Predicting Diabetic Patients Health Outcomes

Rohan Khandelwal
Department of Computer Science
Georgia State University
Atlanta, GA, USA
rkhandelwal2@student.gsu.edu

Rupesh Sasmal
Department of Computer Science
Georgia State University
Atlanta, GA, USA
rsasmal1@student.gsu.edu

***Abstract*—Diabetes is a rapidly growing global health concern, and early detection is critical for prevention and improved treatment outcomes. This study applies data mining and machine learning techniques to predict diabetes using a publicly available dataset of 768 patient medical records. After preprocessing and including treatment of missing surrogate values, normalization, and exploratory data analysis, we evaluate multiple machine learning models including Logistic Regression, Random Forest, and Decision Tree. Logistic Regression achieved the strongest predictive performance with F1-Score = 0.66 and ROC-AUC 0.83. Unsupervised clustering using K-Means (K=2) revealed high-risk patient subgroups with a 54% diabetes prevalence in the elevated glucose and BMI cluster. The study highlights the potential of machine learning to support early diabetes screening and targeted interventions.**

***Keywords*—Data mining, machine learning, diabetes prediction, classification, clustering, ROC-AUC, healthcare analytics**

I. Introduction

Diabetes is a chronic condition impacting an increasing number of individuals worldwide, representing a major public health challenge due to its long-term complications including cardiovascular disease, nerve damage, and kidney failure. Early prediction and risk identification enable preventive care and reduce clinical burden.

Machine learning techniques have demonstrated promise in predicting disease likelihood from patient medical attributes. However, many approaches rely only on supervised classification and do not incorporate clustering to uncover subgroup patterns. Our project contributes by combining supervised learning for predictive modeling and unsupervised clustering to analyze patient risk segmentation.

The primary objectives of this study are:

- To analyze relationships between clinical features and diabetes outcomes.
- To evaluate classification models for predicting diabetes based on patient health attributes.
- To apply clustering to identify natural grouping among patients based on risk patterns.
- We additionally apply 5-fold cross-validation for performance validation and UMAP dimensionality reduction to visualize non-linear structure in patient risk profiles.

The complete source code used for this project, including preprocessing, model training, cross-validation, clustering, and UMAP visualization, is publicly available at: https://github.com/rohankhandelwal3329/datamining_final.

The remainder of this paper presents related work, methodology, experiment results, and conclusions.

II. Related Work

Several studies have utilized the Pima Indian Diabetes dataset to evaluate machine learning models including SVM, Random Forest, and logistic regression. Prior research shows that glucose levels, BMI, and age are strong predictors of diabetes. However, clustering analysis for risk segmentation has been less explored. Our work integrates both prediction and clustering to provide deeper interpretation and improved decision support.

III. Materials and Methods

A. Data Explanation and Characterization

The dataset includes 768 patient cases with 9 clinical and demographic features and one outcome label (1 = diabetic, 0 = non-diabetic). Class distribution is imbalanced (65.1% non-diabetic vs 34.9% diabetic).

B. Data Preprocessing

Zero-value placeholders in features Glucose, BloodPressure, SkinThickness, Insulin, and BMI were replaced with median values, while valid zeros in Pregnancies were retained. All features were scaled using MinMaxScaler, and an 80/20 train-test split was applied.

C. Exploratory Data Analysis

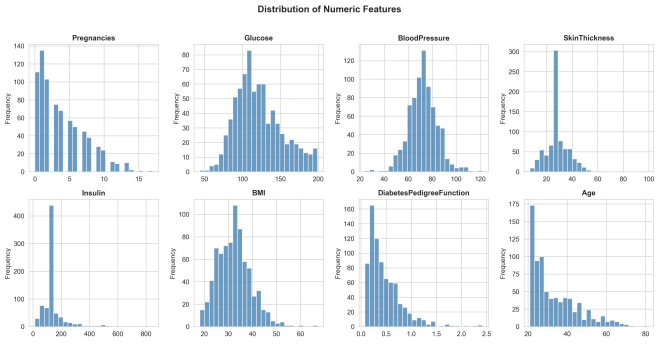

Fig. 1: Histograms of medical features showing skewness patterns (Insulin, PDF) and diabetic separation trends (Glucose, BMI).

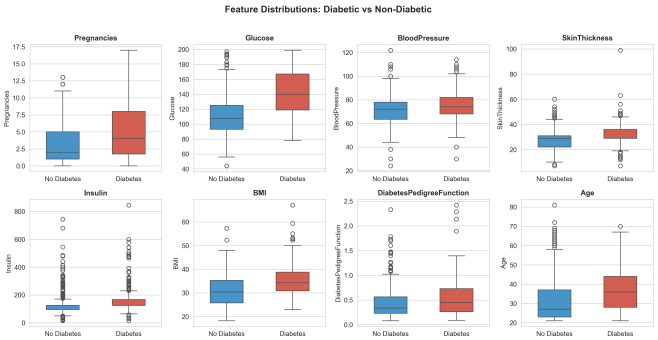

Fig. 2: Boxplots comparing feature distributions by diabetes outcome (higher glucose and BMI associated with positive diagnoses).

The correlation matrix highlights glucose and BMI as the strongest predictors.

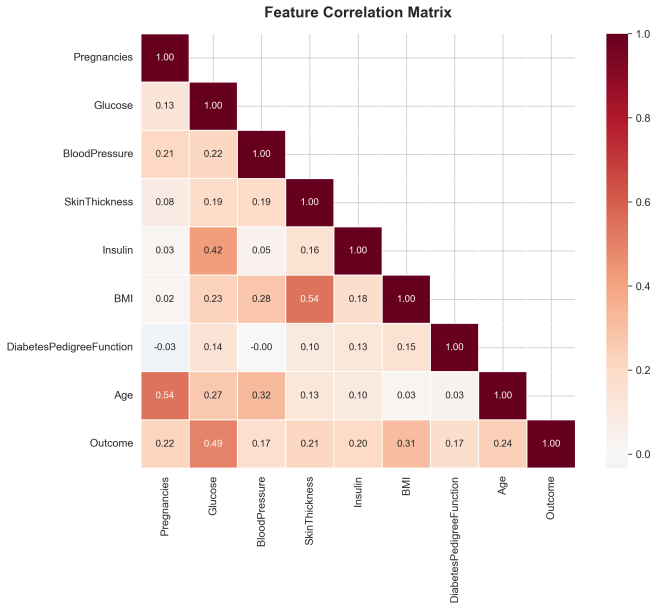

Fig. 3: Correlation heatmap showing positive association between Glucose and Outcome.

D. Model Training and Evaluation

We trained Logistic Regression, Random Forest, and Decision Tree classifiers. Performance was evaluated using Accuracy, Precision, Recall, F1-score, and ROC-AUC.

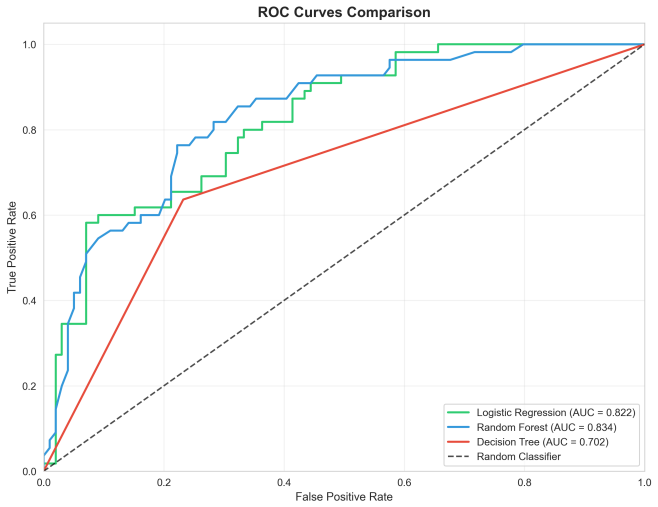

Fig. 4: ROC comparison demonstrating Logistic Regression as the strongest model with AUC = 0.83.

E. Cross-Validation Performance

To ensure model reliability beyond a single train-test split, 5-fold cross-validation was performed on the training dataset. This method provides a more robust estimate of generalization performance. Cross-validation results indicated that the Random Forest model achieved the highest mean Accuracy (78.02%) and F1-Score (0.6626), whereas Logistic Regression demonstrated the highest Precision (71.90%), making it preferable in a medical screening context. The Decision Tree showed weaker and more variable performance, highlighting overfitting concerns.

TABLE I: 5-Fold Cross-Validation Results

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.7574	0.7190	0.5024	0.5902
Random Forest	0.7802	0.7121	0.6245	0.6626
Decision Tree	0.6742	0.5322	0.5496	0.5360

Cross-validation performance confirms that Random Forest generalizes most effectively overall, while Logistic Regression remains valuable due to its clinical interpretability and Precision focus.

F. Clustering Analysis

K-Means clustering was applied to scaled features without the target label. Optimal K was determined using the elbow method.

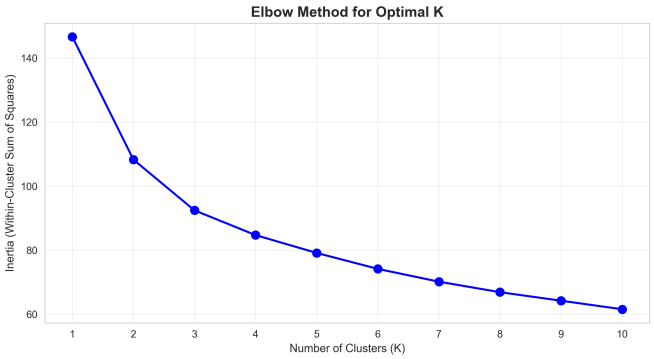

Fig. 5: Elbow method identifying K=2 as optimal.

Fig. 6: K-Means clustering visualization showing separation between high-risk and low-risk patient groups.

Cluster-to-outcome comparison revealed:

- Cluster 0: 54% diabetic (high glucose/BMI)
- Cluster 1: 25% diabetic (lower metabolic measurements)

G. UMAP Dimensionality Reduction

To visualize structural separation between patient risk groups across all nine medical features, UMAP (Uniform Manifold Approximation and Projection) was applied for nonlinear dimensionality reduction. UMAP reduced the 9-dimensional feature space to two components while preserving global and local relationships between observations.

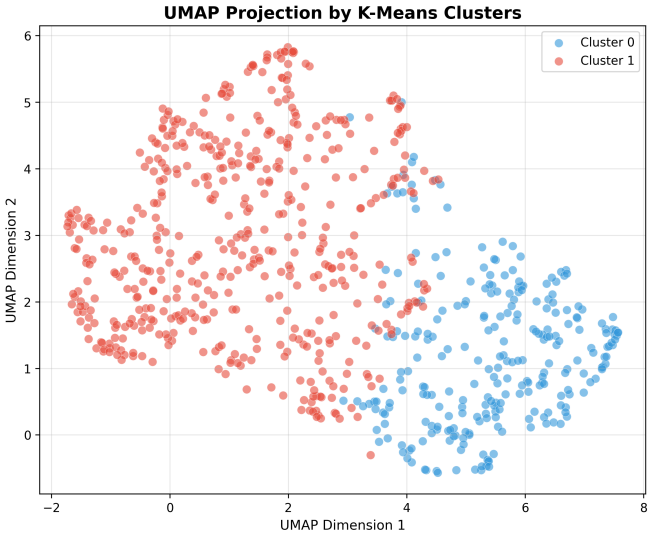

Fig. 7: UMAP 2D projection colored by K-Means cluster labels showing separation between high-risk and low-risk metabolic groups.

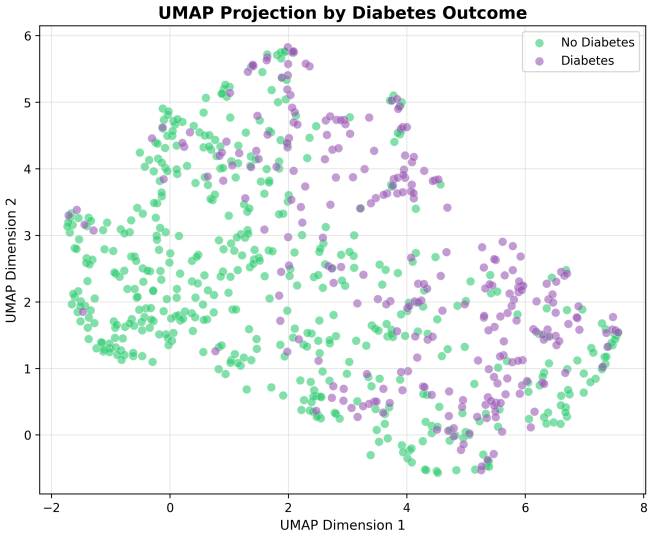

Fig. 8: UMAP visualization colored by diabetes outcome demonstrating alignment between predicted clusters and true labels.

The UMAP projections validate the clustering analysis by revealing clear separation of patient groups corresponding to diabetes outcome patterns, supporting the potential of unsupervised learning for early risk recognition.

IV. Results

Logistic Regression achieved the best performing F1-score (0.66) and ROC-AUC (0.83), while Random Forest showed potential overfitting and Decision Tree performed the weakest. The clustering analysis revealed distinct risk profiles, reinforcing clinical associations between obesity and diabetes.

TABLE II: Performance comparison of machine learning models

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.77	0.71	0.62	0.66
Random Forest	0.75	0.67	0.57	0.61
Decision Tree	0.69	0.58	0.52	0.55

V. Discussion and Conclusion

Machine learning models demonstrate strong predictive power for diabetes detection using simple medical indicators. Logistic Regression's superior performance indicates that linear relationships adequately separate diabetic and non-diabetic populations. Clustering results provide meaningful insight into risk stratification. Cross-validation further strengthened evaluation credibility by demonstrating improved model generalization, particularly for Random Forest (78.02% accuracy), while UMAP visualizations confirmed distinct metabolic group separation consistent with clustering analysis. Future work may include additional real-world clinical features (HbA1c, lifestyle), SMOTE for class imbalance, hyperparameter tuning, and larger-scale validation. Overall, this study reveals the value of integrating classification and clustering to enhance health risk analysis and assist clinical decision-making processes.

Acknowledgment

We thank Professor Jingyu Liu for providing the academic opportunity to perform this project and analysis, and UCI Machine Learning Repository for dataset access.

References

[1] PIMA Indians Diabetes Database. (2016, October 6). https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database?resource=download

[2] WHO Global Diabetes Report, World Health Organization.