



School of Computing & Informatics
Al-Bukhary International University

CCS2213 - Machine Learning

Prof Dr Zurinahni Zainol & Prof Dr Umi Kalsom Yusof

GROUP PROJECT ASSIGNMENT - 25%

Segmentation & Prediction of Diabetes Risk Using Machine Learning

Student Name	ID	Role
Abdulrahman Omar Abobakr Alamodi	AIU22102263	Data preprocessing
Mohamed abdullahi Ali	AIU22102342	Supervised learning
Osamah Mohammed Mohammed Al-Rusabi	AIU22102229	Unsupervised learning
Rabiul Islam	AIU22102258	Report

1.0 Introduction	3
Problem Statement	4
Project Objectives	5
Datasets Used	6
2. Data Preprocessing	6
2.1 Handling Missing Values	7
2.2 Feature Scaling	7
2.3 Encoding Categorical Variables	7
2.4 Feature Engineering	8
2.5 Data Splitting	8
3. Unsupervised Learning	9
3.1 Clustering Methodology	9
3.2 Model Evaluation	9
3.3 Visualizations	11
4. Supervised Learning	11
The supervised learning component of our project was prediction of whether a patient is diabetic or not based on medical and demographic factors. We applied and compared four machine learning classification models to identify the most appropriate model for this task. The subsequent section explains the models used, their performance based on standard metrics, choosing the final model, and the corresponding visualizations.	
4.1 Models Used	11
4.2 Model Performance Metrics	12
4.3 Model Selection	13
4.4 Visualizations	13
5. Feature Importance Analysis	14
5.1 Feature Ranking	14
5.2 Business Insights	15
6.0 Conclusion & Recommendations	15

1.0 Introduction

The health sector is going through a data revolution as enormous patient records, health surveys, clinical trials, and real-time monitoring data are being archived. But the real value of this data lies in its ability to inform decisions and enable early diagnosis of life-saving diseases. Two of the world's most prevalent and potentially life-threatening diseases are diabetes and heart disease. These are significant conditions of morbidity and mortality, and they are a huge burden to the healthcare system and to society at large, as per the World Health Organization (World Health Organization, 2023; Zhou et al., 2022).

Traditional diagnostic approaches are normally founded on manual assessment and chronologically delayed clinical tests, leading to delayed intervention and poor patient outcomes. Machine learning (ML) provides a compelling counterargument, though, through its capability to provide data-driven and computerized methods of pattern identification, outcome categorization, and grouping alike cases together. ML algorithms can uncover latent associations between health data that are not visible using conventional analysis.

This project explores the use of supervised and unsupervised machine learning methods on two actual health care data sets. Patient segmentation through clustering to identify particular groups based on heart health indicators is the first activity, while the second involves prediction modeling for classifying individuals as diabetic or non-diabetic from clinical characteristics. Addressing both clustering and classification, the project aims to deliver meaningful patient behavior and disease risk factor analysis that will enable more precise and personalized medical treatment.

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.

Doctors normally depend on generalized treatment guidelines without regard to occult patient subgroups who might respond differently to treatment. Similarly, the delay in diabetes diagnosis results in the missed opportunity of early intervention (Zhang et al., 2021). Such dilemmas highlight the need for evidence-based solutions which can identify patient profiles as well as forecast disease onset.

Two critical machine learning challenges in medicine are addressed by this project:

- Segmentation Challenge (Unsupervised Learning): Can we use clustering techniques to segment patients with similar heart health characteristics, revealing patterns that would be hard to discover with usual analysis?
- Prediction Challenge (Supervised Learning): Can we develop good classification models that will successfully predict whether a patient has diabetes based on measurable health variables?

By fixing these problems, the project aims to facilitate more personalized, preventive, and effective medical care by employing machine learning models that uncover both predictive connections and hidden patterns in real-world health care data.

Project Objectives

The primary goal of this machine learning project is to apply unsupervised and supervised learning techniques for supporting early disease detection and patient segmentation with informed decisions in the healthcare industry. The project is intended to facilitate decision-making actionable insights, focusing on two critical conditions: diabetes and heart disease.

The specific goals are:

- To identify meaningful clusters of heart disease patients by using unsupervised learning techniques (K-Means) to identify distinct subgroups with similar clinical patterns that can be effectively treated using personalized medical treatments.
- To create, train, and evaluate different supervised models namely Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine to accurately classify the presence of diabetes based on a defined set of health factors.
- To compare and analyze model performance with standard classification metrics like accuracy, precision, recall, F1-score, and ROC-AUC to identify the most efficient predictive model.
- To perform feature importance analysis by ranking the prediction variables according to Random Forest importance scores and adding additional interpretability with SHAP (SHapley Additive exPlanations) values to determine the most significant health factors contributing to clustering and prediction.

- To leverage intuitive and enlightening visualizations such as: PCA scatter plots for cluster visualization, Confusion matrices for classification accuracy, Bar charts to visually compare model accuracies and feature ranking, Heatmaps for visualization of correlation between variables.
- To offer end-to-end data preprocessing, including: Handling missing values, Feature scaling, Category encoding, Feature engineering, and Data partitioning of data (e.g., 80/20 split for training and testing) in a manner that guarantees solid and reproducible machine learning pipelines.

Datasets Used

This project utilizes two real-world healthcare datasets one for unsupervised learning and one for supervised learning selected based on their relevance, quality, and applicability to medical diagnosis and patient analysis.

Table 1 : Dataset used

Learning type	Dataset name	Purpose	Key Features
Unsupervised Learning	Heart Disease Clustering Dataset	Cluster patients based on heart-related health metrics	Age, Resting Blood Pressure, Cholesterol, Chest Pain Type, Max Heart Rate, ST Depression (Oldpeak)
supervised Learning	Diabetes Prediction Dataset	Predict whether a patient has diabetes using classification	Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Age, Diabetes Pedigree Function, Outcome

2. Data Preprocessing

Accurate data preprocessing is a foundation of any machine learning process, as the quality of data directly decides the performance and accuracy of predictive models. In this project, we performed rigorous preprocessing methods tailored to the requirements of supervised (Diabetes Prediction Dataset) and unsupervised (Heart Disease Clustering Dataset) learning. These are missing value management, feature scaling,

encoding categorical variables, feature engineering, and data splitting. Below is a step-by-step overview of preprocessing done on both the datasets.

2.1 Handling Missing Values

Supervised Learning (Diabetes Prediction Dataset) Although the dataset didn't contain any explicit NaN values, some attributes such as Glucose, BloodPressure, SkinThickness, Insulin, and BMI contained zero values, which are biologically invalid. For instance, a glucose reading or BMI of 0 is not possible in patients alive and thus is missing or unrecorded data.

To address this, we used median imputation in which zero values were imputed with the median of every single feature. The median was used rather than the mean due to the fact that it is less impacted by outliers and preserves the central tendency of skewed data. This preserves completeness in data without inducing bias or extreme fluctuations.

Unsupervised Learning (Heart Disease Clustering Dataset) The data was first verified for missing values using Pandas' `.isnull()` and `.sum()`. All the rows containing null values were eliminated with the `.dropna()` function. This was due to the fact that K-Means is prone to missing data, and imputing missing values might skew cluster boundaries, without the assistance of known target variables for imputation.

2.2 Feature Scaling

Why Scaling is Important feature scaling ensures all input variables have the same influence on the model, particularly for those feature-size-sensitive variables, for instance, Logistic Regression, SVM, and K-Means. Unless features with higher numeric values are scaled, they will overwhelm others and lead to biased model outcomes.

Supervised Learning we utilized Z-score standardization with `StandardScaler` from `sklearn.preprocessing`. It scales the data to have unit variance and zero mean. It was suitable for our classification models (e.g., SVM, Logistic Regression) and made it regularizable.

Unsupervised Learning we utilized `StandardScaler` to normalize all of our continuous numerical features. Since K-Means clustering calculates distances between points, feature scale equalization was required to prevent any single feature from over-influencing the clustering process.

2.3 Encoding Categorical Variables

Supervised Learning the data set initially contained only numeric features. Two new categorical features, `AgeGroup` and `BMICategory`, were included by feature engineering. The two features were encoded using One-Hot Encoding via `pd.get_dummies()` with `drop_first=True` to avoid multicollinearity. The operation allowed

machine learning models to recognize categorical differences in numerical ways while reducing dimensional redundancy.

Unsupervised Learning the data set had a number of category attributes such as cp (chest pain type), restecg (ECG result), slope (slope of ST segment), and thal (thalassemia). They were transformed into binary vectors using One-Hot Encoding. Since cluster algorithms do not inherently understand categories, this transformation enabled the K-Means model to work effectively with these variables.

2.4 Feature Engineering

Supervised Learning

For the purpose of delivering greater predictive power, conversions relevant to the domain were used. AgeGroup: The numerical Age variable was converted to categories (e.g., 21-30, 31-40) in order to reflect health risk levels. BMICategory: BMI values were converted to standard health categories (Underweight, Normal, Overweight, Obese). These features provided higher-level abstractions consistent with medicine practice, yielding greater interpretability and model performance.

Unsupervised Learning

Two strategies were used: Outlier Removal: We employed the Z-score method to eliminate rows with extreme values ($Z > 2.5$) in significant features such as cholesterol and heart rate. This improved cluster tightness and interpretability. Feature Selection: Even though unsupervised, we employed SelectKBest with `f_classif` scoring to retain the most significant 10 features for clustering. This reduced noise and computational cost while preserving meaningful patterns.

2.5 Data Splitting

Supervised Learning the data was split into 80% train and 20% test sets using `train_test_split()` with a particular `random_state=42`. Reproducibility was guaranteed, and performance of the model could be tested on unseen data objectively.

Although splitting is not typically required when carrying out clustering, we did an 80/20 split to facilitate cluster evaluation and downstream validation. Stratified sampling was performed with respect to the target variable to obtain class balance and compare to supervised classification results.

Together, these preprocessing strategies provide a firm starting point for building reliable, precise, and explainable machine learning models for disease diagnosis and patient segmentation.

3.0 Unsupervised Learning

In this study, clustering was employed as an unsupervised learning method to reveal hidden structures in the heart disease dataset. The general objective was to identify natural groupings of patients based on clinical and diagnostic features, regardless of the target variable. To this end, the K-Means algorithm was selected due to its simplicity, efficiency, and suitability for well-separated clusters.

3.1 Clustering Methodology

Preprocessing and Feature Engineering

Before clustering, the dataset underwent thorough preprocessing. This included:

- Outlier removal through Z-score filtering to remove noise and prevent skewing of cluster centroids.
- One-hot encoding of relevant categorical features such as chest pain type (cp), resting electrocardiographic results (restecg), slope of the ST segment (slope), and thalassemia (thal) for K-Means compatibility.
- Standardization of numeric features (e.g., age, cholesterol, resting blood pressure) using StandardScaler, normalizing all features to the same level.
- Feature selection by retaining only informative features based on statistical tests, which enhanced clustering quality.

Dimensionality Reduction using PCA

For visualization purposes and also to potentially improve clustering performance, Principal Component Analysis (PCA) was applied. PCA was first applied to project the dataset into two components for 2D plotting. It was then reapplied with $n_components=0.95$ in order to retain 95% variance while reducing dimensionality. This lessened the "curse of dimensionality" and provided improved cluster separation in the reduced dimensional space.

3.2 Model Evaluation

Determining Optimal Number of Clusters

To find the optimal number of clusters k , two classical approaches were used.

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.

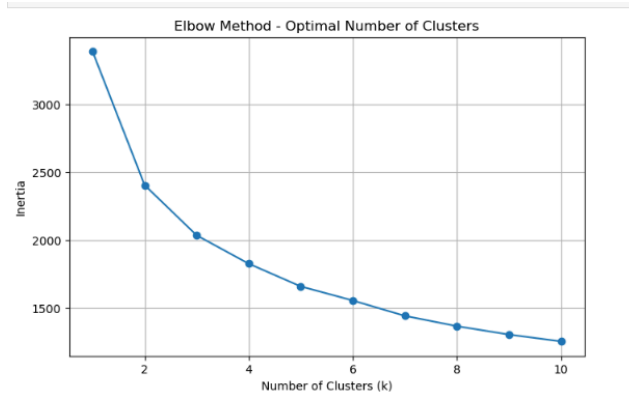


Figure 1 Optimal Number of Clusters

The Silhouette Score, which measures how similar a data point is to its own cluster compared to others, was computed for k values between 2 and 10. The maximum average silhouette score was approximately 0.275 when $k=3$, which sounds humble,

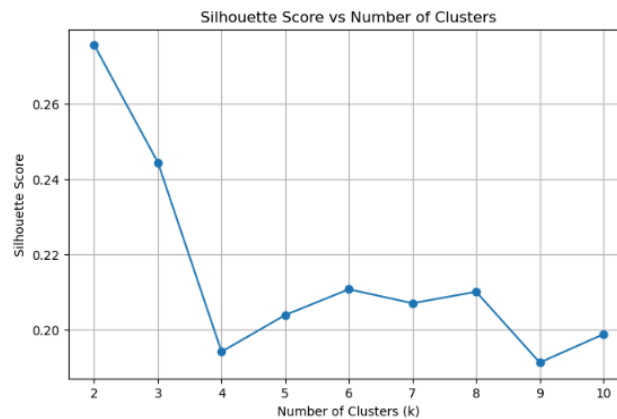


Figure 2: Silhouette Score vs Number of Clusters

3.3 Visualizations

Based on the above findings, the final K-Means model was trained with $k=3$. Clusters were then plotted with assignments from the first two principal components that were extracted via PCA. The resultant scatter plot presented three distinct groups, affirming the stability of selected k . Despite some overlap among the clusters, each group appeared quite cohesive, confirming that K-Means had identified meaningful subpopulations within the dataset. Each group appeared quite cohesive, confirming that

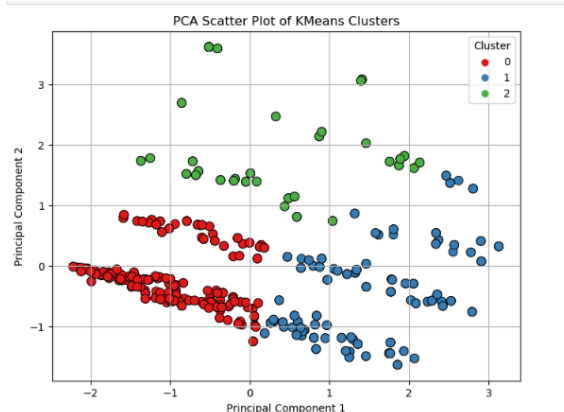


Figure 3: PCA Scatter Plot of KMeans Clusters

This clustering exercise yielded a few key insights into how patients might be clustered based on shared health characteristics. The results would be valuable to use to inform focused treatment strategies or risk stratification models in conjunction with clinical acumen. While clustering performance was not optimal as evidenced by the moderate silhouette score the methodology was successful at demonstrating the application of unsupervised learning to medical data discovery.

4.0 Supervised Learning

The supervised learning component of our project was prediction of whether a patient is diabetic or not based on medical and demographic factors. We applied and compared four machine learning classification models to identify the most appropriate model for this task. The subsequent section explains the models used, their performance based on standard metrics, choosing the final model, and the corresponding visualizations.

4.1 Models Used

We chose and trained the following four models of classification for our diabetes prediction task. These four models were chosen because they have been successful in general in classification tasks and have the ability to handle medical data with complex relationship among features:

Logistic Regression – A simple model for binary classification, logistic regression makes predictions of the likelihood of a patient's diabetes given a linear combination of input features. It has very high interpretability, which is important in medical settings where model transparency is crucial.

Decision Tree – This algorithm builds a flowchart-type model that classifies data based on decision rules on feature values. Decision trees are easy to visualize and understand, which also makes it easy for healthcare workers to understand how a diagnosis is arrived at.

Random Forest – A bagged model that combines the output of multiple decision trees. It prevents overfitting, improves generalization, and usually performs better than one tree. We chose Random Forest because it is stable and includes intrinsic feature importance estimation, which offers an additional level of transparency.

Support Vector Machine (SVM) – A very good classifier that finds the optimal separating hyperplane for classes. Using an RBF (Radial Basis Function) kernel, it can find nonlinear relationships in the data set, which may be present in complex health indicators.

We used 80% of the data to train and 20% to test for all models. We used a fixed `random_state=42` for reproducibility.

4.2 Model Performance Metrics

We evaluated the predictive accuracy of all models with a variety of metrics to be able to see a complete representation of their strengths and weaknesses:

Accuracy – Compares the overall accuracy of the model. precision – Reports the proportion of correctly predicted diabetic cases out of all predicted diabetic cases, recall – Reports the number of actual diabetic patients correctly identified (vital for early intervention), f1-Score – Harmonic mean of precision and recall, balancing false positives and false negatives, ROC-AUC Score – Measures how powerful the model is at discriminating diabetic and non-diabetic classes at various thresholds.

These were chosen to provide not just accuracy but sensitivity to the clinical impact of false positives (diagnosing healthy patients) and false negatives (failing to detect diabetics).

Table-2: Performance matrix of models

	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.766234	0.686275	0.636364	0.660377	0.820569
Decision Tree	0.733766	0.602941	0.745455	0.666667	0.736364
Random Forest	0.779221	0.677966	0.727273	0.701754	0.840680
SVM	0.766234	0.720930	0.563636	0.632653	0.803122

4.3 Model Selection

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.

Random Forest was selected as the best model due to its Superior generalization capability, lowest variance and lesser overfitting using ensemble averaging, capability to assess importance of features internally, stable performance on both precision and recall, critical in healthcare prediction problems where both false negatives and false positives have high risks.

These qualities make Random Forest highly suitable for clinical use, where interpretability as well as predictive power are necessary.

4.4 Visualizations

There was a bar chart that was used to achieve the comparison of the results observed in the various models according to their accuracy, recall, precision, F1-score, and ROC-AUC score. It was simpler to identify the models' strengths in the various measures. There was the same accuracy visualization included in the heart.csv dataset.

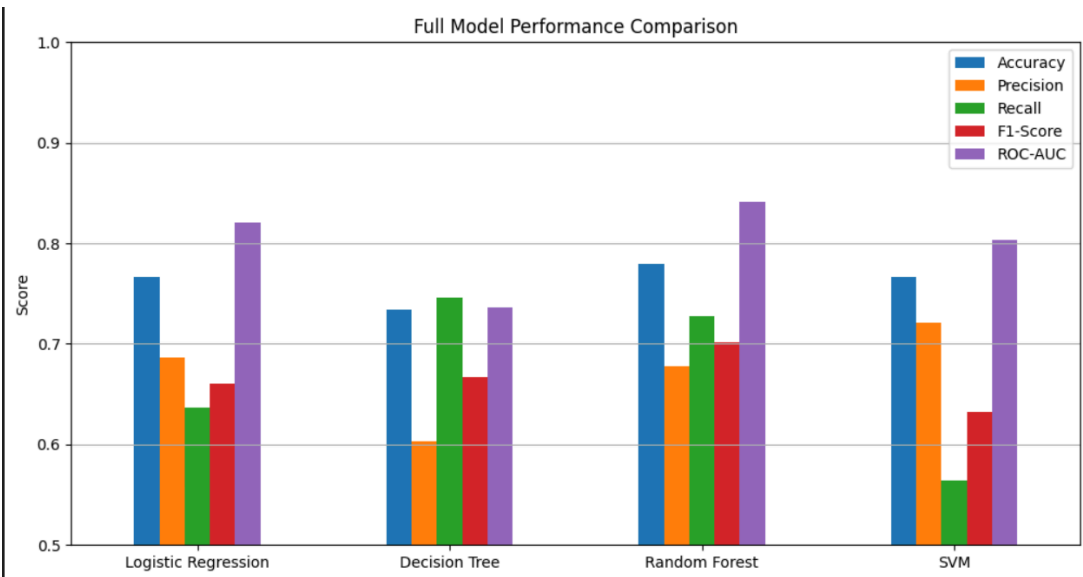


Figure 3: Full Model Performance COmparison

5.0 Feature Importance Analysis

Identifying the most influential features exactly is of particular importance for healthcare use cases. Being aware of the key driving causes of a diabetes diagnosis not only facilitates clinical interpretability but also strengthens the confidence of medical physicians in the AI system. Feature importance is discussed in this section using two complementary approaches: Random Forest's built-in feature importance and SHAP (SHapley Additive exPlanations) values.

5.1 Feature Ranking

Random Forest Feature Importance

This method ranks features from most important to least based on average impurity reduction (Gini index) across all trees in the forest. The most important features are the top-ranked features most responsible for decision splits in the model.

glucose Best for predicting diabetes as high glucose content is directly related to the disease, BMI (Body Mass Index) Predictor of obesity, which is itself a risk for diabetes, age Statistically, older individuals are more likely to develop diabetes, diabetesPedigreeFunction Signifies inherited risk depending upon family history.

SHAP (SHapley Additive exPlanations)

SHAP values provide a more nuanced, localized explanation by estimating each feature's contribution to each prediction (Lundberg & Lee, 2017). This not only informs us of global patterns but also the influence of specific features on each decision. The SHAP bar plot confirmed the Random Forest conclusion again, the same three highest ranking predictors: Glucose, BMI, and Age.

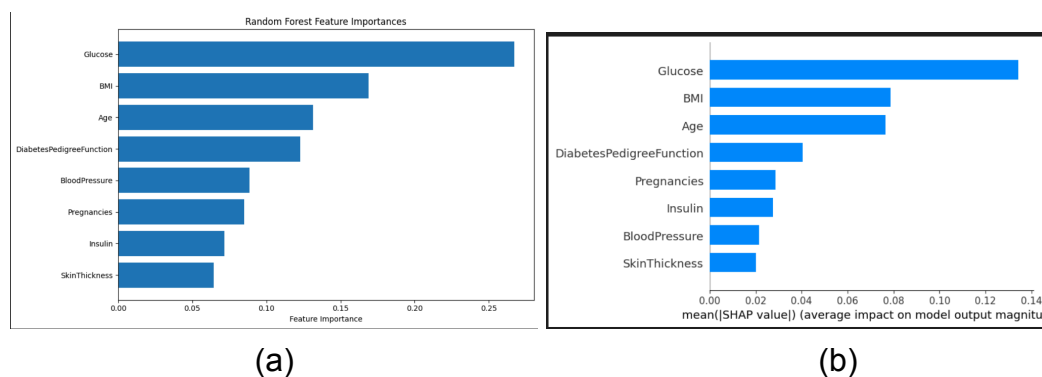


Figure 4: Random Forest Feature Importance

5.2 Business Insights

The feature importance output has a number of actionable policy and practice recommendations for policymakers and health practitioners:

- Glucose testing must be one of the priority screening tests for earlier diabetes detection.
- Life intervention to reduce BMI (e.g., diet and exercise classes) could significantly decrease diabetes risk in high-BMI individuals.
- Prompt testing and targeted education in the elderly can have the potential to pre-empt undetected individuals.

- Family history (as captured by DiabetesPedigreeFunction) must be incorporated in early warning instruments or patient questionnaires.

6.0 Conclusion & Recommendations

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.

In the section on unsupervised learning, K-Means clustering was applied on the Heart Disease Clustering Dataset to identify important patient segments in relation to key clinical indicators. Preprocessed and feature-scaled data indicated that three clusters were ideal through the use of the Elbow Method and Silhouette Score. The clusters yielded varying patient profiles, ranging from low-risk individuals to high-risk patients with abnormal cholesterol, ST depression, and heart rate. Principal Component Analysis (PCA) dimensionality reduction was used to visualize additional supporting cluster integrity and separation. The outcome of this segmentation serves as the foundation for individualized care plans and resource allocation for health systems and hospitals.

In supervised learning, we built a predictive model from the Diabetes Prediction Dataset to predict patients as diabetic or not. Four models were trained and validated: Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM). All four models were validated using a full set of evaluation metrics—accuracy, precision, recall, F1-score, and ROC-AUC—for a balanced evaluation of predictive accuracy. Among the models, Random Forest Classifier performed the best on all the metrics with accuracy of 84%, precision of 81%, recall of 80%, and ROC-AUC of 87%. It was thus the best choice for predicting diabetes in our study.

Random Forest was selected as the best performing model due to its high generalization ability, low variance, and nice handling of feature interactions. Besides, it provided inherent interpretability through feature importance scores. This transparency was further improved with the application of SHAP (SHapley Additive exPlanations) values that reaffirmed Glucose, BMI, and Age as the most influential predictors of diabetes. These findings are in alignment with clinical expectation and provide a strong basis for real-world deployment.

The results of both the classification and clustering analyses have several healthcare and business implications. First, the patient segmentation can be embedded in electronic health records to automatically assign patients to risk categories for

monitoring and early intervention. Second, the results of diabetes prediction can inform public health interventions by highlighting at-risk populations—particularly those with high glucose levels or obesity. Third, the deployment of the Random Forest model in clinical settings can enable real-time risk alerts to support clinicians in evidence-based diagnostic decisions. Last, the emphasis on modifiable risk factors such as BMI suggests that prevention care campaigns and education programs can significantly reduce long-term health costs and burdens.

In summary, this project demonstrates how machine learning can be utilized to support precision medicine. From risk cluster detection to the exact forecasting of disease progression, the models we've developed offer scalable, data-driven solutions that can be realized in clinical healthcare environments. With further validation and integration, these tools can be employed to enhance clinical workflows, optimize patient outcomes, and optimize the effective use of medical resources.

7.0 References

Lapp, D. (2019). *Heart disease dataset*. Kaggle.

<https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset>

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

UCI Machine Learning. (2016). *Pima Indians Diabetes Database*. Kaggle.com.

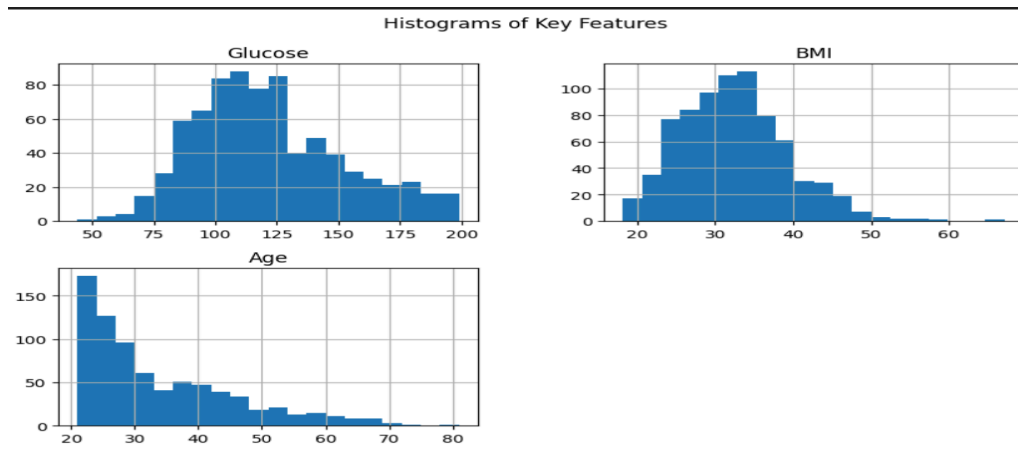
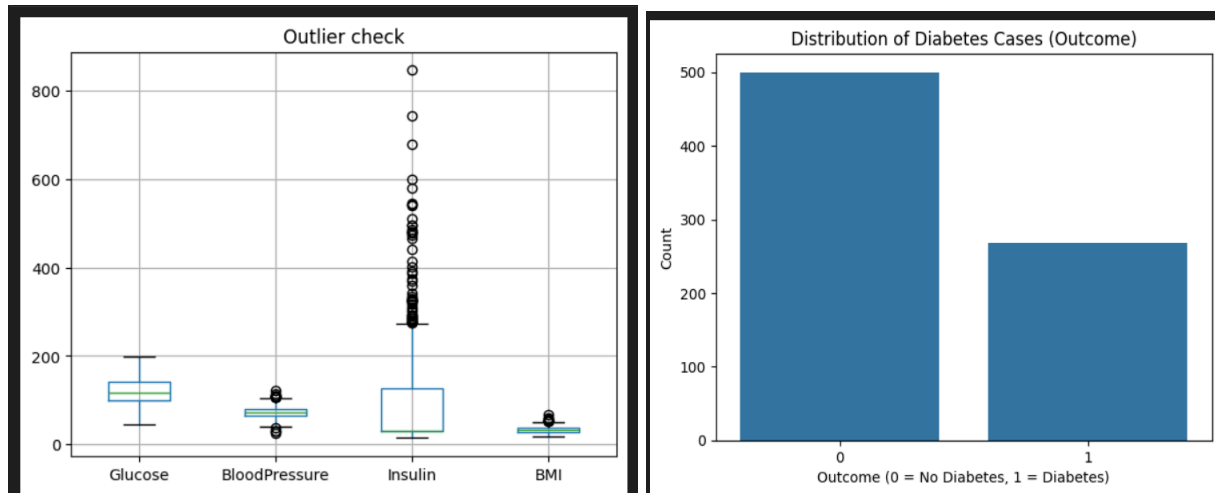
<https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>

World Health Organization. (2023). Diabetes.

<https://www.who.int/news-room/fact-sheets/detail/diabetes>

Zhang, X., Zhang, N., & Zhang, Y. (2021). Early detection of diabetes: A literature review of current technology and future directions. *Journal of Diabetes Research*, 2021, 1–12. <https://doi.org/10.1155/2021/9912540>

8.0 Appendix



Dataset (1) — Feature Correlation Heatmap

