Clustering Analysis of Diabetes Risk and Progression in Pima Indian Females

By Hetu Virajkumar Patel, Student Number: 501215707
CPS 803 - Machine Learning

Introduction:

The dataset used here originates from the National Institute of Diabetes and Digestive and Kidney Diseases and serves as a diagnostic tool to predict diabetes onset among the specific population. The data of **770 samples** comprises medical diagnostic measurements collected from females aged 21 years or older, all from the Pima Indian Heritage(North Americans Indians traditionally from Arizona, US).

This demographic specificity offers valuable insights into genetic, physiological, and lifestyle factors influencing diabetes risk. The dataset includes predictor variables such as **pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, BMI, age, and the Diabetes Pedigree Function, a measure of familial diabetes history.**

These features are paired with a **binary target variable, Outcome**, where "1" indicates diabetes presence and "0" its absence. Through clustering analysis, this data allows us to uncover patterns and subgroup characteristics associated with diabetes risk and progression. Such insights can support early detection, personalized interventions, and improved disease management strategies.

I chose to work with this specific database because, as an Indian female in North America with family members affected by severe diabetes, it resonates personally. It also aligns with my passion for applying ML in healthcare to identify patterns for early detection and improved management of diabetes.

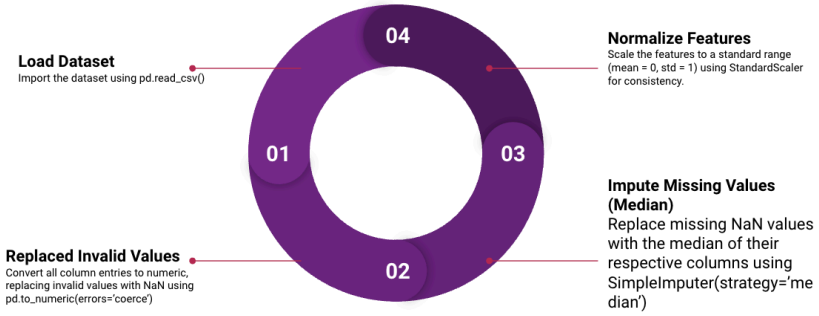
Methodology:

Loading and Preprocessing the Dataset:

The dataset was loaded using the pandas library, and the head() function was used to inspect the first few rows. Data preprocessing involved replacing invalid zero values with NaN in columns such as Glucose, Insulin, and BMI, where zeros are not valid. Missing values were then imputed using the median of each column to preserve the integrity of the dataset.

$$X_{imputed} = \begin{cases} X & \text{if } X \neq NaN \\ Median(x) & \text{if } X = NaN. \end{cases}$$

Feature scaling was performed using the MinMaxScaler to normalize all numerical columns to a range of [0, 1]. This ensured that features like BMI and glucose levels, which have different ranges, contributed equally to the clustering process.



Exploratory Data Analysis:

To understand the dataset better, I generated pair-plots and correlation heatmaps:

- Pair-plots: These visualized relationships between features such as BMI vs Glucose or Insulin vs Age. Patterns identified in these plots hinted at potential clustering tendencies among data points.
- Correlation Heatmap: This helped depict the relationship between variables. Highly correlated features were identified, which could influence the clustering process.

Clustering Methods Used:

K-Means Clustering:

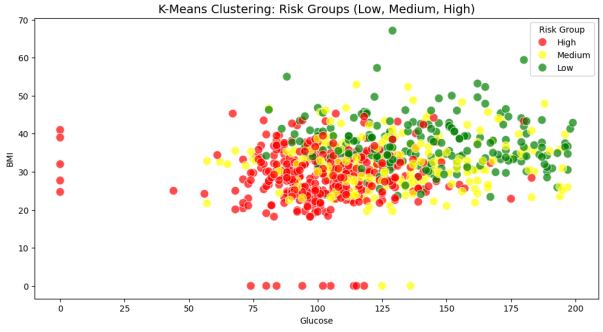
K-Means is a partitioning algorithm that divides data into clusters by minimizing intra-cluster variance. The Elbow Method was employed to determine the optimal number of clusters, which was found to be **k=3**, aligning with the objective of grouping data into three categories. K-Mean formula used to minimize the within-cluster sum of squares (WWCC)::

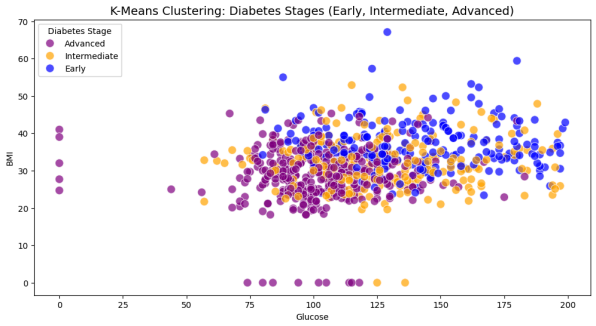
$$WWCS = \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - \mu_k\|^2$$

K-Means was applied to classify:

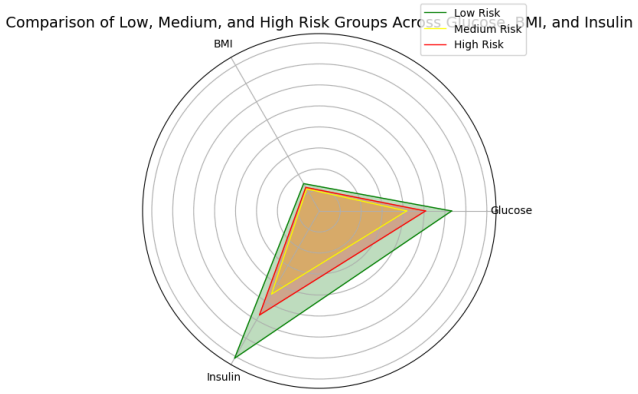
- **Risk Groups**: Low, Medium, High Risk.
- **Diabetes Stages**: Early, Intermediate, Advanced.

A scatter plot was used to visualize the clusters in two dimensions (**Glucose** and **BMI**), with distinct colors representing each cluster.





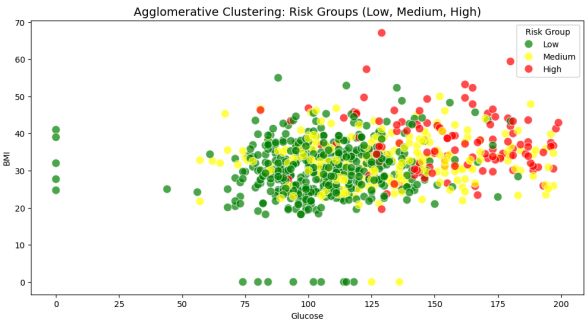
Group analysis provided insights into how the indicators varied among the clusters, highlighting trends in diabetes risk. A radar chart was then employed to visualize these variations across the clusters, enabling clear comparisons of the key metrics for each risk group. This visualization effectively illustrated the differences in health profiles, aiding in intuitive understanding of diabetes risk categorization.

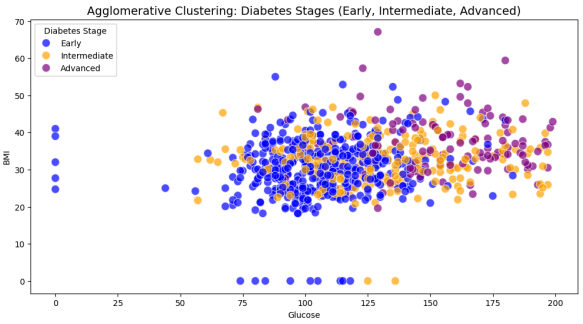


Hierarchical Clustering:

Hierarchical clustering groups data based on a tree-like structure (dendrogram), using the Ward linkage method to minimize the variance within clusters. The dendrogram was analyzed to identify clusters, with the tree being "cut" at a level corresponding to three clusters. Hierarchical clustering was also applied to classify the data into **Risk Groups** and **Diabetes Stages**.

The dendrogram provided a visual representation of how the data points were grouped and their hierarchical relationships.





Results:

In the K-Means clustering analysis, the dataset was effectively divided into three distinct clusters: a Low-Risk Group (Cluster 0) with lower glucose, BMI, and insulin levels; a Medium-Risk Group (Cluster 1) with elevated glucose or BMI, indicating prediabetes; and a High-Risk Group (Cluster 2) showing significantly high glucose, insulin, and BMI levels, indicative of advanced diabetes. Scatter plots illustrated clear separations between these groups, particularly in glucose and BMI values.

Hierarchical clustering, on the other hand, produced a detailed dendrogram highlighting the hierarchical relationships between data points. When the dendrogram was cut into three clusters, it revealed similar groupings to K-Means: Low, Medium, and High Risk, each representing varying degrees of diabetes progression. The dendrogram, however, provided a more nuanced view of how these clusters were formed, offering a granular perspective on the data's structure.

K-Means clustering is efficient for large datasets, offering clear cluster boundaries but requiring the number of clusters to be pre-specified and being sensitive to initial centroids. Hierarchical clustering avoids pre-specifying clusters and reveals detailed data relationships but is computationally intensive and can produce overly complex dendrograms with large datasets. While K-Means was more effective for this dataset due to its speed, hierarchical clustering provided valuable insights into the data's structure.

Conclusion:

In conclusion, the clustering analysis of diabetes risk and progression in Pima Indian females provided valuable insights into the varying degrees of diabetes. Through K-Means clustering, the dataset was segmented into three risk groups: Low, Medium, and High Risk, based on glucose, BMI, and insulin levels. Hierarchical clustering, although computationally intensive, offered additional insights into the data's hierarchical relationships, validating the findings from K-Means. Both methods highlighted the importance of these health indicators in predicting diabetes onset, with K-Means being more efficient and hierarchical clustering offering a deeper understanding of the data structure. These analyses contribute to early detection and personalized interventions.

References:

Pima Indian Heritage Diabetes Dataset from Kaggle:
https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database?resource=download

The Google Colab Notebook I created:
script_HetuPatel

Facts about Diabetes understood from:
https://www.hopkinsmedicine.org/health/conditions-and-diseases/diabetes