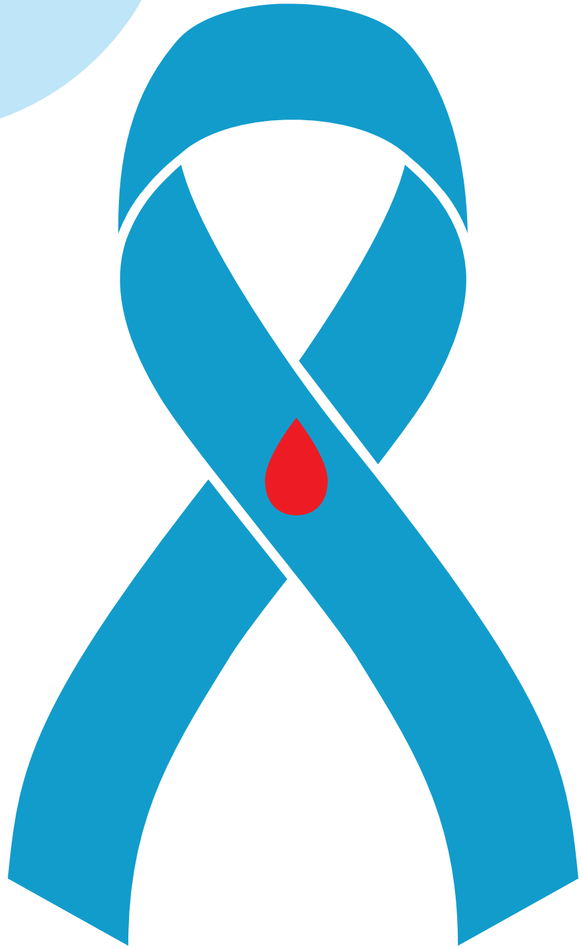
DIABETES PREDICTION

Using Data Mining Techniques



Problem

Diabetes is one of the most widespread and serious global health concerns, often developing silently with symptoms that go unnoticed until complications arise. This project aims to analyze patient health data to predict the risk of diabetes at an early stage, enabling individuals and healthcare professionals to take timely preventive actions.

Data

The dataset used in this project contains

- 1879 records (tuples)
- 46 attributes , for example:

Age

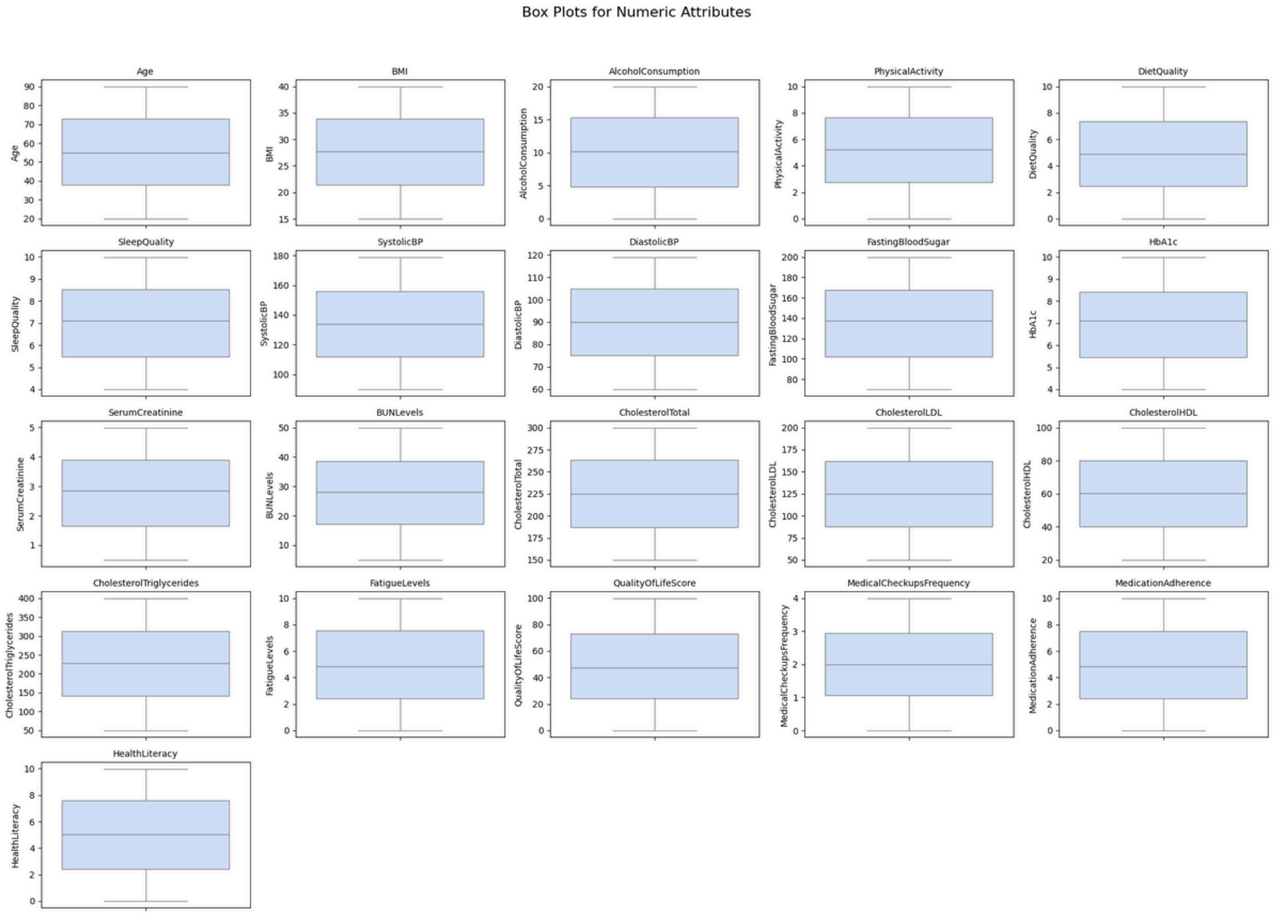
BMI

FatigueLevels

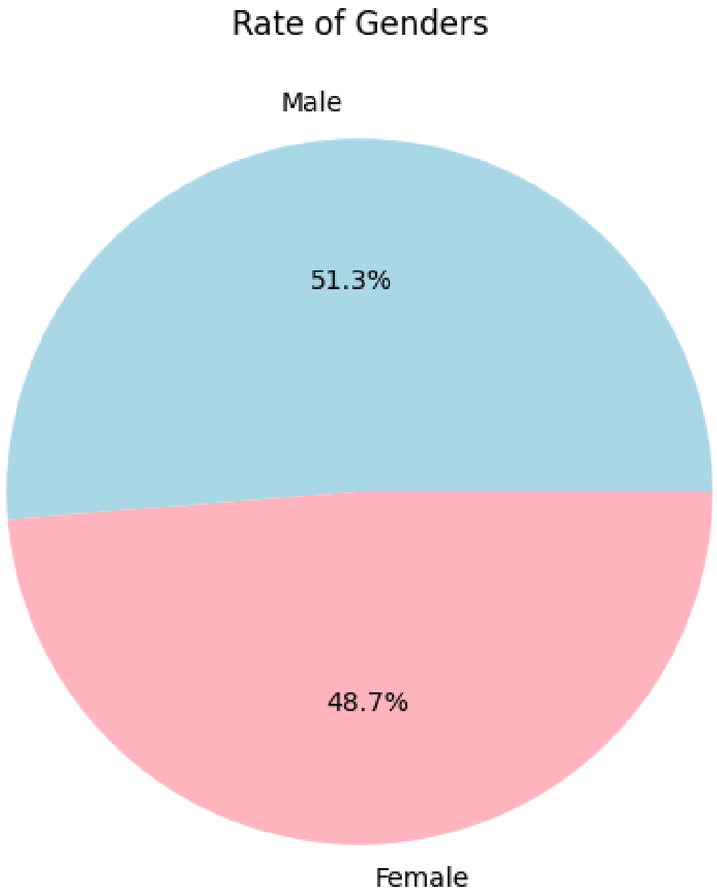
HbA1c

DiastolicBP

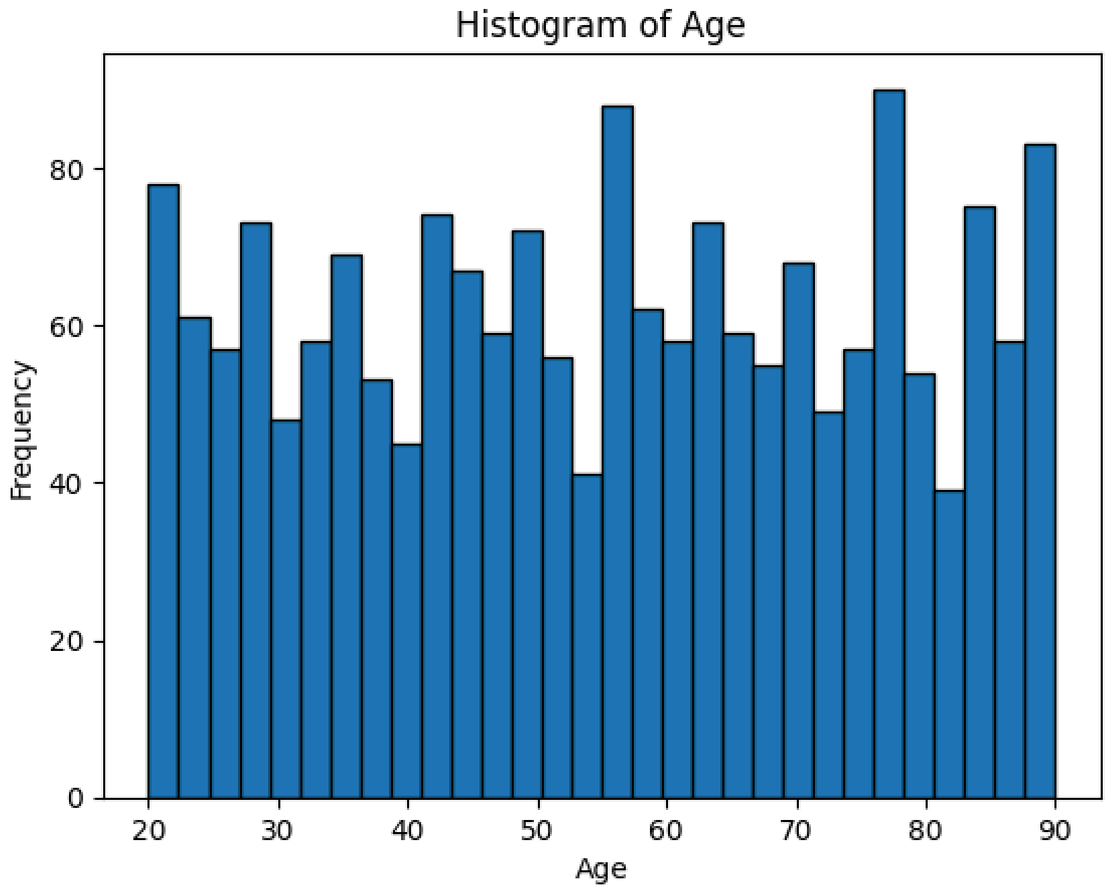
Data graphs


Box Plots for Numeric Attributes
Age
BMI
AlcoholConsumption
PhysicalActivity
DietQuality
SleepQuality
SystolicBP
DiastolicBP
FastingBloodSugar
HbA1c
SerumCreatinine
BUNLevels
CholesterolTotal
CholesterolLDL
CholesterolHDL
CholesterolTriglycerides
FatigueLevels
QualityOfLifeScore
MedicalCheckupsFrequency
MedicationAdherence
HealthLiteracy

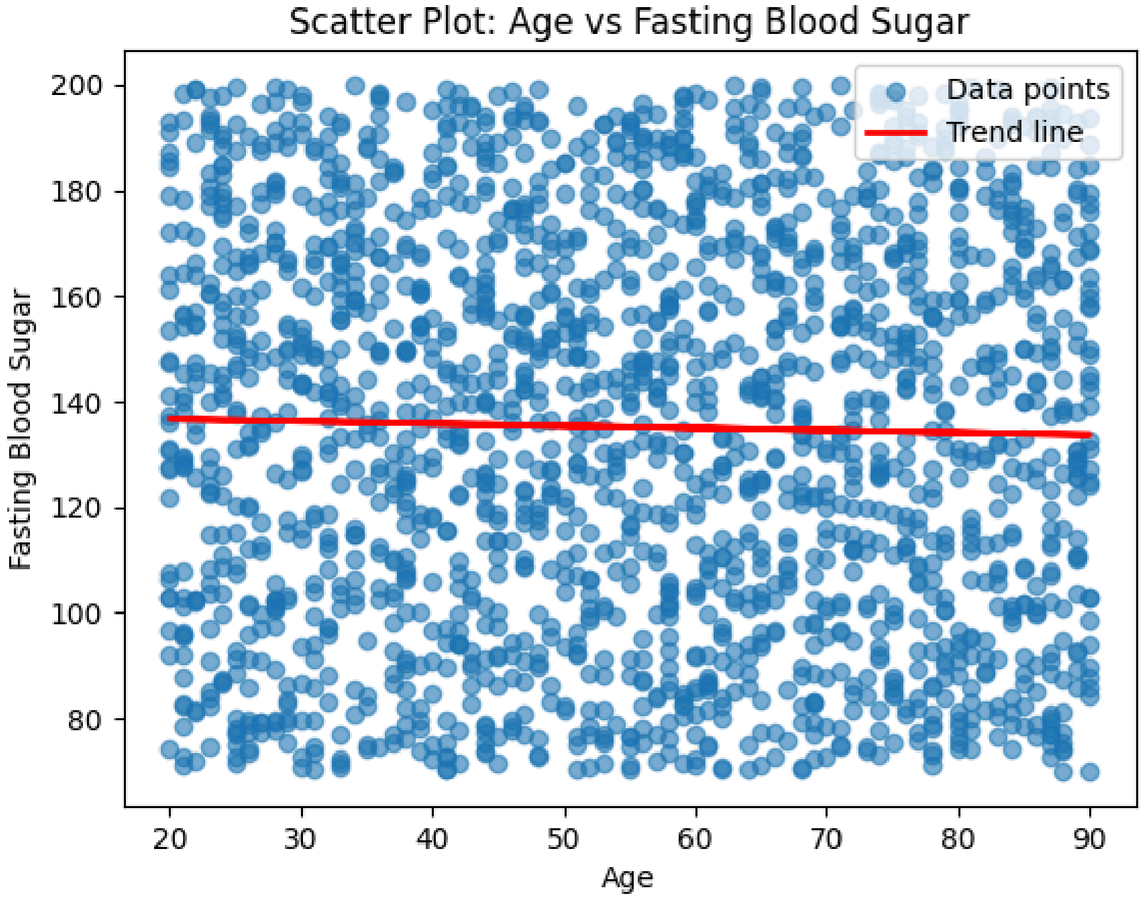
Data graphs



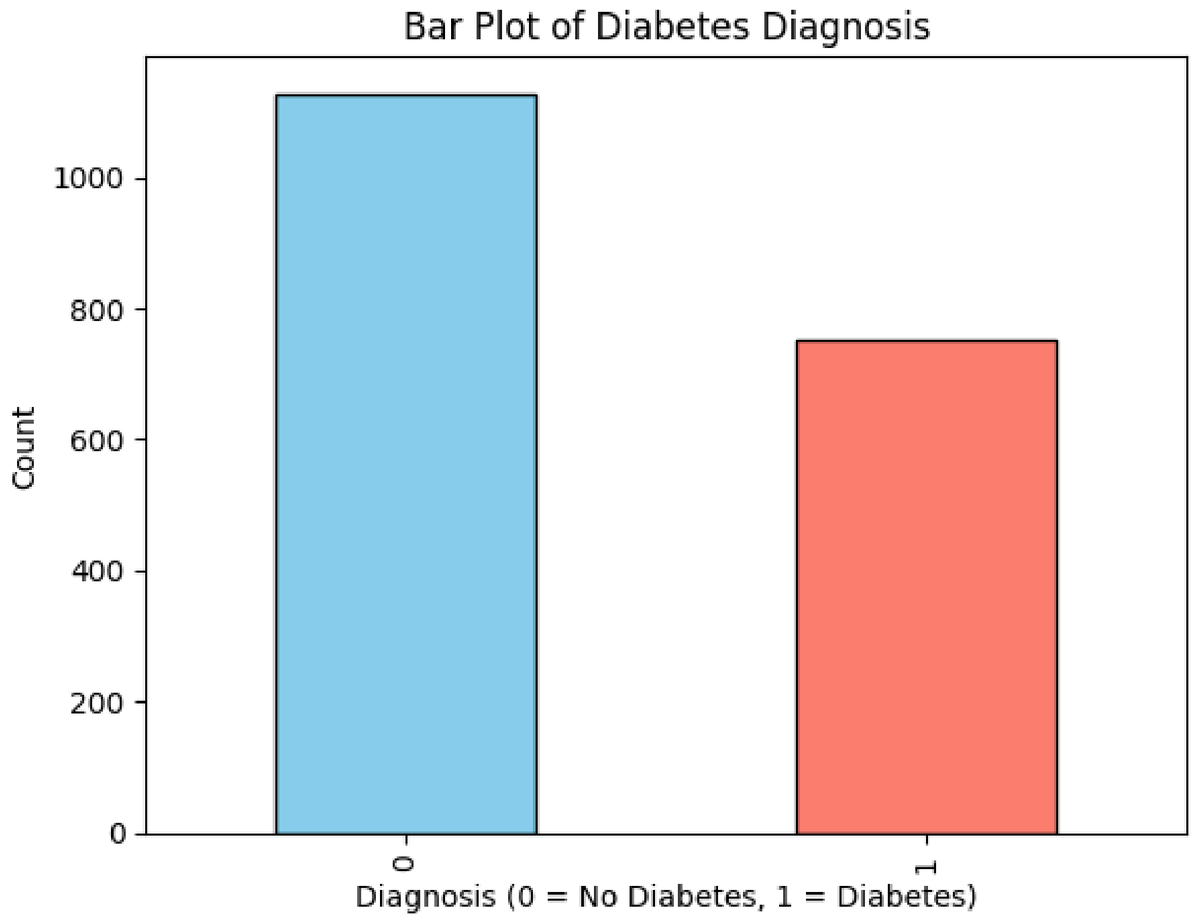
Data graphs


Histogram of Age
Frequency
0
20
40
60
80
20
30
40
50
60
70
80
90
Age

Data graphs

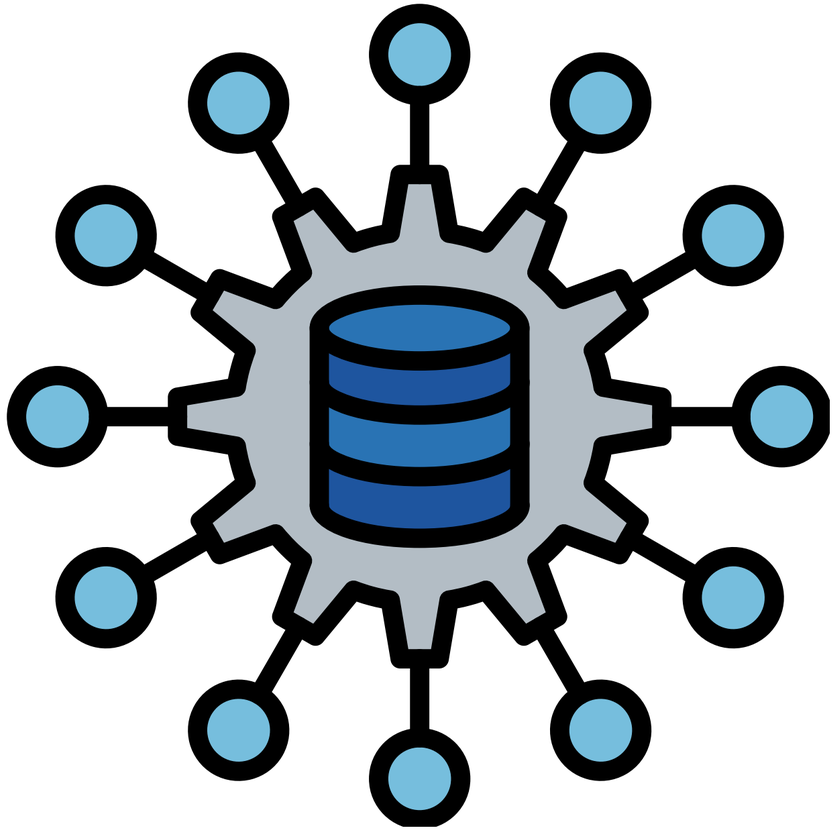

Scatter Plot: Age vs Fasting Blood Sugar
Data points
Trend line
Fasting Blood Sugar
200
180
160
140
120
100
80
20
30
40
50
60
70
80
90
Age

Data graphs


Bar Plot of Diabetes Diagnosis
Count
1000
800
600
400
200
0
0
1
Diagnosis (0 = No Diabetes, 1 = Diabetes)

Data pre-processing

Data preprocessing is a crucial step in any data mining project because real world datasets often contain noise, missing values, duplicates, or inconsistent formats. By cleaning and preparing the data before applying machine learning techniques, we ensure that the model learns from accurate and reliable information.



Data pre-processing

In this project, preprocessing helped improve the quality of the diabetes dataset by handling:

Data Cleaning

Data Transformation

Feature selection

Data cleaning

- **Missing Values:**

The dataset was examined using **df.isnull()** and no missing values were found, so no imputation was required.

- **Outlier Detection:**

Outliers were checked using the **Five-Number Summary, Boxplots, and the IQR method** → No significant outliers were detected in the 1879 records.

- **Initial Distribution Analysis:**

Histograms, scatter plots, and pie charts were used to confirm that the data distribution was stable before further

Data Transformation

- **Normalization:**

Numerical attributes (blood pressure, cholesterol, blood sugar, lifestyle scores) were scaled to [0,1] using Min-Max Scaling. This prevents any single feature from dominating the model and ensures fair contribution from all features for accurate predictions.

- **Discretization:**

Continuous features like *Age and BMI* were grouped into meaningful categories. Age :Children, Young Adults, Older Adults, Seniors. BMI : Underweight, Normal, Overweight, Obese. This simplifies analysis and highlights clear trends.

Feature Selection

Type of Feature Selection:

We used a Filter Method because it is simple, fast, and evaluates each feature independently from the model using statistical measures.

Top 10 Most Important Features:

These features were selected as they are most correlated with diabetes and have the highest Spearman correlation with the target variable meaning they have the strongest impact on predicting the disease.

```
Top 10 features selected by the Filter Method (Spearman Correlation):
FastingBloodSugar: 0.4495
HbA1c: 0.4256
FrequentUrination: 0.1515
Hypertension: 0.1319
ExcessiveThirst: 0.0735
UnexplainedWeightLoss: 0.0610
DiastolicBP: 0.0552
Smoking: 0.0538
SystolicBP: -0.0520
FamilyHistoryDiabetes: 0.0477
```

Feature Selection

Technique Used:
The Spearman Rank Correlation (Filter Method) was applied to measure the strength of the relationship between each feature and the target (Diagnosis).

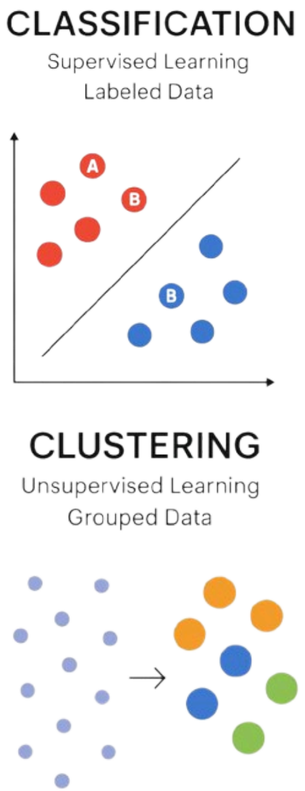
Benefits:

- Reduces model complexity
- Improves accuracy and performance
- Focuses on the most influential features

Data Mining Technique

In this project, we applied two main data mining techniques to analyze and understand the diabetes dataset.

We used **Classification** because our dataset includes labeled outcomes (Diabetic = 1, Non-diabetic = 0). This technique allows the model to learn from known labels and predict the class of new patients.



We used **Clustering** to explore the hidden structure within the data.

Since Clustering does not rely on labels, it helps reveal natural patient groups based on similarities in their medical attributes.

Data Mining Technique

Why Both?

- **Classification** gives us accurate predictions.
- **Clustering** gives us a better understanding of how patients are grouped.

Which Technique Fits Our Data Best?

Because our dataset contains clear labels, Classification is the most suitable technique for prediction.

However, Clustering still adds valuable insights by showing hidden patterns that may not be obvious from labels alone.

Classification

Decision Tree Classifier was used to predict whether a patient is diabetic or non-diabetic based on medical attributes.

Two methodologies with different splitting criteria were applied:

Decision Tree Classifier
(Entropy – Information Gain)

Decision Tree Classifier
(Gini Index)

Decision Tree Classifier (Entropy – Information Gain)

Entropy based classification measures the amount of uncertainty or impurity in the data.
In a Decision Tree, the model selects the feature that provides the highest Information Gain, meaning the feature that reduces uncertainty the most after splitting.

Different splits allowed us to observe how the model behaves with different amounts of training data and how stable its predictions are across various evaluation scenarios:

80% Training
20% Testing

75% Training
25% Testing

70% Training
30% Testing

Decision Tree Classifier (Gini Index)

The Gini Index is a measure of impurity used by Decision Trees to determine how well a feature can split the data.
It evaluates how “pure” the groups become after each split lower Gini values indicate better separation between classes.

To evaluate the model fairly and study its consistency, the Gini based Decision Tree was tested using the same three data splits:

80% Training
20% Testing

75% Training
25% Testing

70% Training
30% Testing

Findings

We evaluated our model by calculating these measures:

Decision Tree Results (Entropy – Information Gain)

Metric	80% train / 20% test	75% train / 25% test	70% train / 30% test
Accuracy	**0.9202**	**0.8957**	**0.9043**
Error Rate	0.0798	0.1043	0.0957
Sensitivity (Recall)	0.8933	0.8880	0.8933
Specificity	0.9381	0.9007	0.9112
Precision	0.9054	0.8564	0.8707

Decision Tree Results (Gini Index)

Metric	80% train / 20% test	75% train / 25% test	70% train / 30% test
Accuracy	**0.9096**	**0.9000**	**0.9060**
Error Rate	0.0904	0.1000	0.0940
Sensitivity (Recall)	0.8667	0.8989	0.9115
Specificity	0.9381	0.9007	0.9024
Precision	0.9028	0.8579	0.8619

Clustering

- Clustering was applied to explore natural patterns in the patient data and to group individuals with similar health characteristics without using the diabetes label

- We tested multiple clustering models and evaluated them using different performance measures to identify the most meaningful grouping structure in the dataset

K-Means Clustering

K-Means clustering groups patients by placing them into K clusters based on similarity in their attributes.

The algorithm assigns each patient to the nearest cluster center, then recalculates the centers until the clusters stabilize and no further changes occur

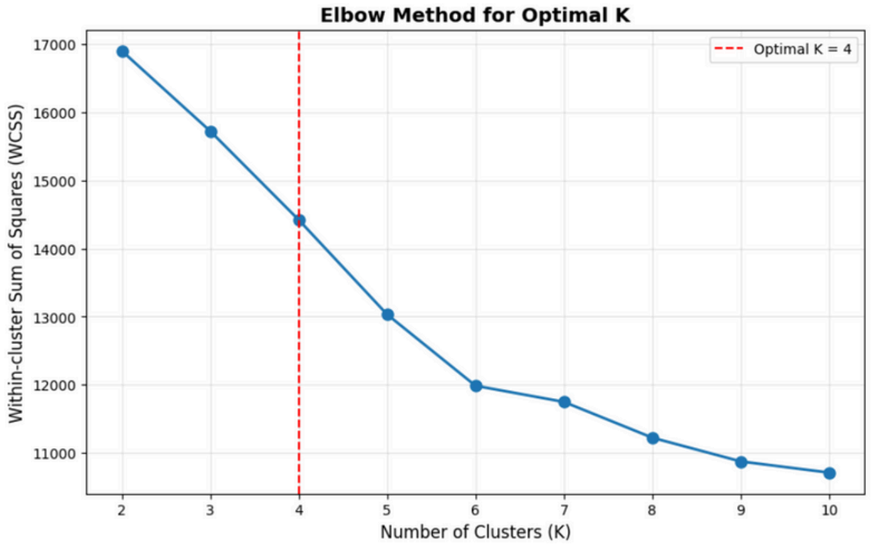
K = 2 Clusters

K = 3 Clusters

K = 4 Clusters

Findings

To evaluate the performance of each model, we calculated the average silhouette width and examined the Elbow Method



Optimal K based on Elbow Method: 4

K = 2 provided the strongest and clearest separation between clusters, with the highest silhouette score and the least overlap

Methods Comparison and Evaluation

- Decision Tree achieved 89–92% accuracy, with Entropy showing higher sensitivity. Glucose is the key factor, providing clear and interpretable predictions.
- K-Means identified 2 clusters (low- and high-risk) using Glucose, BMI, and Age, highlighting patterns and risk segmentation rather than prediction.
- Decision Tree excels in classification; K-Means reveals patient group patterns.

Final Conclusion

- **Decision Tree → accurate & interpretable predictions**
  Reliable for direct diabetes detection.

- **K-Means → reveals natural patient groups**
  Helps understand patient risk segmentation.

- **Top features → Glucose, HbA1c, BMI**
  Key indicators for diabetes risk.

- **Combined approach → supports early diagnosis & risk assessment**
  Enhances decision-making and preventive care.

DO YOU HAVE ANY QUESTIONS ?

THANK YOU!

Team Members:

Noura Almuayli
Alanoud Alsanad
Remas Ayidh
Lubna Alqifari