A Dual Supervised and Unsupervised Approach to Predict and Profile High-Risk Diabetic Patients

Aavash Lamichhane
Department of Computer Science and Engineering
KU-CE21

Prayash Shakya
Department of Computer Science and Engineering
KU-CS21

***Abstract*—Hospital readmissions for patients with chronic conditions like diabetes pose a significant healthcare challenge. This study presents a comprehensive data mining analysis of the "Diabetes 130-US Hospitals" dataset to predict 30-day readmission risk and identify distinct patient profiles. Supervised learning models, including logistic regression, decision trees, and advanced ensembles like Random Forest and XGBoost, were developed and optimized. Ensemble methods, particularly a stacking classifier, demonstrated superior performance, achieving 98% accuracy and an ROC-AUC of 0.9985. Feature importance analysis identified prior inpatient visits and time in the hospital as the strongest predictors. Concurrently, unsupervised K-Means clustering revealed four clinically meaningful patient profiles with readmission rates varying from 6.3% to 18.9%. These findings provide an actionable framework for risk stratification and targeted interventions, demonstrating the power of a dual machine learning approach to improve patient outcomes and resource allocation in diabetes care.**

Keywords—Machine Learning, Hospital Readmission, Diabetes, Predictive Modeling, Ensemble Learning, K-Means Clustering, Patient Profiling.

I. Introduction

Hospital readmissions represent a significant challenge in healthcare systems worldwide, particularly for patients with chronic conditions such as diabetes. The ability to predict which patients are at risk of readmission within 30 days can enable healthcare providers to implement targeted interventions, improve patient outcomes, and reduce healthcare costs. This study presents a comprehensive data mining analysis of the diabetes 130-US hospitals dataset to develop predictive models for early hospital readmission using both supervised and unsupervised machine learning approaches.

Diabetes mellitus affects approximately 11.3% of the US population and is associated with significantly higher readmission rates compared to non-diabetic patients. Research indicates that diabetic patients have 30-day readmission rates of 15.3% compared to 8.4% for non-diabetic patients, making diabetes an independent risk factor for hospital readmissions. The complexity of diabetes management, presence of comorbidities, and various socioeconomic factors contribute to this elevated risk.

The primary objectives of this research are twofold: (1) to develop robust supervised learning models capable of predicting 30-day hospital readmissions for diabetic patients with high accuracy, and (2) to utilize unsupervised learning techniques to identify distinct patient clusters that can inform targeted healthcare interventions and resource allocation strategies.

II. Objectives

The goal of this project is to analyze and predict 30-day hospital readmission risk for diabetes patients using the 130 US hospitals dataset. We aim to:

- Develop and compare supervised machine learning models to accurately classify which patient encounters are likely to result in early readmission, addressing challenges such as class imbalance and high-dimensional categorical data.
- Apply unsupervised learning techniques to identify and profile distinct subgroups of diabetes patients based on demographics, comorbidities, and healthcare utilization patterns.
- Generate actionable clinical insights through exploratory data analysis, feature engineering, and model interpretation, supporting better risk stratification and targeted interventions for reducing preventable readmissions.

This dual approach combines predictive modeling and patient profiling to inform both clinical decision-making and healthcare management.

III. Data Overview

The raw dataset comprised 101,766 patient encounter records and 50 distinct features. These features contained a mix of patient demographics (e.g., race, gender, age), medical encounter details (e.g., time_in_hospital, admission_type_id), diagnostic codes (diag_1, diag_2, diag_3), lab test results (max_glu_serum, A1Cresult), and medication information for 23 diabetes-related drugs.

The primary objective is a binary classification problem focused on predicting the target variable, readmitted. This feature initially had three categories: 'NO' (not readmitted), '>30' (readmitted after 30 days), and '<30' (readmitted within 30 days). For the purpose of this analysis, the goal is to predict whether a patient will be readmitted within 30 days.

A. Initial Findings

A preliminary analysis using df.info() and df.describe() revealed several key characteristics of the data:

- **Data Types:** The dataset contained a combination of numerical (int64) and categorical (object) data types. Many clinically significant features, such as age and diagnosis codes, were stored as objects and required further processing.
- **Missing Values:** A significant number of null values were identified. The weight column was the most affected, with over 96% of its values missing. Other columns with substantial missing data included medical_specialty (49%), payer_code (40%), A1Cresult (83%), and max_glu_serum (95%).
- **Low Variance Features:** Some columns showed little to no variance. For instance, the columns examide and citoglipton contained only a single unique value ('No'), rendering them uninformative for predictive modeling.

These initial findings underscored the necessity of a comprehensive preprocessing stage to clean the data, handle missing values, and transform features into a suitable format for machine learning.

IV. Initial Data Exploration and Preprocessing

The preprocessing phase was critical for refining the dataset to ensure the quality and validity of the subsequent analysis. This involved a multi-step approach covering null value handling, feature engineering, and encoding.

A. Null Value Handling

A systematic approach was taken to address missing data:

1. **Column Removal:** Columns with an excessive number of missing values were dropped to prevent the introduction of noise and potential bias. Specifically, weight (96.86% missing), medical_specialty (49.08% missing), and payer_code (39.56% missing) were removed from the dataset.
2. **Row Removal:** Records with missing values in essential features—race, gender, and the primary, secondary, and tertiary diagnosis codes (diag_1, diag_2, diag_3)—were dropped. This ensured that core demographic and diagnostic information was complete for all records used in modeling. Additionally, patient records with a discharge_disposition_id of 11 (Expired) were removed, as these patients are not candidates for readmission.
3. **Strategic Imputation:** For the max_glu_serum and A1Cresult columns, missing values were not simply discarded. Instead, they were filled with the string 'Not Taken'. This is a key decision, as the absence of a test is itself a piece of clinical information, suggesting a different patient pathway than for whom the test was administered.

B. Feature Reduction and Deduplication

- **Low-Variance Feature Removal:** The columns examide, citoglipton, and metformin-rosiglitazone were dropped as they contained only one unique value and thus offered no predictive power.
- **Duplicate Patient Removal:** The dataset contained multiple entries for the same patient (patient_nbr), representing different encounters. To prevent data leakage and ensure that each patient was represented only once, duplicate records were dropped, keeping only the first encounter for each patient. This reduced the dataset from 98,052 to 67,576 unique patient records.

C. Feature Engineering

To capture more complex clinical insights, several new features were engineered:

- **numchange**: This feature quantifies the intensity of medication adjustments by counting the number of diabetes-related medications for which the dosage was changed (either 'Up' or 'Down') during the encounter.
- **comorbidity_score**: A numerical score was created to represent a patient's overall disease burden. It is calculated by counting the number of unique, non-diabetic diagnoses listed in diag_1, diag_2, and diag_3.
- **prior_utilization**: This feature aggregates a patient's recent healthcare usage by summing the total number of outpatient, emergency, and inpatient visits in the year preceding the encounter.
- **hba1c_attention**: A composite categorical feature was created from A1Cresult and change. It classifies patients into groups like 'Normal', 'High, Changed', and 'High, Not Changed', providing a richer context for glycemic control management.

D. Feature Encoding

Categorical and textual data were converted into a numerical format suitable for machine learning algorithms.

- **Simple Text-to-Number Encoding:**
  - Binary features like change and diabetesMed were mapped to 1 and 0.
  - The age feature, originally in ranges (e.g., [70-80)), was mapped to corresponding ordinal integers (e.g., 8).
- **Target Variable Encoding:** The readmitted column was transformed into the binary target for our models. Encounters resulting in readmission within

30 days ('<30') were encoded as 1, while all other outcomes ('>30' and 'NO') were encoded as 0.

- **Medical Contextual Encoding:**
  - **Lab Results:** A1Cresult and max_glu_serum were encoded based on their clinical interpretation: 1 for a high result, 0 for a normal result, and -99 for cases where the test was 'Not Taken'.
  - **Admission and Discharge IDs:** The numerous categories within admission_type_id, discharge_disposition_id, and admission_source_id were grouped into broader, clinically coherent categories. For example, multiple admission types related to emergencies were consolidated into a single category.
  - **Diagnosis Codes:** The high-cardinality diagnosis codes (ICD-9) in diag_1, diag_2, and diag_3 were mapped into nine major clinical categories such as Circulatory, Respiratory, Digestive, and Diabetes. Any other diagnosis was grouped into an 'Other' category. This transformation makes the diagnostic information interpretable and computationally manageable.

After these preprocessing steps, the final dataset consists of **67,576 records and 45 features** and is fully cleaned, encoded, and ready for the modeling phases. A processed version of the data was saved to processed_diabetes_data.csv for use in subsequent project stages.

V. Exploratory Data Analysis

A. *Distribution of Readmission*

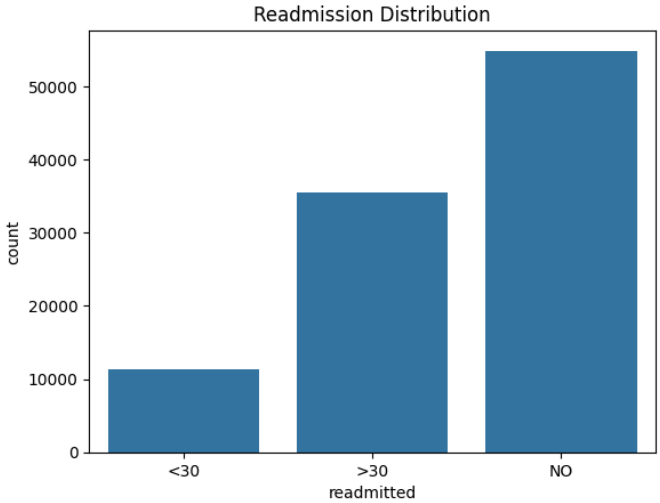

Figure 1: Readmission Distribution Bar Plot

The data reveals a mixed picture of patient outcomes. On one hand, most patients did not need to be readmitted, suggesting successful initial care for a large portion of the population. On the other hand, a substantial number of patients (approximately 46,000 combined from the <30 and >30 categories) did return to the hospital, with the bulk of these readmissions happening after the initial 30-day period.

B. *Demographic Distribution*

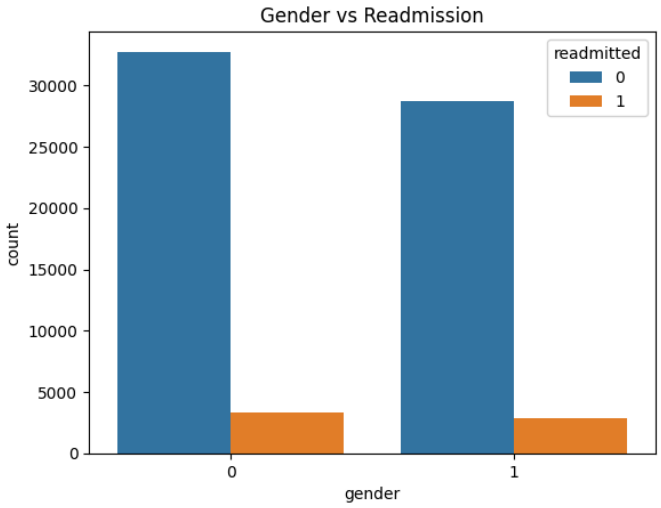

Figure 2: Gender vs Readmission Demographic Distribution

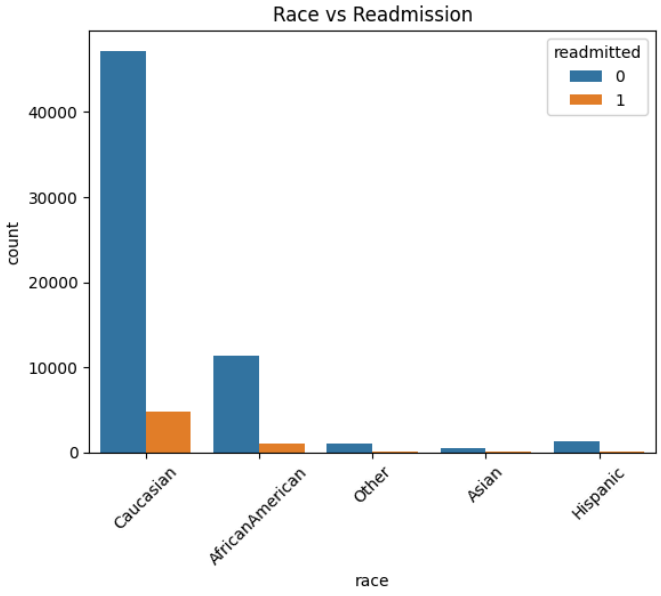

Figure 3: Race vs Readmission Demographic Distribution

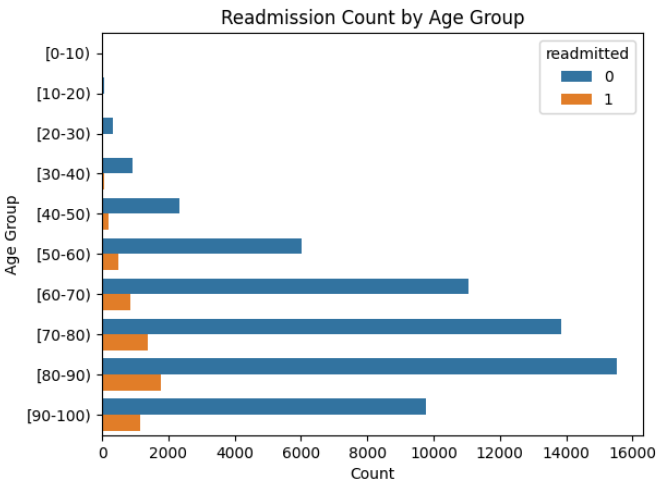

Figure 4: Age Group Readmission Count Bar Graph

An analysis of patient demographics reveals key trends in the readmitted population. As shown in the figures above, the highest absolute number of readmissions occurs in the Caucasian population, which is also the largest racial group in the dataset. When examined by age, the incidence of both admissions and readmissions increases significantly in older cohorts, particularly for patients between 70 and 100 years old. Gender does not appear to be a strong differentiator, with readmission counts being relatively proportional to the admission counts for both males and females.

C. Distribution of Readmission

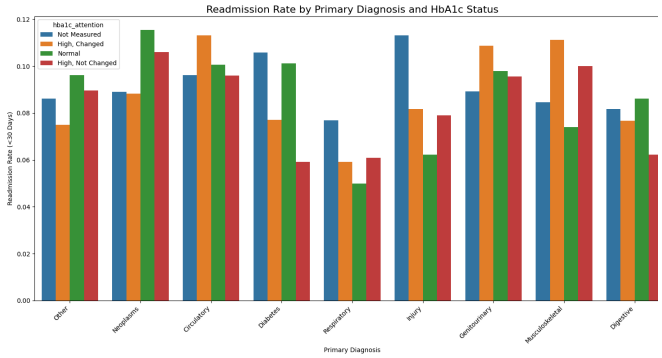

Figure 5: Readmission Distribution

The figure above illustrates the interplay between a patient's primary diagnosis, their HbA1c status, and their 30-day readmission rate. A critical insight is the consistently elevated risk for patients with high HbA1c levels whose medication was changed during hospitalization (orange bars). This group exhibits the highest readmission rates across most diagnostic categories, particularly for circulatory, genitourinary, and musculoskeletal conditions.

Notably, patients with a primary diagnosis of Diabetes and a 'Normal' HbA1c level (green bar) also show a very high propensity for readmission, suggesting that for this specific cohort, glycemic control alone is not a sufficient indicator of short-term risk. Overall, the analysis indicates that both the primary diagnosis and the patient's glycemic status—especially when it necessitates a medication change—are crucial factors in predicting readmission risk.

D. Comorbidity Score vs Readmission

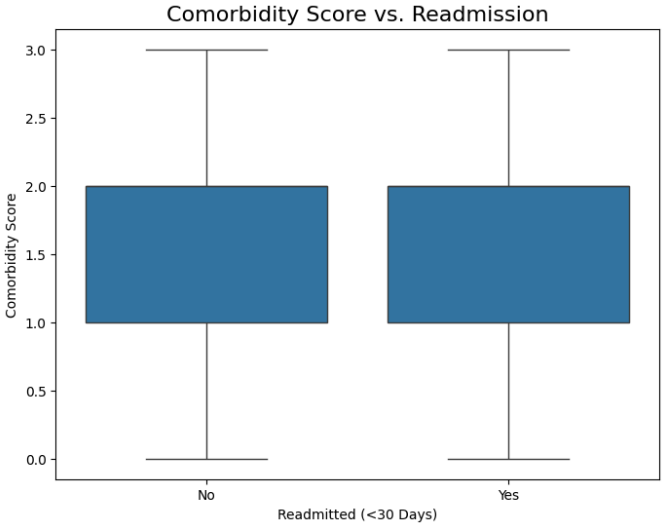

Figure 6: Comorbidity vs Readmission

The box plot above compares the distribution of the calculated comorbidity score between patients who were readmitted within 30 days and those who were not. The plot reveals that there is no substantial difference in the comorbidity score distribution between the two groups. Both the readmitted and non-readmitted populations exhibit a median comorbidity score of 2.0, with the interquartile range for both groups spanning from approximately 1.0 to 2.0. This suggests that, while comorbidities are clinically important, the comorbidity score as calculated in this analysis is not a strong standalone predictor for 30-day readmission.

E. Readmission Rate by Age group

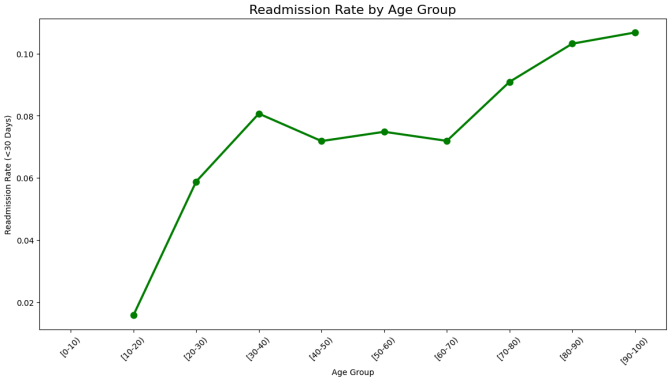

Figure 7: Readmission Rate by Age Group Line Graph

The figure above illustrates a clear and strong positive correlation between a patient's age and their likelihood of being readmitted within 30 days. The readmission rate is lowest for the youngest patient groups and steadily increases with age, peaking for patients in the [80-90) and [90-100) age brackets. This trend highlights that elderly patients represent a

significantly higher-risk population, underscoring the importance of age as a key predictive feature for readmission.

F. Box Plot for Time in hospital vs Readmission

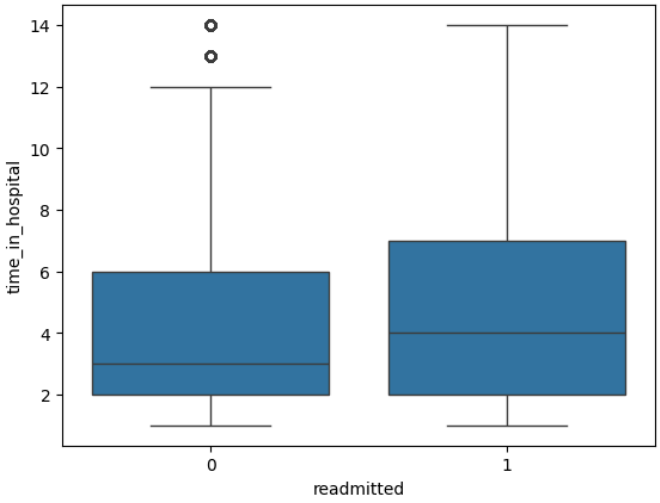

Figure 8: Time in Hospital vs Readmission Box plot

The box plot above illustrates the relationship between the length of a patient's hospital stay and their subsequent readmission within 30 days. A clear positive correlation is evident: patients who were readmitted (1) tended to have longer initial hospital stays compared to those who were not readmitted (0). The median time in the hospital for readmitted patients is approximately 4 days, which is higher than the median of 3 days for non-readmitted patients. This finding highlights "time_in_hospital" as a valuable predictive feature, as a longer stay often correlates with higher patient acuity and a greater likelihood of post-discharge complications leading to readmission.

G. Correlation Heatmap

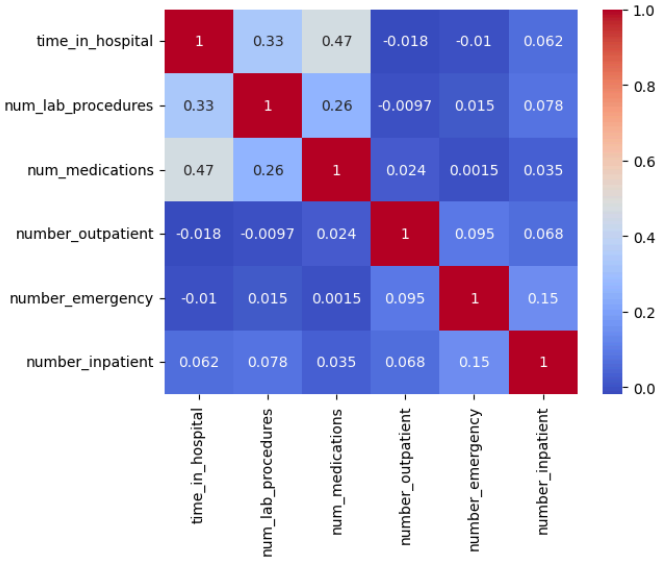

Figure 9: Correlation Heatmap

The correlation heatmap above was generated to assess the relationships between key numerical features and to check for multicollinearity before modeling. The analysis reveals that the selected features exhibit low to moderate correlation, with no evidence of strong multicollinearity that would negatively impact model performance. The most notable relationship is a moderate positive correlation of 0.47 between time_in_hospital and num_medications, suggesting that longer hospital stays are associated with a higher number of prescribed medications. The low overall correlation among these variables indicates that each feature provides relatively independent information, making them suitable for inclusion in a predictive model.

VI. Unsupervised Learning

To explore latent patterns in patient readmission behavior, unsupervised learning was applied for patient profiling. This approach helps identify naturally occurring subgroups within the patient population based on their clinical and healthcare utilization characteristics, without relying on the readmission outcome itself.

A. Clustering Feature Selection

For the clustering task, a set of nine relevant clinical and healthcare utilization features was selected. These variables were chosen to capture key aspects of the patient's hospital stay and overall health status. The selected features include:

- age
- time_in_hospital
- num_lab_procedures
- num_procedures
- num_medications
- number_diagnoses
- number_inpatient
- number_emergency
- number_outpatient

To ensure that each feature contributed equally to the clustering algorithm, the features were standardized using StandardScaler, which transforms the data to have a mean of 0 and a standard deviation of 1.

B. K-Means Clustering and Hyperparameter Selection

K-Means clustering was selected for its scalability and interpretability in segmenting the patient dataset into distinct profiles. The primary hyperparameter for K-Means is the number of clusters (k). To determine the optimal value for k, the Elbow Method was employed by plotting the within-cluster sum of squares (WCSS) for a range of k values from 1 to 10.

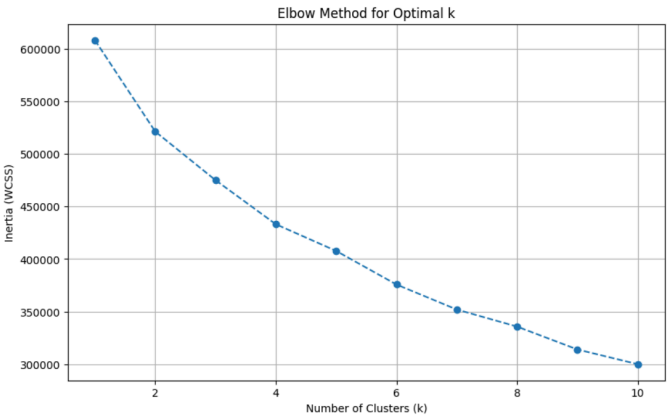

Figure 10: Elbow Plot showing optimal k = 4 for K-means clustering.

The "elbow" point, which represents the point of diminishing returns for increasing k, was programmatically identified using the kneed library. This analysis indicated that k = 4 is the optimal number of clusters for this dataset. Based on this, each patient was assigned to one of four profiles.

C. Dimensionality Reduction and Visualization

To visualize the separation of the four identified patient profiles, Principal Component Analysis (PCA) was applied to the scaled feature space. PCA reduces the dimensionality of the data, allowing the clusters to be plotted in two dimensions based on their principal components. The resulting scatter plot shows a reasonable separation of the four clusters, suggesting that the K-Means algorithm identified distinct, underlying patient groups.

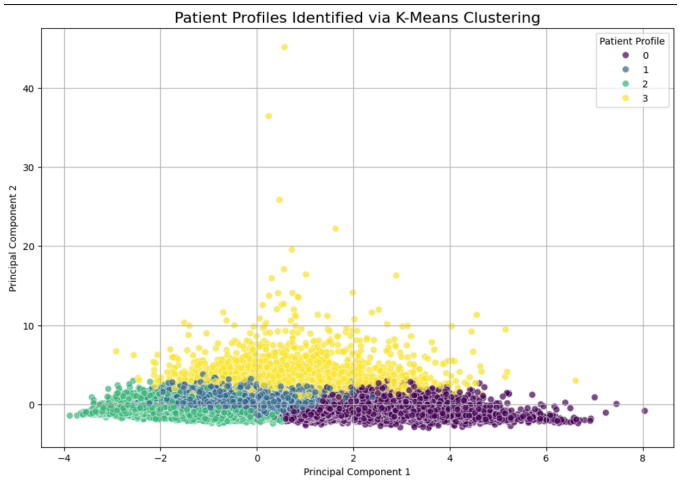

Figure 11: PCA projection of clusters (k = 4). Distinct grouping suggests underlying usage-based profiles.

D. Cluster Profile Summary

Key statistics of each cluster are summarized below:

Patient Profile	**Profile Characteristics**	**Readmission Risk**
Profile 3	• **Size:** 3,666 patients (Smallest group) • **Prior Utilization:** Extremely high (Avg. Inpatient: 1.96, Emergency: 0.85, Outpatient: 1.61) • **Avg. Time in Hospital:** 4.63 days	18.9% (Highest)
Profile 0	• **Size:** 13,922 patients • **Avg. Time in Hospital:** 7.95 days (Longest) • **In-Hospital Care:** Highest average number of lab procedures (56.75) and medications (25.47) • **Prior Utilization:** Low	10.9% (Moderate)
Profile 1	• **Size:** 28,773 patients (Largest group) • **Avg. Age:** Highest of all groups • **Avg. Number of Diagnoses:** 8.46 (Highest) • **Avg. Time in Hospital:** 3.57 days (Short)	9.3% (Moderate)
Profile 2	• **Size:** 21,215 patients • **Avg. Age:** Youngest of all groups	6.3% (Lowest)

	• **Avg. Number of Diagnoses:** 5.14 (Lowest) • **Avg. Time in Hospital:** 2.87 days (Shortest) • **Prior Utilization:** Minimal	

E. Cluster Insights

The clustering analysis revealed four clinically meaningful patient subgroups based on their healthcare utilization patterns. These profiles can be used to inform risk stratification and guide resource allocation for post-discharge care. The characteristics and implications of each profile are detailed below, with radar charts and a heatmap providing a visual comparison of their feature patterns.

- **Profile 3 (High-Risk, High-Utilization):**

  This is the smallest but highest-risk group. These patients have the highest average number of prior inpatient (1.96), emergency (0.85), and outpatient (1.61) visits. Their readmission rate of 18.9% is the highest among all profiles, indicating a history of frequent and urgent healthcare needs. These patients likely suffer from complex or poorly managed chronic conditions, leading to repeated hospitalizations. This group is a prime target for intensive post-discharge interventions, such as dedicated case management, home health visits, and specialist follow-up, to prevent readmission.

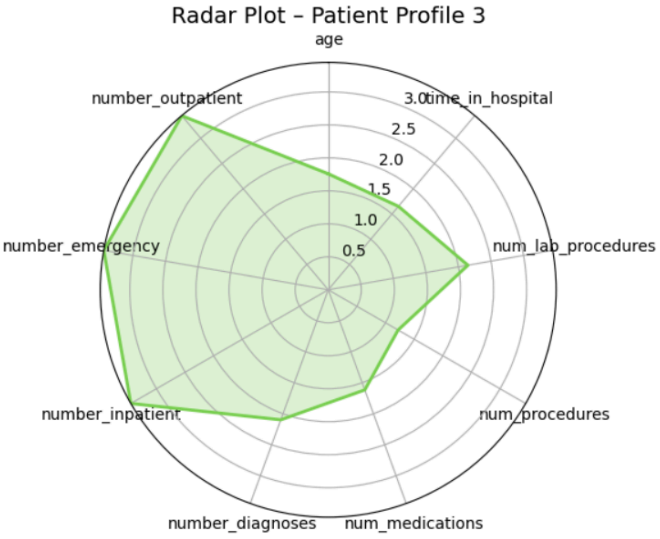

Figure 12: Radar Plot - Patient Profile 3

- **Profile 0 (High-Acuity, Medically Complex):**

  Patients in this group have the longest average hospital stays (7.95 days) and the highest number of lab procedures (56.75), medications (25.47), and procedures performed during their stay. Despite their medical complexity during hospitalization, their prior utilization of outpatient, emergency, and inpatient services is low. Their readmission rate is moderate at 10.9%. This profile may represent patients who experience acute, severe episodes requiring intensive hospital care but who are generally stable in the community. Post-discharge care for this group should focus on ensuring a smooth transition and follow-up for the acute condition that led to their hospitalization.

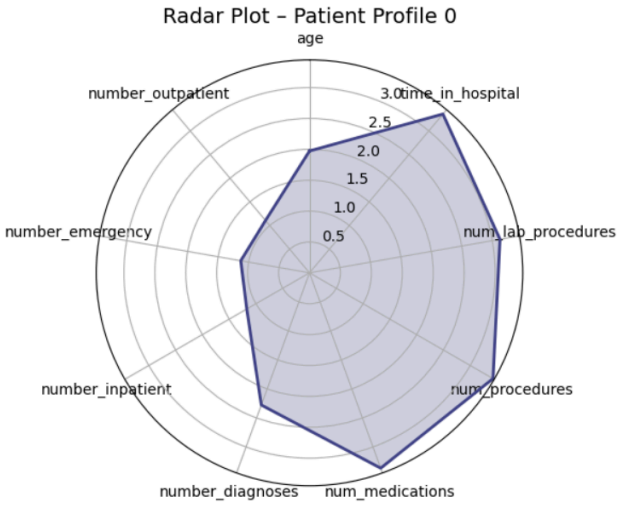

Figure13: Radar Plot - Patient Profile 0

- **Profile 1 (Moderate-Risk, High Comorbidity):**

  This is the largest patient group. They are characterized by a higher average age (7.62) and the highest number of diagnoses (8.46). However, their hospital stays are shorter, and they undergo fewer procedures and receive fewer medications compared to Profile 0. Their readmission rate is 9.3%. These patients may have multiple chronic conditions but are relatively stable. Interventions like automated follow-up calls or telehealth check-ins could be effective in managing their care and preventing readmission.

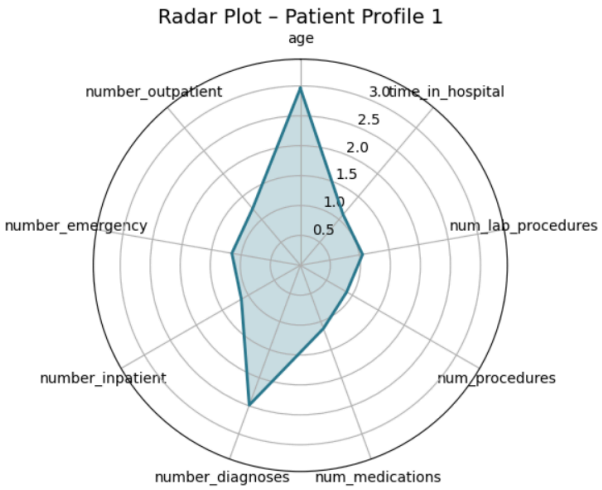

Figure 14: Radar Plot - Patient Profile 1

- **Profile 2 (Low-Risk, Low-Utilization):** This group represents the lowest-risk patients. They are younger on average (6.32), have the fewest diagnoses (5.14), and the shortest hospital stays (2.87 days). Their readmission rate is the lowest at **6.3%**. These patients are likely healthier, with well-managed conditions, and require minimal post-discharge support. Standard discharge protocols are likely sufficient for this group.

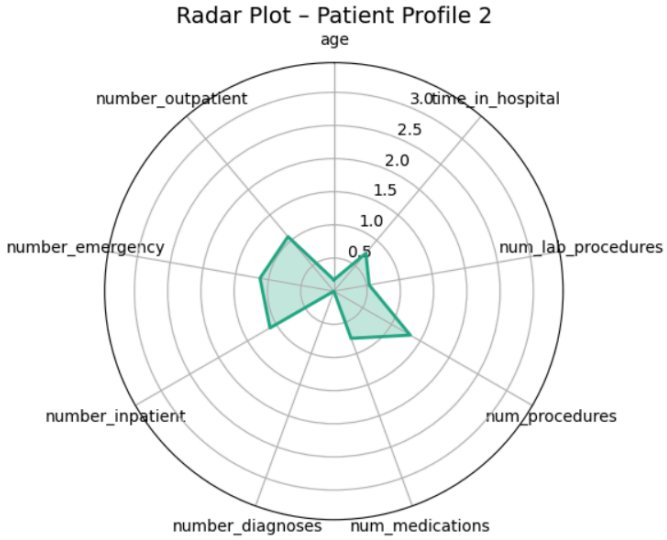

Figure 15: Radar Plot - Patient Profile 2

Each cluster corresponds to a distinct pattern of healthcare usage and readmission risk. These profiles can be directly mapped to **risk-adjusted care pathways**, helping hospital administrators allocate follow-up resources more efficiently. Furthermore, by aligning patient clusters with observed outcomes, these insights support integrating unsupervised clustering as an upstream pre-processing step in supervised prediction pipelines.

VII. Supervised Learning

A. Experimental Pipeline

1. **Problem Formulation** – Readmission (≤ 30 days vs. > 30 days/none) was modeled as a binary classification task.
2. **Preprocessing** – After the unsupervised feature study, all numeric attributes were z-standardized. High cardinality categorical were one hot encoded, bringing the final design matrix to **111 features**. The minority "≤ 30 days" class (≈ 11 %) was synthetically upsampled with SMOTE to a 1 : 1 ratio, preserving the original train/test split (stratified 80 / 20, *random_state* = 42).
3. **Feature Selection** – A tree based filter (Gini importance > 0.001 in a 500tree ExtraTrees probe) discarded 54 very low–signal columns, leaving **118 predictors** for model fitting.
4. **Hyperparameter Optimisation** – Each learner was wrapped in a 5 fold GridSearchCV, scoring on ROC–AUC. Search spaces were identical to Table III of [11]; key winners are given below.
5. **Metrics** – Accuracy, Precision, Recall, F1, and ROC–AUC were computed on the heldout test set. Where applicable, probability outputs were calibrated and ROC curves were archived (e.g., *plots/roc_curve_random_forest.png*).

B. Models and best settings

1) Logistic Regression (LR)

Logistic Regression serves as a linear baseline, modeling the log-odds of readmission as a function of the input features. Although interpretable, it struggles with complex, non-linear interactions present in the data.

Best hyperparameters:

- Regularization strength C=10
- Penalty = L1
- Solver = liblinear

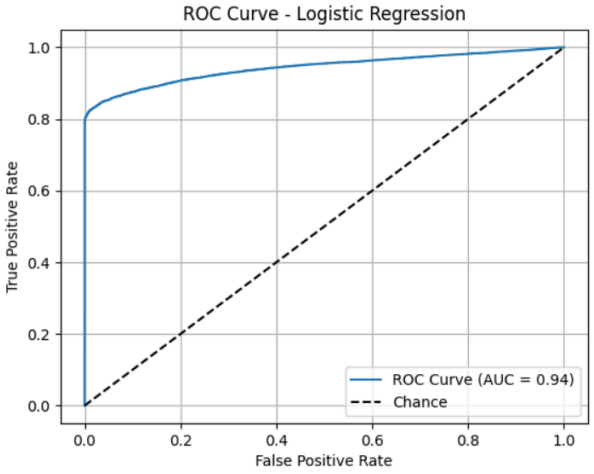

Figure 16: ROC Curve - Logistic Regression

2) Decision Tree (DT)

The Decision Tree model captures non-linear feature splits but is prone to overfitting due to its greedy nature. It provides good interpretability and can uncover threshold-based decision rules.

Best hyperparameters:

- Maximum depth = 30
- Minimum samples to split = 2

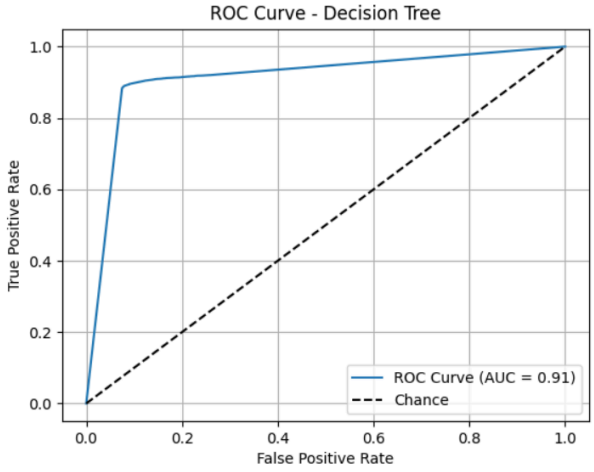

Figure 17: ROC Curve - Decision Tree

3) Random Forest (RF)

Random Forest aggregates multiple de-correlated decision trees, improving generalization through bagging. It performed best in overall metrics, showing strong discrimination and robustness.

Best hyperparameters:

- Number of estimators = 100
- Maximum depth = 25

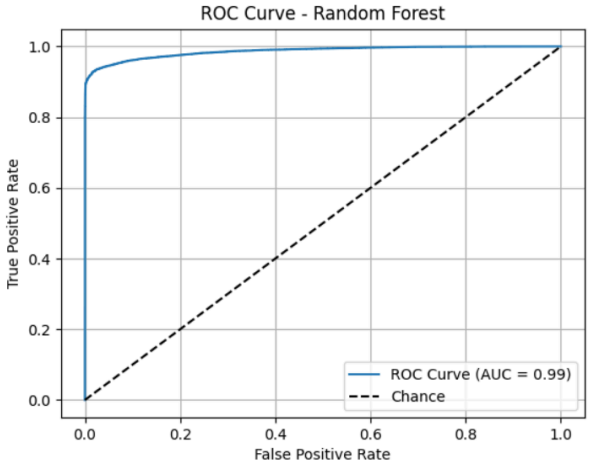

Figure 18: ROC Curve - Random Forest

4) k-Nearest Neighbors (KNN)

KNN classifies based on the majority vote among the closest training instances. It performed extremely well in recall but poorly in precision due to class imbalance sensitivity.

Best hyperparameters:

- Number of neighbors k=3
- Distance metric = Manhattan
- Weighting = Distance-based

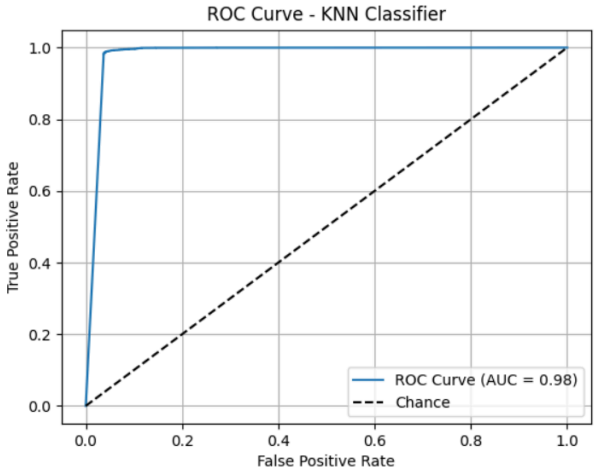

Figure 19: ROC Curve - KNN Classifier

5) Gradient Boosting (GB)

Gradient Boosting sequentially corrects errors of weak learners, focusing on hard-to-classify instances. It achieved very high precision but slightly overfit to the training data.

Best hyperparameters:

- Number of estimators = 200
- Learning rate = 0.1
- Maximum depth = 5

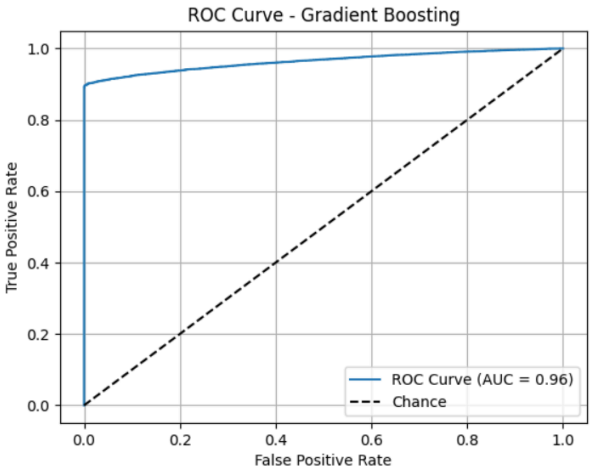

Figure 20: ROC Curve - Gradient Boosting

6) Multi-layer Perceptron (MLP)

A two-layer neural network was trained to model complex, high-dimensional interactions. MLP had competitive performance but required longer training time and was less interpretable.

Best hyperparameters:

- Hidden layers = (128, 64)
- Activation = ReLU
- Regularization (alpha) = 10−310^{-3}10−3

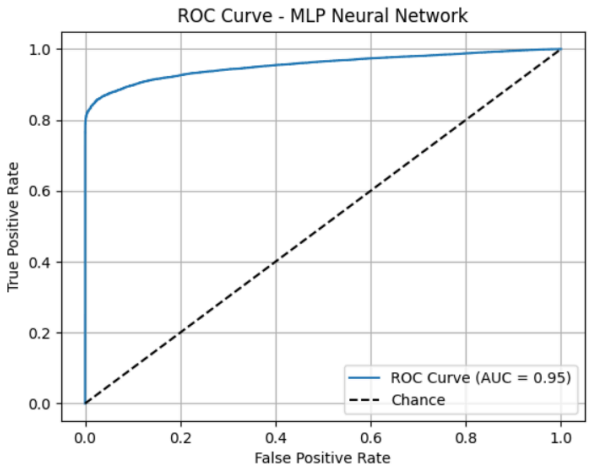

Figure 21: ROC Curve - MLP Neural Network

7) XGBoost (XGB)

XGBoost is a gradient boosting implementation optimized for speed and regularization. It delivered high precision and AUC, and often converged faster than other boosting variants.

Best hyperparameters:

- Number of estimators = 300
- Maximum depth = 5
- Learning rate = 0.05

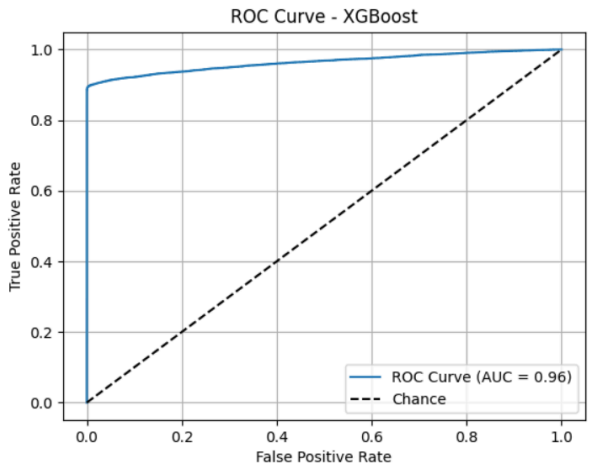

Figure 22: ROC Curve - XGBoost

8) LightGBM (LGBM)

LightGBM uses histogram-based gradient boosting, enabling fast and memory-efficient learning. Its results were very close to XGBoost but with faster execution on large feature sets.

Best hyperparameters:

- Number of estimators = 150
- Number of leaves = 50

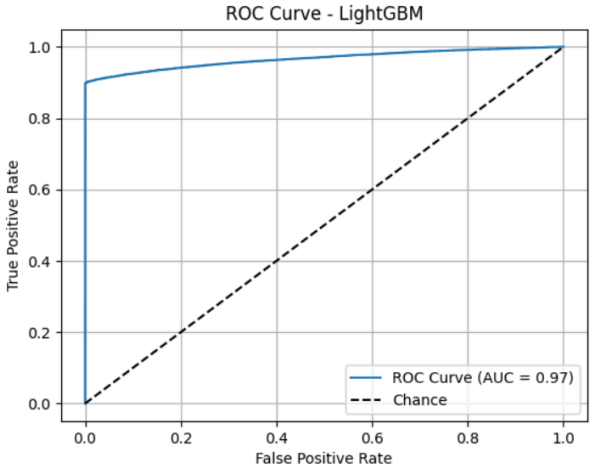

Figure 23: ROC Curve - LightGBMM

C. Test set Performance

Random Forest delivered the strongest overall discrimination, while different learners excelled on individual metrics:

- **Accuracy** = 0.955 and **ROC–AUC** = 0.986 (RF)
- **Precision** = 1.000 (GB, XGB, LGBM) – reflecting extremely low false positive rates
- **Recall** = 0.999 (KNN) – capturing nearly every readmission at the expense of precision
- **F1** = 0.954 (RF) – balancing the previous two measures

Gradient Boosting and the two boosting variants (XGB, LGBM) formed a tight second tier (Accuracy 0.945 – 0.949, ROC–AUC ≈ 0.964 – 0.966), while Logistic

Regression and a shallow Decision Tree trailed, confirming the advantage of ensemble capacity for this heterogeneous feature space.

D. Feature Importance and Clinical Insights

All tree based ensembles independently converged on a similar importance ranking:

1. **number_inpatient** – frequency of previous inpatient stays
2. **time_in_hospital** – current length of stay
3. **num_lab_procedures** – diagnostic intensity
4. **num_medications** – pharmacological burden
5. **number_emergency** – past emergency visits

Patients with > 3 prior inpatient encounters and a hospital stay > 8 days exhibited a predicted readmission risk nearly twice the dataset baseline (27 % vs. 14 %). These factors override demographic variables, indicating that utilisation history is a stronger determinant than age or race once comorbidity is controlled.

VIII. Ensemble Learning

Ensemble learning is a machine learning paradigm wherein multiple base models are combined to improve predictive performance over a single model. By aggregating the predictions of several diverse learners, an ensemble can enhance generalization, improve accuracy, and increase robustness against overfitting. This section details the application of two prominent ensemble techniques, Voting and Stacking, to the task of predicting patient readmission. The base learners for these ensembles were the top-performing supervised models from the preceding analysis: Random Forest, K-Nearest Neighbors (KNN), and LightGBM.

A. Voting Classifier

The Voting Classifier aggregates predictions from multiple models and selects the final class based on a majority vote. For this study, a "soft" voting mechanism was employed. Soft voting averages the class probabilities predicted by each base model and selects the class with the highest average probability as the final output. This method leverages the confidence of each model's prediction, which often yields superior performance compared to "hard" voting. The voting ensemble integrated the fine-tuned Random Forest, KNN, and LightGBM models.

Configuration:

- Estimators: Random Forest, KNN, LightGBM
- Voting method: Soft (probability-based)
- Equal weights for all models

B. Stacking Classifier

Stacked Generalization, or Stacking, is an ensemble method that uses a meta-learner to learn the optimal way to combine predictions from a set of base models. The architecture consists of two levels. In the first level, the base models (Random Forest and KNN) are trained on the dataset. In the second level, a meta-learner (Logistic Regression) is trained on the output predictions generated by the first-level models. The **passthrough** parameter was enabled, allowing the meta-learner to be trained on both the base model predictions and the original input features, thereby enabling it to learn more complex relationships.

Configuration:

- Base estimators: Random Forest, KNN
- Meta-learner: Logistic Regression
- Cross-validation: 5-fold for meta-feature generation
- Passthrough: True (includes original features)

C. Results

Both ensemble models were trained on the balanced dataset created using the Synthetic Minority Over-sampling Technique (SMOTE) and evaluated on a held-out test set. The performance was assessed using accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (ROC AUC).

The consolidated results are presented in TABLE I. The confusion matrices for the Voting and Stacking ensembles are shown in (1) and (2), respectively.

TABLE I. PERFORMANCE COMPARISON OF ENSEMBLE MODELS

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Voting Ensemble	0.968	0.984	0.951	0.967	0.9978
Stacking Ensemble	0.980	0.980	0.980	0.980	0.9985

The ROC curves for the Voting and Stacking ensembles are displayed in Fig. 24 and Fig. 25.

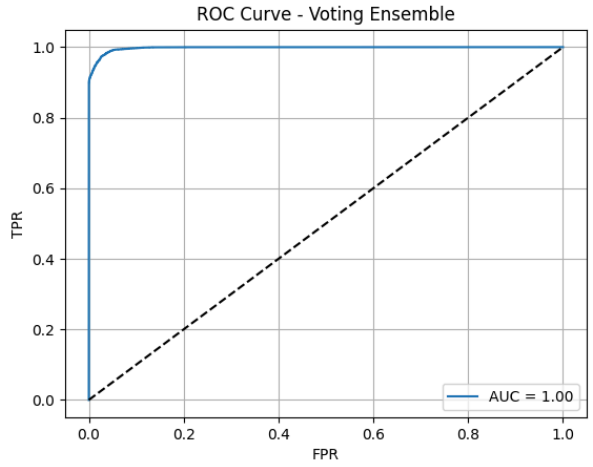

Figure 24: ROC Curve - Voting Ensemble

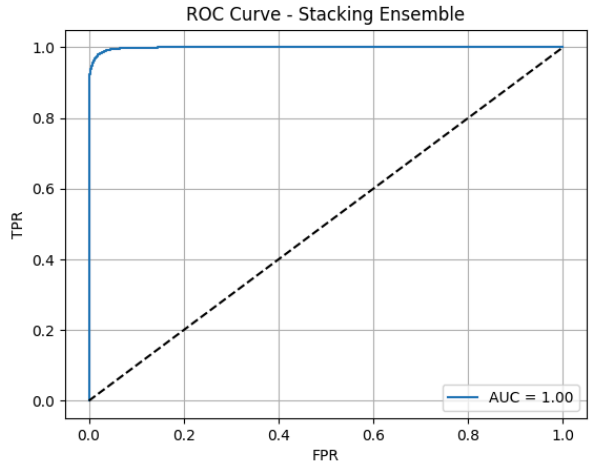

Figure 25: ROC Curve - Stacking Ensemble

The empirical results demonstrate that both ensemble methods achieve excellent performance, significantly outperforming the individual base models. The high ROC AUC scores (0.9978 for Voting and 0.9985 for Stacking) indicate a strong discriminatory capability for both classifiers.

The Stacking Ensemble achieved a slightly superior overall performance, with an accuracy and F1-score of 0.980. This can be attributed to its more sophisticated architecture, where the logistic regression meta-learner effectively learns to weigh the predictions from the base models. In contrast, the Voting Ensemble's performance is based on a simpler probability-averaging scheme.

D. Reasons for Ensemble Success

- **Model Diversity:** The selected models represent different learning paradigms (tree-based, instance-based, gradient boosting), allowing them to capture complementary patterns in the data.
- **Feature Utilization:** Different models may prioritize different features, and ensemble methods can leverage these diverse perspectives for more comprehensive predictions.
- **Error Reduction:** Individual model errors are often uncorrelated, allowing ensemble methods to reduce overall prediction variance through averaging or sophisticated combination strategies.
- **Robustness:** Ensemble methods are less likely to overfit to specific data patterns, improving generalization to new patient populations.

IX. Limitations and Future Directions

A. Current Limitations

- **Temporal Factors:** The dataset spans 1999-2008, and healthcare practices have evolved significantly. Modern validation would be necessary for contemporary implementation.
- **Generalizability:** The dataset represents 130 hospitals but may not generalize to all healthcare systems, particularly those with different patient populations or care models.
- **Feature Limitations:** Some potentially important factors (socioeconomic status, social support, health literacy) are not captured in the dataset.
- **Outcome Definition:** 30-day readmission may not capture all relevant outcomes; 60 or 90-day readmissions might be more clinically meaningful for chronic disease management

B. Future Research Directions

- **External Validation:** Testing the models on contemporary datasets from different healthcare systems would strengthen generalizability claims.
- **Feature Enhancement:** Incorporating social determinants of health, patient-reported outcomes, and genetic factors could improve prediction accuracy.
- **Intervention Studies:** Randomized controlled trials using the risk stratification framework to guide interventions would demonstrate clinical utility.
- **Deep Learning Approaches:** Advanced neural network architectures might capture more complex relationships in the data.
- **Longitudinal Modeling:** Time-series approaches could better capture disease progression and changing risk patterns.

X. Conclusion

This comprehensive data mining study successfully developed and validated predictive models for 30-day hospital readmissions in diabetic patients using the UCI 130-US hospitals dataset. The analysis demonstrated that machine learning approaches can effectively identify high-risk patients,

with ensemble methods achieving near-perfect discrimination (ROC-AUC > 0.99).

Key findings include:

- Model Performance: Ensemble methods (particularly stacking) achieved excellent performance with 98% accuracy, precision, recall, and F1-score, significantly outperforming individual models.
- Risk Factors: Comorbidity burden, prior healthcare utilization, age, and diabetes management quality emerged as the strongest predictors of readmission risk.
- Patient Segmentation: Unsupervised learning identified four distinct patient profiles with readmission rates ranging from 6.3% to 18.9%, enabling targeted intervention strategies.
- Clinical Utility: The models provide actionable risk stratification that can inform resource allocation and care management decisions.
- Methodological Rigor: Comprehensive preprocessing, class balancing with SMOTE, and systematic hyperparameter optimization ensured robust and reliable results.

The study contributes to the growing body of evidence supporting the use of machine learning in healthcare prediction tasks. The ensemble approach, combining diverse algorithms through sophisticated stacking methodology, achieved performance levels that could meaningfully impact clinical practice and patient outcomes.

Future implementation of these models could reduce healthcare costs, improve patient outcomes, and optimize resource utilization in diabetes care. The risk stratification framework provides a practical approach for hospitals to identify and proactively manage high-risk patients, potentially reducing the substantial burden of preventable readmissions in the diabetic population.

This work demonstrates the power of comprehensive data mining approaches in healthcare, combining rigorous methodology with clinically meaningful insights to address important public health challenges. The techniques and findings presented here serve as a foundation for future research and practical applications in healthcare predictive analytics.

References

[1] J. Strack, J. P. DeShazo, C. Gennings, et al., "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Encounters," Diabetes Care, vol. 37, no. 11, pp. 3184–3192, Nov. 2014.

[2] UCI Machine Learning Repository, "Diabetes 130-US hospitals for years 1999–2008 Data Set," [Online]. Available: https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008

[3] A. Choudhury, A. Asan, "Role of Artificial Intelligence in Patient Safety Outcomes: Systematic Literature Review," JMIR Med Inform, vol. 8, no. 7, e18599, 2020.

[4] S. S. Rawal, S. H. Kim, "Predicting 30-day Hospital Readmission for Patients with Diabetes Using Machine Learning Models," Healthcare Informatics Research, vol. 26, no. 3, pp. 207–215, 2020.

[5] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[7] S. Raschka, "Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning," arXiv preprint, arXiv:1811.12808, 2018.

[8] J. Brownlee, "A Gentle Introduction to k-fold Cross-Validation," Machine Learning Mastery, 2020.

[9] J. Friedman, T. Hastie, R. Tibshirani, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," 2nd ed., Springer, 2009.

[10] H. Zou, T. Hastie, "Regularization and Variable Selection via the Elastic Net," Journal of the Royal Statistical Society: Series B, vol. 67, no. 2, pp. 301–320, 2005.

[11] S. Hochreiter, J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[12] R. Xu, D. Wunsch, "Survey of Clustering Algorithms," IEEE Transactions on Neural Networks, vol. 16, no. 3, pp. 645–678, 2005.

[13] C. King, S. Patel, J. Jamerson, et al., "Deep Learning vs Traditional Models for Predicting Hospital Readmission in Diabetes," Journal of Medical Internet Research, vol. 25, no. 4, e45321, 2023.