



SAIT

UNSUPERVISED LEARNING - CLUSTERING

DATA SCIENCE

Diabetes Dataset

Written by:
Shubh Desai

July 20, 2024

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.

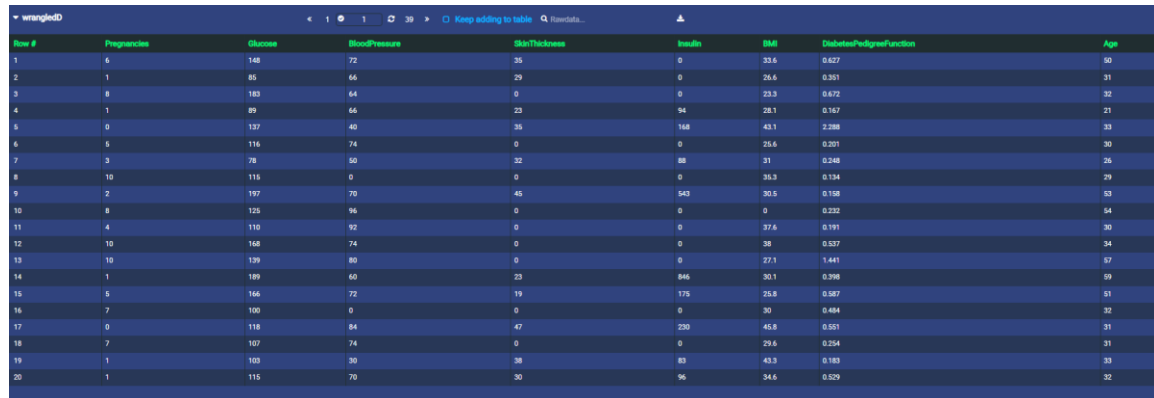
- **Data Source and Description**

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](#), 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:



Row #	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
1	6	146	72	35	0	33.6	0.627	30
2	1	85	66	29	0	26.6	0.351	31
3	8	183	64	0	0	23.3	0.672	32
4	1	89	66	23	94	28.1	0.167	21
5	0	137	40	35	168	43.1	2.288	33
6	5	116	74	0	0	25.6	0.201	30
7	3	78	60	32	88	31	0.248	26
8	10	115	0	0	0	35.3	0.134	29
9	2	197	70	46	943	30.5	0.168	53
10	8	125	96	0	0	0	0.232	54
11	4	110	92	0	0	37.4	0.191	30
12	10	168	74	0	0	38	0.537	54
13	10	129	80	0	0	27.1	1.441	57
14	1	189	60	23	846	30.1	0.398	59
15	5	166	72	19	175	25.8	0.587	51
16	7	100	0	0	0	30	0.484	32
17	9	118	84	47	230	45.8	0.551	31
18	7	107	74	0	0	29.4	0.254	31
19	1	103	30	38	83	43.3	0.183	33
20	1	115	70	30	96	34.6	0.529	32

- 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 available(Unsupervised)

Define Dataset

Apply Computational Settings :

Random State/Seed : ☒ 123456

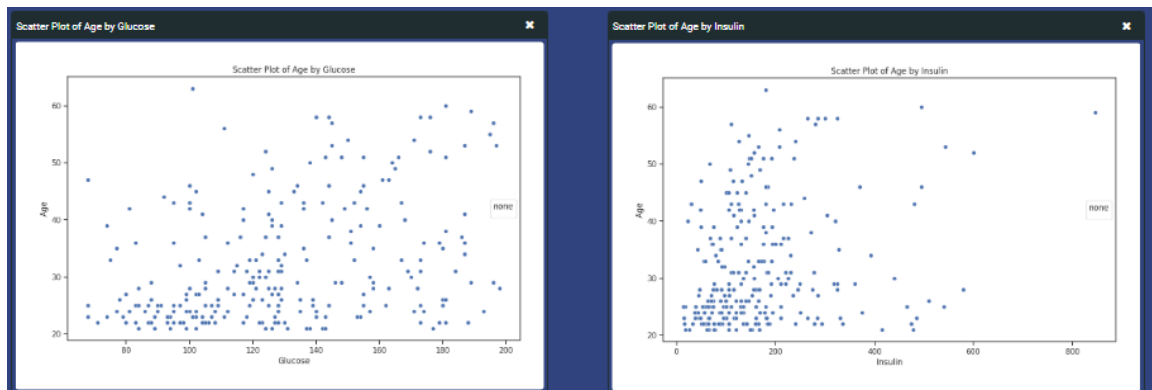
K-Fold Crossvalidation (K=): ☒ Default ▼

☒ No target available (Unsupervised Learning)

☐ Email me when dataset is computed

☐ Do not precompute model

- There are several missing or zero values in columns like Glucose, BloodPressure, SkinThickness, Insulin, and BMI.



Observations (Count of zero value)

Glucose: 5 instances

BloodPressure: 35 instances

SkinThickness: 227 instances

Insulin: 374 instances

BMI: 11 instances

Row #	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
1	1	89	66	23	94	28.1	0.167	21	0
2	3.28	137	40	35	168	43.1	2.288	33	1
3	3	78	50	32	88	31	0.248	26	1
4	2	197	70	45	543	30.5	0.158	53	1
5	1	189	60	23	846	30.1	0.398	99	1
6	5	166	72	19	175	25.8	0.587	51	1
7	3.28	118	84	47	230	45.8	0.551	31	1
8	1	103	30	38	83	43.3	0.183	33	0
9	1	115	70	30	96	34.6	0.529	32	1
10	3	126	88	41	235	39.3	0.704	27	0
11	11	143	94	33	146	36.6	0.254	51	1
12	10	125	70	26	115	31.1	0.205	41	1
13	1	97	66	15	140	23.2	0.487	22	0
14	13	145	82	19	110	22.2	0.245	67	0
15	3	198	76	36	245	31.6	0.851	28	1
16	3	88	58	11	54	24.8	0.267	22	0
17	4	103	60	33	192	24	0.966	33	0
18	4	111	72	47	207	37.1	1.39	56	1
19	3	180	64	25	70	34	0.271	26	0
20	9	171	110	24	240	45.4	0.721	54	1

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:

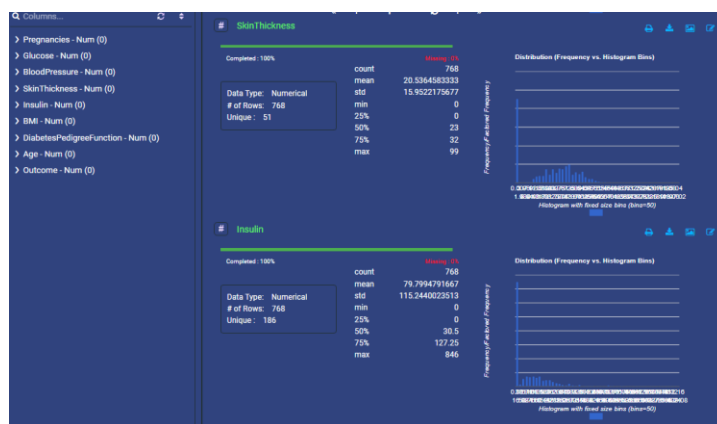
- **K-means Clustering**

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

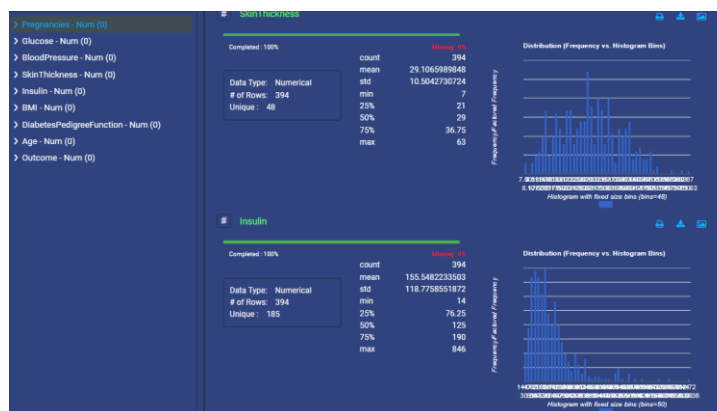
- **Affinity Propagation**

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:

Column Name...	Feature (Input)	Target (Output)	Data Type	Missing Values	Stat
Pregnancies	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Num	0	
Glucose	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Num	0	
BloodPressure	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Num	0	
SkinThickness	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Num	0	
Insulin	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Num	0	
BMI	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Num	0	
DiabetesPedigreeFunction	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Num	0	
Age	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Num	0	

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

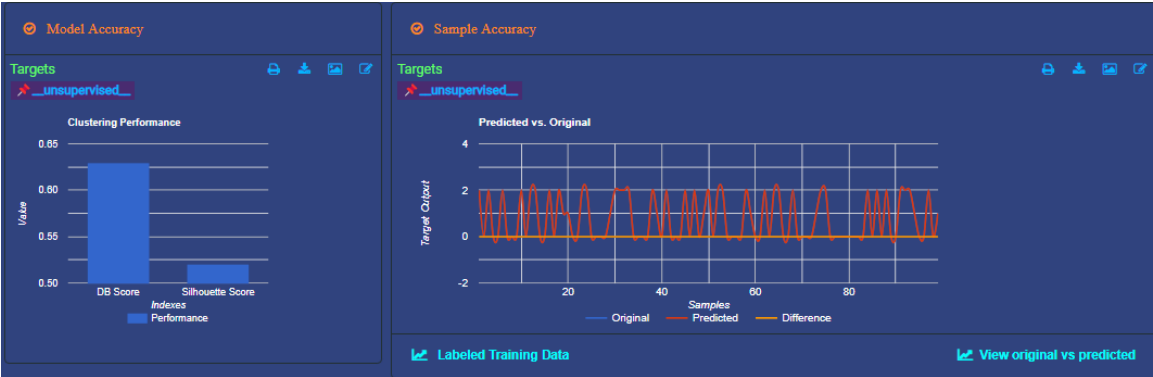
Version-Tag	Dataset	Algorithm	Rank	Cluster	Doc.	Publish	Delete
<input type="checkbox"/> v.6-v.92e	diabetesWR-DiabetesWR	KMeansClustering	1	0.63	 	↶	✖
<input type="checkbox"/> v.1-v.a06	diabetesWR-DiabetesWR	KMeansClustering	2	0.74	 	↶	✖
<input type="checkbox"/> v.2-v.afb	diabetesWR-DiabetesWR	AffinityPropagation	3	1.15	 	↶	✖

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

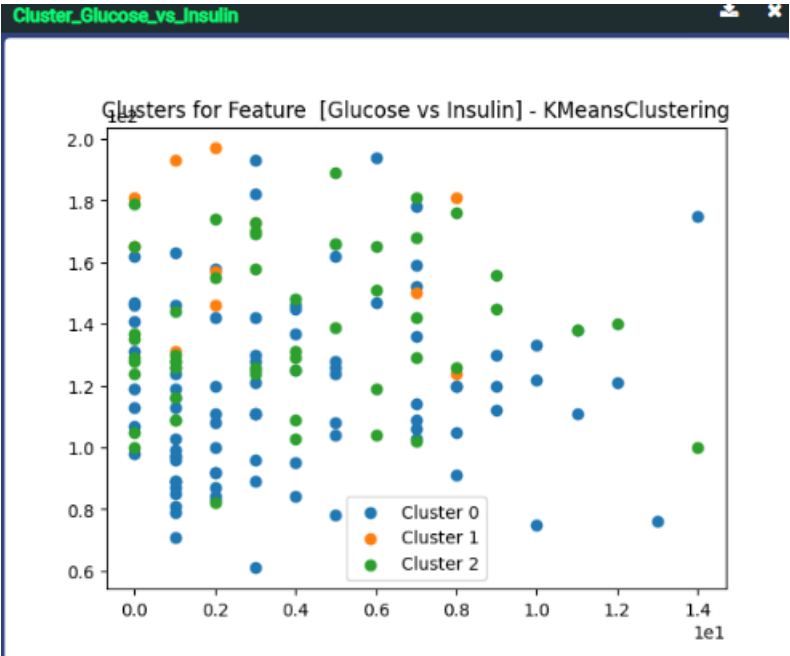
Row #	Model	Tag	Status	Clustering Algo	Davies Bouldin Score (DB Score)	Sort value using - Model Silhouette Score	Variance Ratio Criterion	Pearson-Matthews Scores	Time	Size	Hyperparameters
1	v1	v.a06	Active	KMeansClustering	0.74	0.54	257.03	0	0.09	0	Unknown
2	v2	v.a0b	Active	AffinityPropagation	1.15	0.15	161.56	0	0.23	2.88	damping(0.5)(max_iter(200)(convergence_iter(10)(affinity(xcclidean))
3	v3	v.92e	Active	KMeansClustering	0.63	0.52	367.26	0	0.08	0	Unknown

Model Accuracy

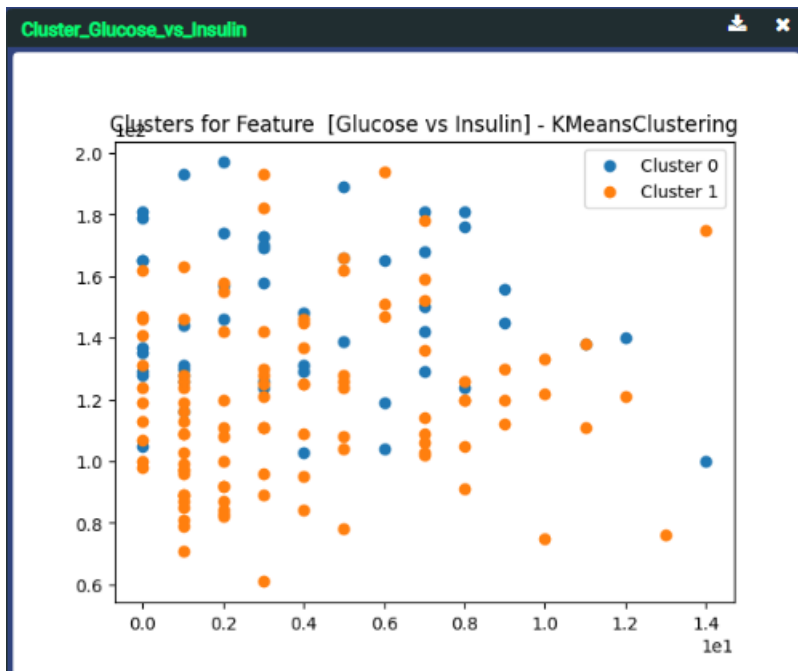


Modeling Details

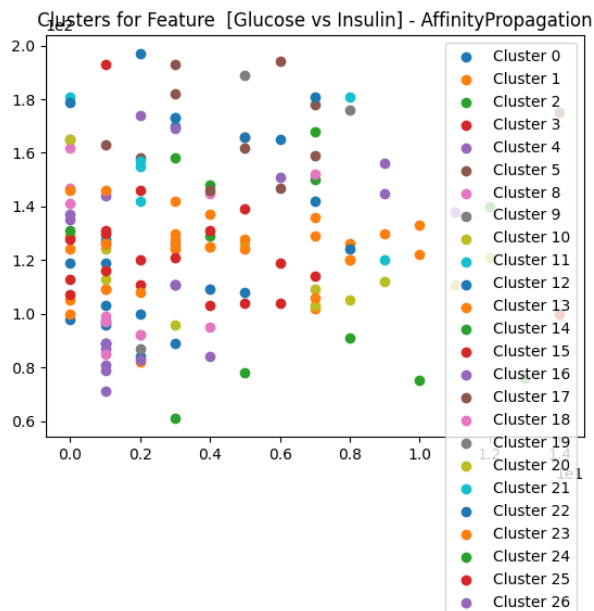
Cluster Glucose vs Insulin (3 Cluster)



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.