SEGMENTATION & PREDICTION OF DIABETES RISK USING MACHINE LEARNING

Members

Abdulrahman Omar Abobakr Alamodi (AIU22102263) – Data Preprocessing
Mohamed Abdullahi Ali (AIU22102342) – Supervised Learning
Osamah Mohammed Mohammed Al-Rusabi (AIU22102229) – UL
Rabiul Islam (AIU22102258) – Report

INTRODUCTION

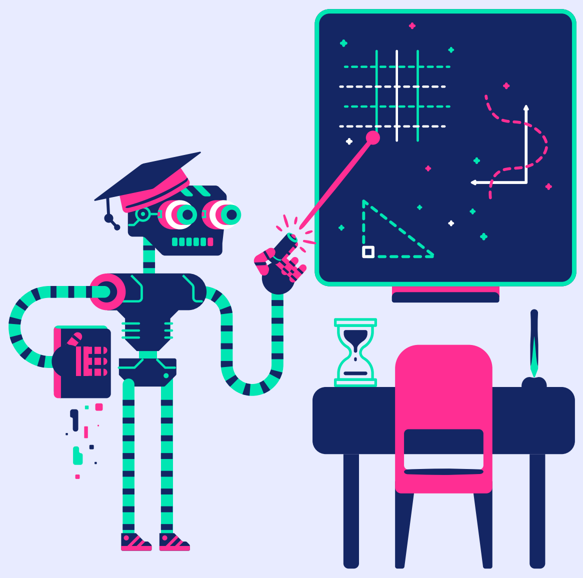
Machine Learning (ML) revolutionizes healthcare by enabling early disease detection. This project uses ML to classify diabetic patients and segment heart disease groups for targeted interventions.

PROBLEM STATEMENT

In the current healthcare sector, early diagnosis and personalized treatment are paramount in minimizing the effects of chronic diseases. Diabetes and heart disease are two of the most common and potentially fatal diseases globally, causing millions of avoidable deaths worldwide each year. However, early diagnosis and personalized treatment are still a mirage, given the nature of the symptoms, variability among patients, and scale of patient data.

OBJECTIVES

- Use K-Means for heart patient clustering.
- Apply LR, DT, RF, SVM for diabetes prediction.
- Evaluate models using Accuracy, Precision, Recall, F1, ROC-AUC.
- Perform feature analysis using RF & SHAP.
- Use visual aids (PCA, confusion matrix, bar charts).



PREPROCESSING

Missing Values: Median imputation (for 0-values in clinical data).
Scaling: StandardScaler used for both datasets.
Encoding: One-Hot Encoding for categorical features.
Feature Engineering: AgeGroup, BMICategory added.
Split: 80% training, 20% testing.

UNSUPERVISED LEARNING

Method: K-Means
Dataset: Heart Disease Dataset
Tools: PCA (for dimensionality reduction), Elbow & Silhouette Method (k=3)

SUPERVISED LEARNING

Dataset: Diabetes Dataset

Models

- Logistic Regression
- Decision Tree
- Random Forest
- SVM

Evaluation Metrics

- Accuracy,
- Precision,
- Recall,
- F1-Score
- ROC-AUC

RESULT AND ANALAYSIS

RESULT FOR SL

Model	Accuracy	Precision	recall	f1-score
Logistic regression	0.77	0.69	0.64	0.83
Decision tree	0.74	0.61	0.75	0.74
Random Forest	0.78	0.68	0.73	0.85
SVM	0.77	0.73	0.57	0.71

RESULT FOR UL

claster	Silhouette Score
2	0.29
3	0.28
4	0.22
8	0.21
10	0.21

ANALYSIS OF SL

In all four models, Random Forest outperformed the others on all evaluation metrics. It recorded the highest accuracy, precision, recall, and ROC-AUC scores. This would imply that Random Forest did not only accurately predict the majority of cases but also handled imbalanced data and high-order feature interactions better

ANALYSIS OF UL

The Elbow Method is to plot the within-cluster sum of squares (inertia) for k values ranging from 1 to 10. An "elbow" in the plot, where the slope of the curve suddenly levels off, typically indicates the optimal k. In this case, the elbow occurred at k=3, which indicated that three clusters were a reasonable balance between underfitting and overfitting.The Silhouette Score, which measures how similar a data point is to its own cluster compared to others

CONCLUSION

The project was successfully conducted to investigate the use of unsupervised and supervised machine learning techniques to solve two important healthcare concerns: patient segmentation and disease forecasting. Using medical information on heart disease and diabetes, we constructed end-to-end pipelines that uncovered previously unknown trends in patient groups and accurately predicted diabetes risk. This work not only demonstrates the clinic translational relevance of machine learning but also presents opportunities for improving healthcare decision-making and prevention medicine.