SemiAuto clustering Report

Dataset: DIABETES_DATASET

Generated on: 2025-05-29 12:36:36

Project Flow

Table of Contents

1. Data Ingestion

This step involves loading and analyzing the original dataset to understand its structure and characteristics.

Dataset Overview

Dataset: diabetes_dataset

Train samples: 8000, Test samples: 2000

Target column: N/A

Column Types

Original Columns:

Unnamed: 0, Age, Sex, Ethnicity, BMI, Waist_Circumference, Fasting_Blood_Glucose, HbA1c, Blood_Pressure_Systolic, Blood_Pressure_Diastolic, Cholesterol_Total, Cholesterol_HDL, Cholesterol_LDL, GGT, Serum_Urate, Physical_Activity_Level, Dietary_Intake_Calories, Alcohol_Consumption, Smoking_Status, Family_History_of_Diabetes, Previous_Gestational_Diabetes

Numerical Columns:

Age, BMI, Waist_Circumference, Fasting_Blood_Glucose, HbA1c, Blood_Pressure_Systolic, Blood_Pressure_Diastolic, Cholesterol_Total, Cholesterol_HDL, Cholesterol_LDL, GGT, Serum_Urate, Dietary_Intake_Calories, Family_History_of_Diabetes, Previous_Gestational_Diabetes

Categorical Columns:

Sex, Ethnicity, Physical_Activity_Level, Alcohol_Consumption, Smoking_Status

Skewed Columns:

None

Normal Columns:

Age, BMI, Waist_Circumference, Fasting_Blood_Glucose, HbA1c, Blood_Pressure_Systolic, Blood_Pressure_Diastolic, Cholesterol_Total, Cholesterol_HDL, Cholesterol_LDL, GGT, Serum_Urate, Dietary_Intake_Calories, Family_History_of_Diabetes, Previous_Gestational_Diabetes

Columns with Nulls:

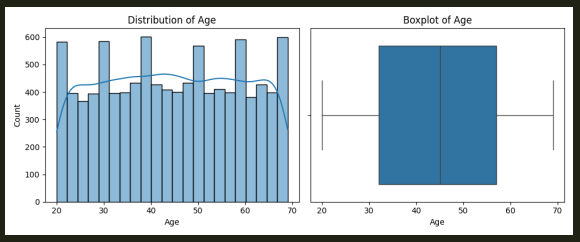
Alcohol_Consumption

Columns with Outliers:

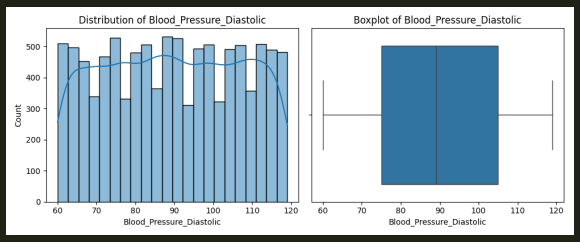
None

Feature Distributions

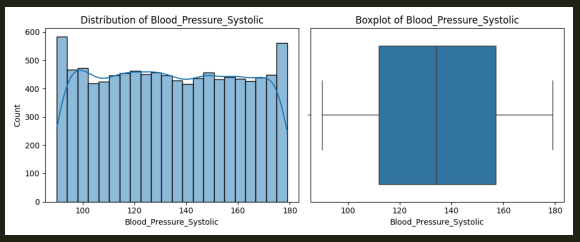
Age



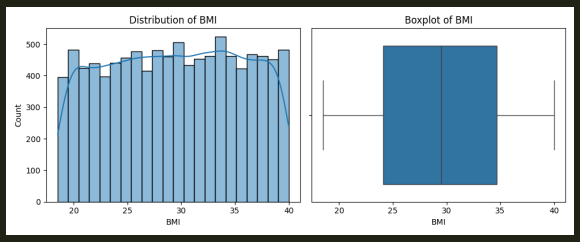
Blood_Pressure_Diastolic



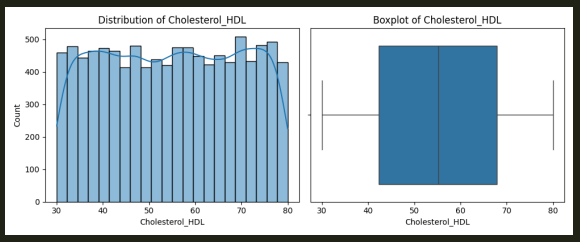
Blood_Pressure_Systolic



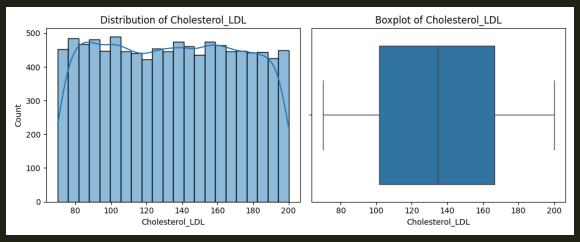
BMI



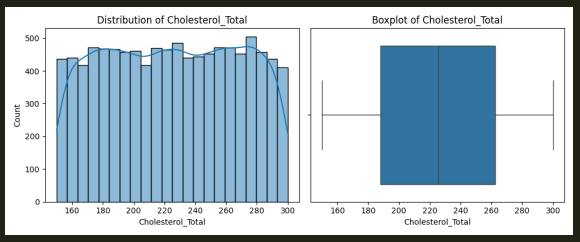
Cholesterol_HDL



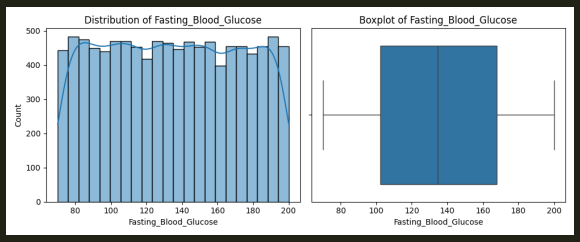
Cholesterol_LDL



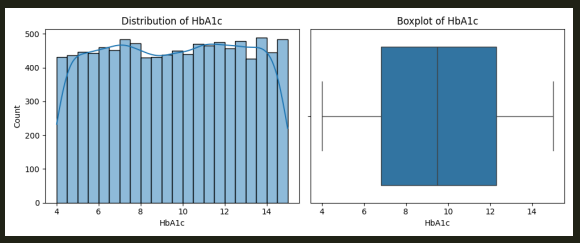
Cholesterol_Total



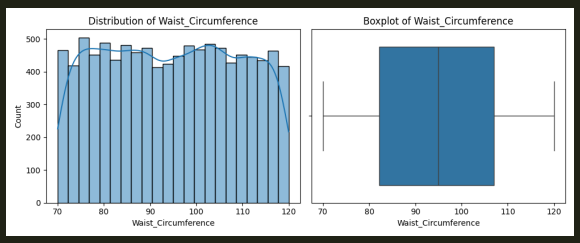
Fasting_Blood_Glucose



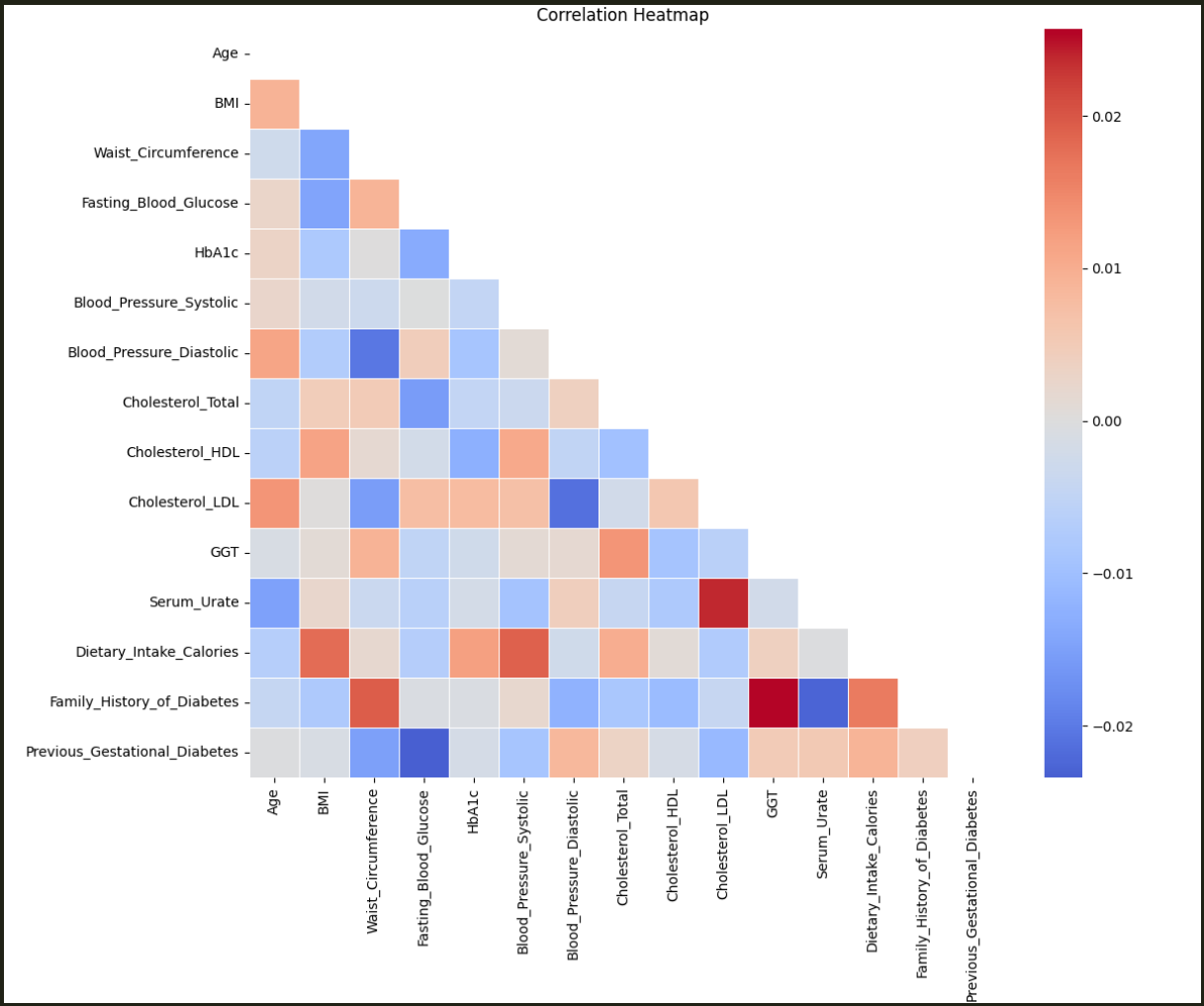
HbA1c



Waist_Circumference



Correlation Heatmap



2. Data Preprocessing

This step involves cleaning the dataset and preparing it for model training.

Preprocessed Data Preview

Training Data Sample (First 5 rows):

Age	BMI	Waist_Circ..	Fasting_Bl..	HbA1c	Blood_Pres..	Blood_Pres..	Cholestero..	Cholestero..	Cholestero..
0.56	0.29523809..	-0.0668016..	-0.3161094..	0.61818181..	-0.5333333..	-0.1724137..	-0.7287449..	0.24218749..	-0.4814814..
0.88	0.06666666..	-0.2287449..	0.60182370..	-0.4363636..	-0.4444444..	-0.5172413..	0.05802968..	0.65234374..	-0.3302469..
0.68	-0.7619047..	-0.2246963..	-0.6458966..	0.29090909..	0.84444444..	0.75862068..	0.31848852..	0.89453124..	-0.2854938..
-0.28	0.53333333..	-0.1639676..	-0.4407294..	-0.9090909..	0.11111111..	-0.5517241..	0.73819163..	0.83593749..	0.65432098..
0.6	-0.5238095..	-0.3623481..	-0.9118541..	0.32727272..	0.64444444..	-0.2413793..	-0.8933873..	0.74609374..	-0.1743827..

Test Data Sample (First 5 rows):

Age	BMI	Waist_Circ..	Fasting_Bl..	HbA1c	Blood_Pres..	Blood_Pres..	Cholestero..	Cholestero..	Cholestero..
0.2	-0.1714285..	0.52429149..	0.57902735..	0.87272727..	0.33333333..	0.93103448..	0.76383265..	-0.7421874..	0.27160493..
-0.8	-0.1809523..	-0.3825910..	-0.5881458..	0.69090909..	-0.2666666..	0.13793103..	-0.6302294..	0.01562499..	0.33024691..
0.44	0.55238095..	-0.4514170..	0.16261398..	0.81818181..	0.97777777..	0.13793103..	-0.5816464..	0.33984374..	-0.0092592..
0.36	0.98095238..	-0.8117408..	-0.4361702..	-0.9454545..	0.24444444..	-0.8965517..	-0.8798920..	-0.4960937..	-0.6682098..
0.2	0.23809523..	-0.4068825..	0.24468085..	-0.4727272..	-0.4	-0.3793103..	-0.4156545..	0.35156249..	-0.8996913..

3. Feature Engineering

This step involves creating new features or selecting the most important ones.

Feature Engineering Configuration

Applied Techniques:

Automated Feature Engineering: Yes

SHAP-based Feature Selection: No

Transformed Data Preview

Transformed Training Data Sample (First 5 rows):

distance_t..	distance_t..	distance_t..	distance_t..	cluster_8 ..	distance_t..	distance_t..	cluster_8 ..	distance_t..	distance_t..
-9.7567528..	-9.8604071..	-9.2751739..	-63.349431..	-6.9166666..	-64.022446..	-60.222597..	-4.1166666..	-5.8070312..	-5.8687242..
-18.087533..	-16.325287..	-16.464806..	-18.545619..	-21.318493..	-16.738742..	-16.881795..	86.1627906..	73.1042459..	65.9817892..
-10.386467..	-9.9247077..	-9.6001662..	-21.534286..	-15.025862..	-20.576918..	-19.904045..	21.9788135..	15.1926215..	14.5171904..
20.4692768..	20.9500581..	20.4252752..	-28.276404..	17.6595744..	-28.940559..	-28.215620..	5.41864716..	6.28077358..	6.42829607..
-6.1016742..	-5.8037203..	-5.5846922..	-13.701339..	-7.3741362..	-13.032282..	-12.540453..	-6.7160120..	-5.5571143..	-5.2857520..

Transformed Test Data Sample (First 5 rows):

distance_t..	distance_t..	distance_t..	distance_t..	cluster_8 ..	distance_t..	distance_t..	cluster_8 ..	distance_t..	distance_t..
-23.441122..	-22.427902..	-21.902116..	9.01384236..	-34.721115..	8.62422770..	8.42204676..	9.16431095..	6.18706317..	5.91963332..
14.3769992..	13.1597854..	13.1885085..	-11.711048..	16.0438144..	-10.719544..	-10.742941..	-7.9336188..	-7.1093836..	-6.5074750..
-19.812173..	-18.252301..	-18.298288..	-9.9763412..	-21.996466..	-9.1908737..	-9.2140304..	-8.5962877..	-7.7426592..	-7.1330563..
13.2070154..	12.3801456..	12.3514129..	-6.2858351..	12.9417879..	-5.8922892..	-5.8786139..	-5.6825153..	-5.7989721..	-5.4359079..
-10.653598..	-9.6930439..	-9.7163529..	-10.767821..	-12.158203..	-9.7969678..	-9.8205267..	-12.029220..	-10.540578..	-9.5902136..

4. Model Building

This step involves training the clustering model on the transformed data.

Model Selection

Selected Model:

KMeans

Training timestamp: 2025-05-29 12:35:36

5. Model Evaluation

This step involves evaluating the performance of the trained clustering model.

Clustering Metrics

Original Model Performance:

Evaluation timestamp: 2025-05-29 12:36:05

Metric	Value
Silhouette Score	0.80199
Calinski-Harabasz Score	278.39420
Davies-Bouldin Score	0.40065
Number of Clusters	6.00000
Outlier Ratio	0.00000

6. Model Optimization

This step involves tuning the hyperparameters of the model to improve performance.

Error: Could not decode hyperparameters file.

Optimization timestamp: 2025-05-29 12:36:04

7. Final Evaluation Results

This section presents the final performance of the optimized clustering model.

Optimized Model Performance

Metric	Value
Silhouette Score	0.95005
Calinski-Harabasz Score	258.88698
Davies-Bouldin Score	0.02999
Number of Clusters	2.00000
Outlier Ratio	0.00000

Evaluation timestamp: 2025-05-29 12:36:05

Performance Comparison

Metric	Original Model	Optimized Model	Improvement
Silhouette Score	0.80199	0.95005	+18.46%
Davies-Bouldin Score	0.40065	0.02999	+92.52%
Calinski-Harabasz Score	278.39420	258.88698	-7.01%

Conclusion

Summary of the clustering model development and performance.

This report summarizes the development of a clustering model for the diabetes_dataset dataset. A KMeans clustering model was trained and optimized using hyperparameter tuning. The optimization process improved the model's Silhouette Score from 0.80199 to 0.95005, representing a 18.46% improvement.

This automatic report was generated to provide insights into the model development process and performance metrics. It includes details about data preprocessing, feature engineering, model selection, and evaluation results.