**Unsupervised Learning -- Clustering Use Case with Diabetes Dataset**

**Business Understanding**

* **Problem Statement**

The problem we are addressing is the prediction of diabetes in patients based on various medical attributes. This is a classification problem where the goal is to accurately categorize patients as diabetic or non-diabetic.

* **Importance of the Problem**

Predicting diabetes is critical due to its increasing prevalence and the severe health complications associated with it. Early detection can lead to better management and prevention strategies, improving patient outcomes and reducing healthcare costs.

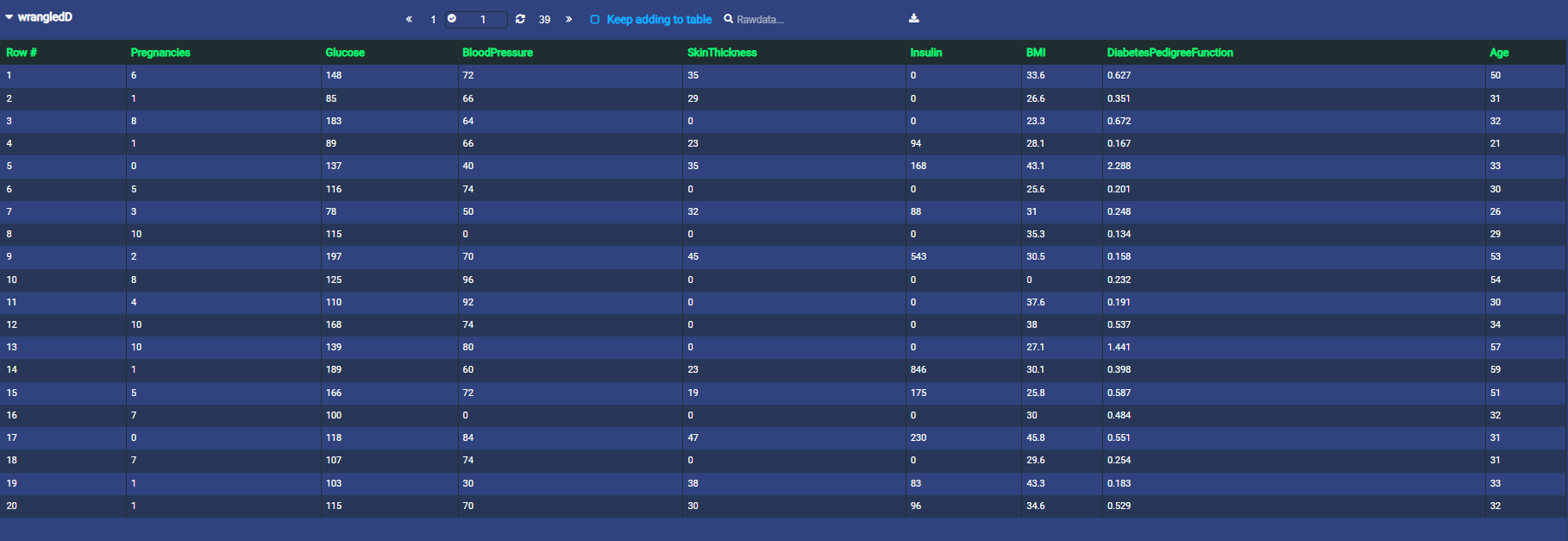
* **Data Source and Description**

The dataset used for this analysis is sourced from the National Institute of Diabetes and Digestive and Kidney Diseases. The data includes various medical predictors such as age, body mass index (BMI), blood pressure, and others, which are used to predict whether a patient has diabetes (Outcome variable).

**Data Collection**

The dataset was downloaded from [Kaggle](https://www.kaggle.com/datasets/mathchi/diabetes-data-set), a popular data science and machine learning platform. The specific dataset used is the "Diabetes" dataset.

**The dataset is named diabetes.csv and contains the following columns:**

****

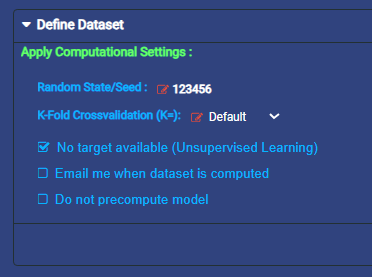
* Pregnancies: Number of times pregnant
* Glucose: Plasma glucose concentration
* BloodPressure: Diastolic blood pressure (mm Hg)
* SkinThickness: Triceps skinfold thickness (mm)
* Insulin: 2-Hour serum insulin (mu U/ml)
* BMI: Body mass index (weight in kg/(height in m)^2)
* DiabetesPedigreeFunction: Diabetes pedigree function
* Age: Age (years)

**Data Understanding**

**Exploratory Data Analysis (EDA)**

Initial data exploration reveals the following:

* The dataset contains 768 rows and 9 columns.
* There is no target variable so while defining the dataset checked the no target availbale(Unsupervised)



* There are several missing or zero values in columns like Glucose, BloodPressure,

SkinThickness, Insulin, and BMI.



**Obervations (Count of zero value)**

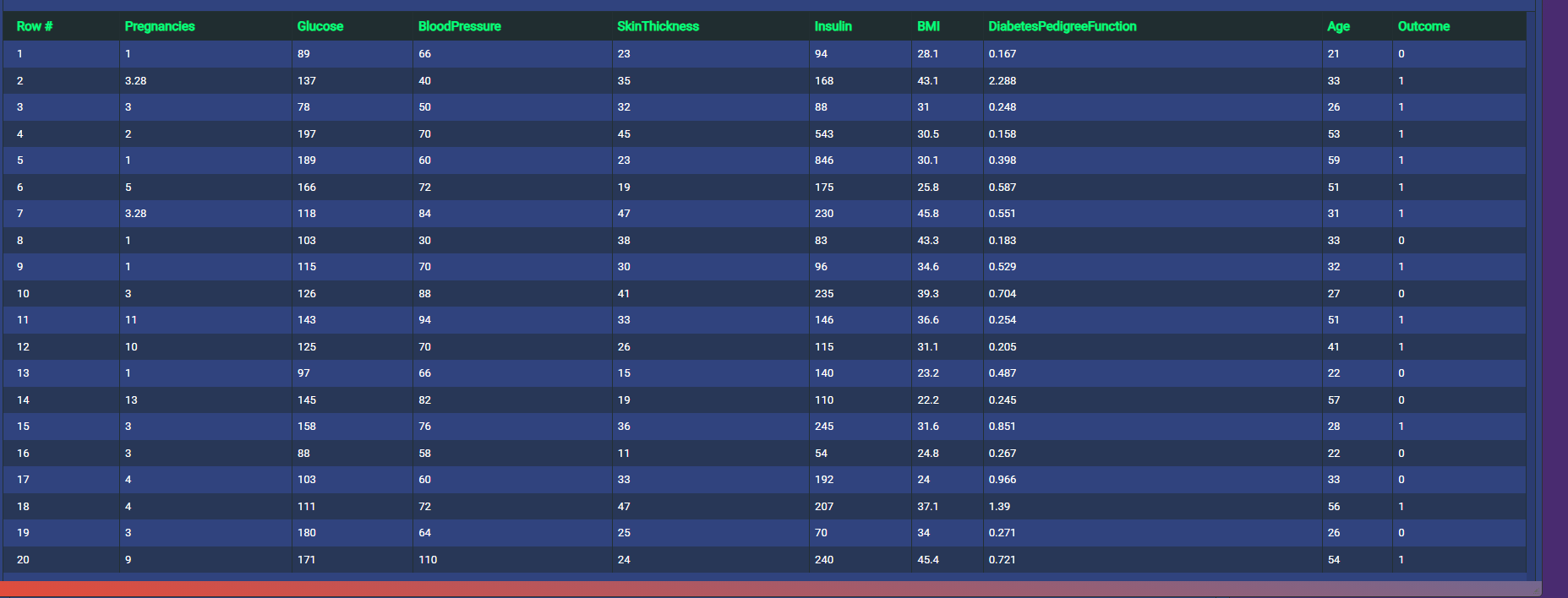
**Glucose**: 5 instances

**BloodPressure**: 35 instances

**SkinThickness**: 227 instances

**Insulin**: 374 instances

**BMI**: 11 instances

****

**Data Preparation**

**Handling Missing Values**

Replaced zero values in Glucose, BloodPressure, SkinThickness, Insulin, and BMI with the mean values of their respective columns.

**Data Splitting**

The dataset was split into training and testing sets:

* 80% of the data was used for training.
* 20% of the data was reserved for testing.

**Methodology**

**Model Selection**

For Clustering, we used the following algorithms:

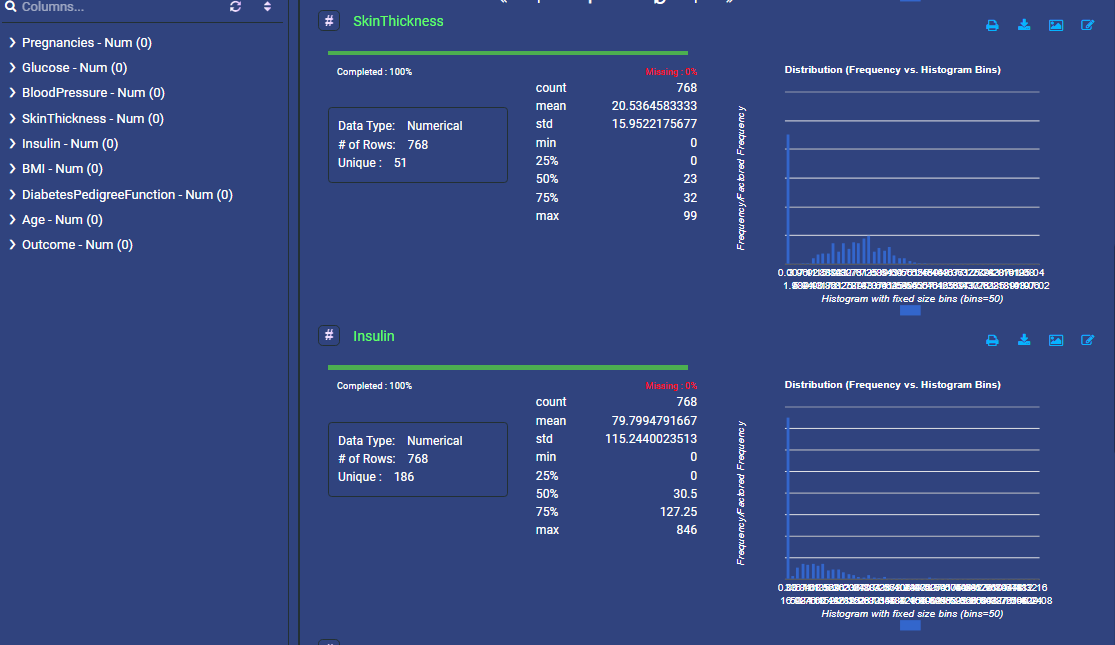
* **K-means Clustering**

K-means clustering is a popular unsupervised machine learning algorithm used for partitioning a dataset into a pre-defined number of clusters.

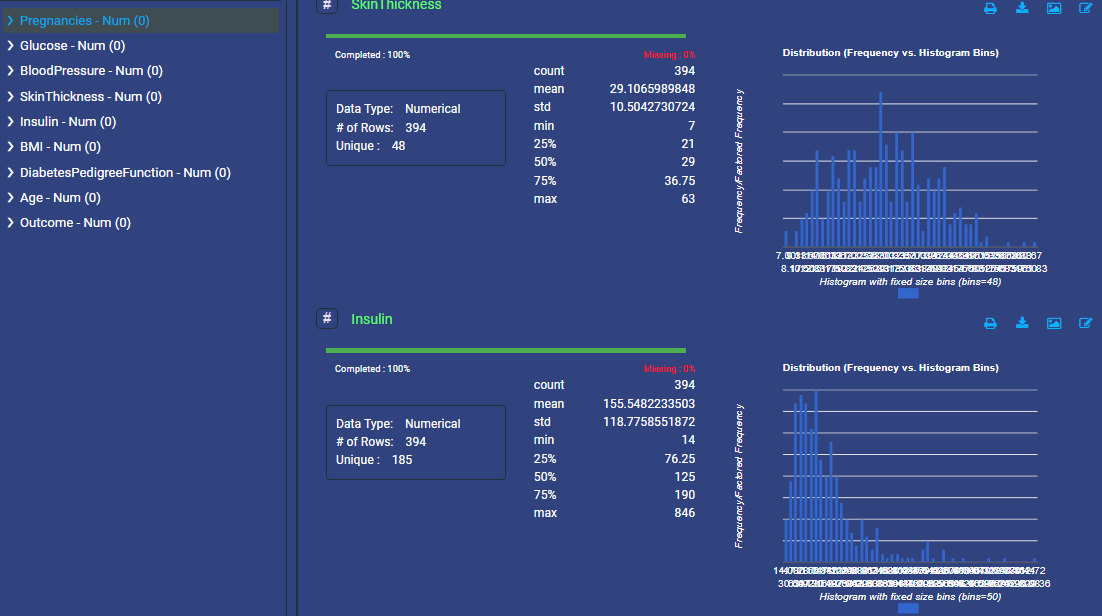
* **Affinity Propagation**

The method works on simple estimators as well as on nested objects (such as Pipeline).

**Original Data Distribution of Skin Thickness and Insulin**

****

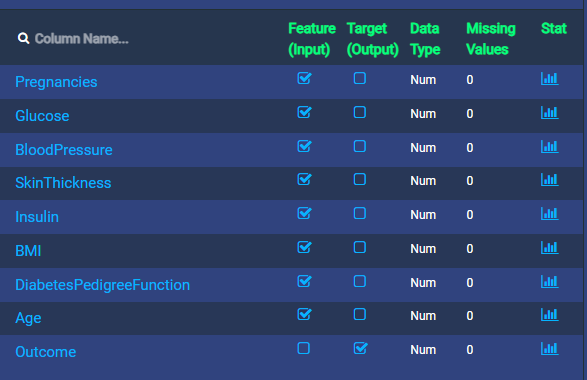
**Wrangled Data Distribution of Skin Thickness and Insulin**

****

**Model Training**

Each model was trained using the training dataset.

**Feature and Target:**



**Model Evaluation**

The models were evaluated on the test dataset using the following metrics:

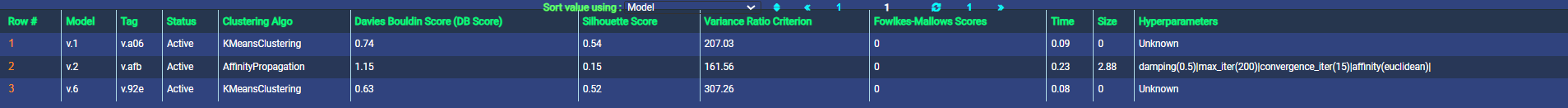
* Davies Boulding Score(DB Score)
* Silhouette Score
* Variance Ratio Criterion
* Fowlkes-Mallows Scores

**Results**

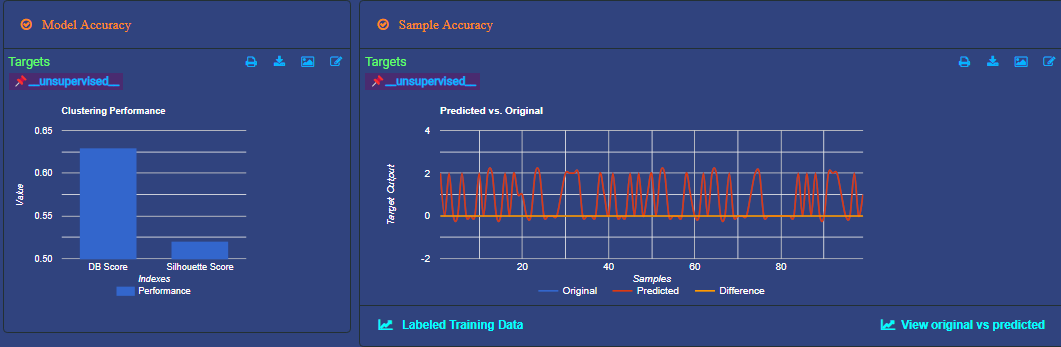
****

Rank 1 and Rank 2 represent the KMeansClustering; however, Rank 1 has 2 clusters and Rank 2 has 3 clusters, which shows the cluster 0.63 and 0.74 individually, and Rank 3 indicates affinity propagation with 1.15 cluster

**Performance Metrics**

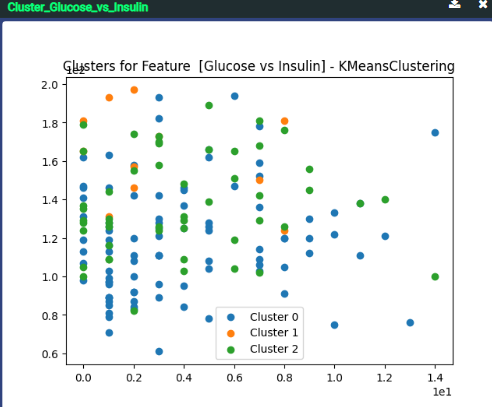
****

**Model Accuracy**

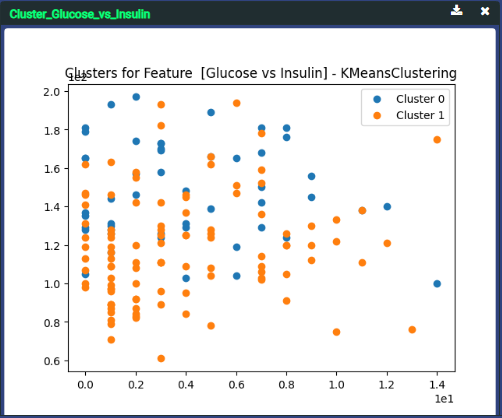
****

**Modeling Details**

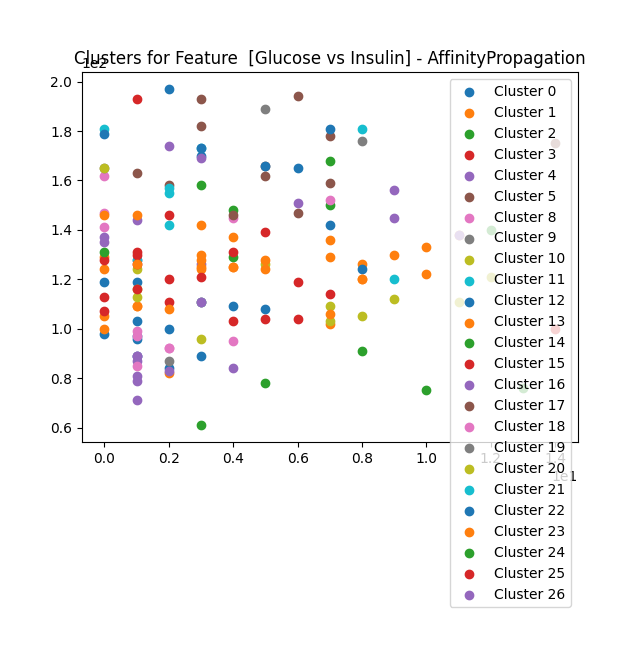
**Cluster Glucose vs Insulin (3 Cluster)**

****

**Cluster Glucose vs Insulin (2 Cluster)**

****

**Cluster Glucose vs Insulin (Affinity Propagation / 26 Cluster)**



**Conclusions**

**Improvements**

**What improvements would you like to make in future?**

Include more relevant features to improve the model's predictive power.

Apply advanced techniques like ensemble learning to boost model performance.

**Real-life Application**

**How do you think the solution could be used in real life?**

The solution can be integrated into healthcare systems to assist doctors in early diabetes diagnosis, leading to timely intervention and better patient management.

**Value to Client**

**What value do you think the solution will have to the client?**

The model provides a reliable tool for predicting diabetes, which can enhance patient care and reduce long-term healthcare costs by enabling early detection and prevention strategies.

**Key Learnings**

**What did you learn through this project?**

The importance of data preprocessing and feature engineering in building effective models.

The value of model evaluation metrics in selecting the best model for deployment.

This project provided valuable insights into the application of machine learning in healthcare, demonstrating the potential impact of data-driven solutions in real-world scenarios.