggle, for an exploratory data analysis (EDA) in this study. Python libraries like matplotlib, seaborn, and pandas will be used. Understanding the correlations between various factors, exploring and analyzing the dataset, and obtaining useful analysis are our goals.
2. **Dataset Overview**
   The information contained in the dataset "diabetes.csv" that follows pertains to the circumstances in which a person's likelihood of having diabetes is determined, regardless of their actual diabetes status. Though it's possible that none of the columns causes diabetes—the majority of them reveal accurate conceptions all of them might not.

3. **Exploratory Data Analysis:**
   In order to obtain correct results, many analyses of this dataset have been coded using algorithms. The whole patient record whether or not they have diabetes is displayed in the file. It explains what kinds of documents should be submitted in order to verify that the patient has diabetes. Also used the correlation function to find the relation between the factors which helped to offer deeper insights into the probability of diabetes.

4. **Visualization of Graph:**
   - **Bar Chart**
     Generally, millimeters of mercury (mmHg) are used to measure blood pressure. Less than 120/80 mmHg is the normal blood pressure measurement. Blood pressure measurements on the graph are probably not typical because the blood pressure scale only ranges from 0 to 200. Milligrams per deciliter, or mg/dL, is the standard unit of measure for blood glucose. The typical range for normal blood sugar levels is 90 mg/dL, though this might vary based on several factors. The normal range of the blood glucose axis is shown by the scale, which runs from 0 to 200.

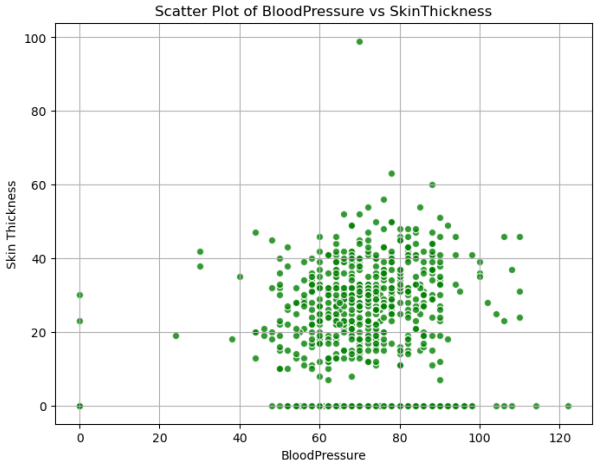


S

Fig 1: Bar graph of Glucose vs Blood Pressure

- **Scatter Plot**

  Blood pressure and skin thickness do not clearly correlate linearly, according to the scatter plot. On the other hand, the lower left corner clustering implies that decreased skin thickness and lower blood pressure may be related. Skin thickness variability is implied by varied data points across blood pressure levels. Since



  correlation does not imply causality, more research is necessary. Results may vary depending on sample size, demography, and health. In-depth investigation is necessary to identify plausible causative connections between skin thickness and blood pressure.

Fig 2: Scatter plot of Blood Pressure vs Skin Thickness

- **Heatmap plot**

  The relationships between different characteristics in a diabetes-related dataset are displayed visually in the heatmap. Pregnancies and age, insulin and glucose, skin thickness and insulin, and BMI and skin thickness are among the strong relationships. To establish causal linkages, more research is required as correlation does not automatically imply causation. To further emphasize the necessity for thorough research and interpretation when deriving conclusions from the heatmap, information on the dataset and the distribution of each characteristic is also missing.

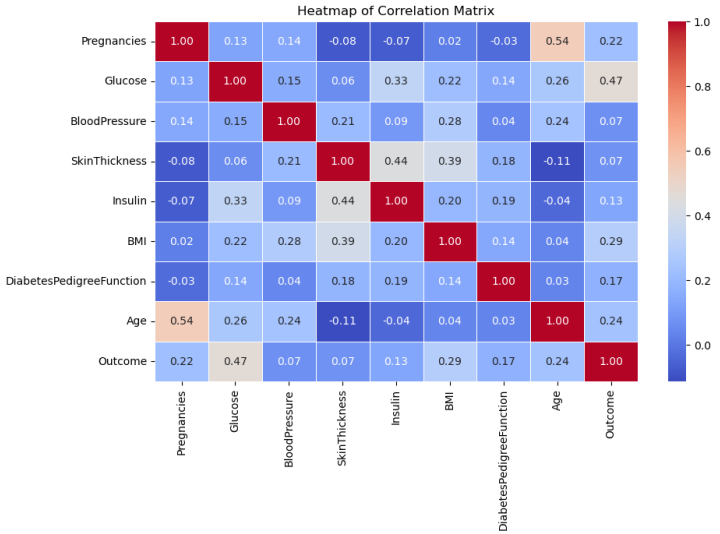
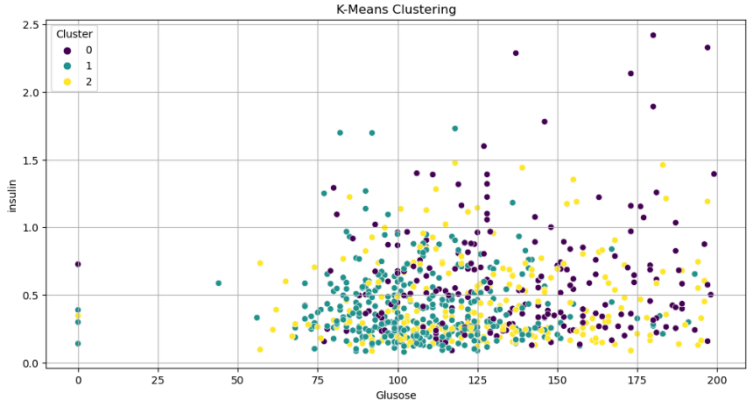


Fig 3: Heatmap of Correlation Matrix

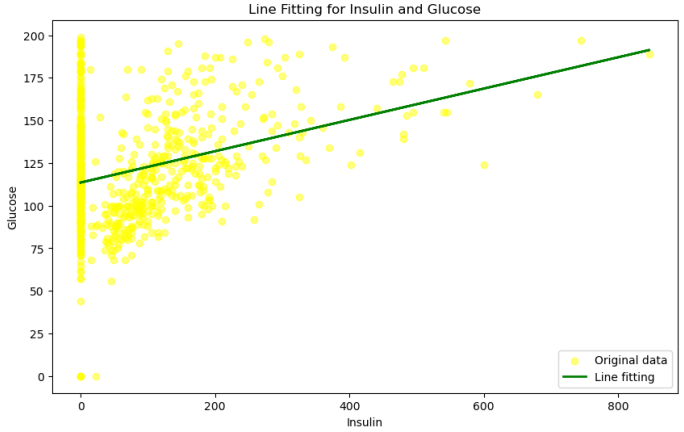
- **K-Means Clustering**



The scatter figure displays the data point distribution according to glucose and diabetes pedigree function variables, illuminating the clustering outcomes of the K-means algorithm. The use of unique colors for each cluster makes it easier to spot patterns or groupings within the collection. In addition, the elbow plot shows the within-cluster sum of squares (WCSS) for various cluster sizes, which aids in figuring out how many clusters are best for the dataset. When it comes to clustering techniques, these visualizations help with successful data analysis and interpretation by providing insights into the underlying structure of the data and helping to comprehend how data points are grouped or clustered depending on their qualities.

- **Line Fitting**

  The graph shows that insulin and glucose levels are positively correlated; as insulin levels rise, so do glucose levels, as seen by the upward-sloping line. This is consistent with the function of insulin in promoting the uptake of glucose by cells. Notwithstanding, the dispersed data points surrounding the line indicate probable factors influencing individual responses to insulin and reveal diversity in the amount glucose levels increase for a given insulin level.



5. Conclusion

This study examined a dataset connected to diabetes using EDA and machine learning approaches. Relationships between variables including age, insulin, and glucose were potentially shown through visualization. Line fitting verified a positive association between insulin and glucose, while K-means clustering found categories within the data. It is advised to conduct additional research to examine causal links and include new aspects in order to develop a more thorough understanding of diabetes.