A close up of a logo

AI-generated content may be incorrect.A blue and white logo

AI-generated content may be incorrect.

**CBIO313: Data Mining and Machine Learning**

**Hospital Readmission in Diabetic Patients Using Machine Learning: A Comparative Model Analysis and Deployment Approach**

**Farid Maged Ghattas**

**221000545**

**Under the Supervision of:**

**Dr. Mohamed Elsayeh**

**Abstract**

Hospital readmissions among diabetic patients pose significant challenges in terms of cost, patient care, and resource allocation. This project applies machine learning techniques to predict 30-day hospital readmission using a structured dataset of over 100,000 diabetic patient encounters. The goal was to develop and evaluate classification models capable of identifying high-risk patients based on clinical, demographic, and treatment-related features.

After thorough data preprocessing, including handling missing values, one-hot encoding, and feature selection informed by exploratory data analysis, multiple models were trained and compared, namely Logistic Regression, Random Forest, and XGBoost. Key visualizations such as histograms and correlation heatmaps were used to highlight data distribution and feature relationships, guiding both model selection and feature selection. Due to severe class imbalance, Synthetic Minority Oversampling Technique (SMOTE) was used to enhance minority class representation.

Evaluation metrics included accuracy, precision, recall, and F1 score, with special emphasis on the latter due to the critical impact of false negatives in healthcare settings. XGBoost achieved the highest F1 score and was selected as the final model. It was then deployed using a Flask-based API (app.py) to provide real time predictions in a scalable web application format.

While the model achieved promising results, limitations such as class imbalance and missing clinical context remain. This study still demonstrates the potential of machine learning to support proactive healthcare decision-making and reduce preventable hospital readmissions through data-driven decisions.

1. **Introduction**

The exponential growth of healthcare data presents a unique opportunity for using machine learning (ML) techniques to derive insights and support data-driven decision-making in a clinical setting. One especialy critical area is the prediction of patient readmission, which has substantial implications for hospital resource management, cost control, and patient outcomes. Accurately identifying patients at high risk of readmission enables timely interventions, potentially reducing avoidable hospitalizations and improving the quality of care.

This project explores the use of supervised machine learning models to predict hospital readmission among diabetic patients based on various clinical, demographic, and treatment-related features. Diabetes, being a chronic condition with high readmission rates, serves as an ideal case study for predictive modeling in healthcare. The dataset used in this project comprises thousands of patient encounters, including variables such as age, gender, admission type, number of procedures, medication changes, and discharge disposition.

The primary objective of this study is to build and evaluate machine learning models that can predict whether a patient will be readmitted to the hospital. This is formulated as a binary classification task, where the target variable indicates whether a patient was readmitted within a defined timeframe. To accomplish this, the project involves several phases: data preprocessing, exploratory data analysis (EDA), feature selection, model training and comparison, and deployment of the final model in a real-world application setting.

A wide range of classification algorithms were explored, including logistic regression, decision trees, random forests, and XGBoost. Hyperparameter tuning was performed to optimize performance, and models were assessed using accuracy, F1 score, precision, and recall metrics. Among the evaluated models, the highest performance and was selected as the final model for deployment. The trained model was then integrated into a web application using a lightweight app.py Flask script to facilitate practical usage.

In addition to model development, exploratory data analysis was conducted to understand the relationships among variables and identify patterns relevant to readmission risk. Visualizations provided key insights into variable distributions and interdependencies, informing both preprocessing strategies and feature engineering decisions.

This report presents a comprehensive account of the project, including the methodology, results, and implications of the findings. The goal is to demonstrate how machine learning, when applied thoughtfully and rigorously, can contribute to solving complex challenges in healthcare.

1. **Methodology**

The methodology for this project follows a structured pipeline encompassing data acquisition, preprocessing, exploratory data analysis (EDA), feature selection, model development, evaluation, and deployment. Each stage was designed to ensure the resulting machine learning models are both accurate and generalizable.

* 1. **Data Source and Structure**

The dataset used in this project originates from a healthcare repository containing electronic health record data for over 100,000 diabetic patient encounters. Each record includes variables such as demographic data (age, gender), hospitalization metrics (length of stay, admission source, discharge disposition), clinical indicators (number of medications, procedures, diagnosis codes), and administrative flags (readmission status, medication changes).

The target variable was binary: whether a patient was readmitted or not. This required some preprocessing to convert categorical “<30”, “>30”, and “NO” labels into a binary classification problem (readmitted = 1 for "<30", 0 otherwise).

* 1. **Data Cleaning and Preprocessing**

The preprocessing phase included the following critical steps:

* **Missing Values**: Variables with excessive missingness such as weight and payer\_code were dropped. For other columns, missing values were imputed using mode or a special category ("Unknown") for categorical data.
* **Irrelevant Features**: Non-informative columns such as encounter\_id were excluded to avoid data leakage or model bias.
* **Categorical Encoding**: Nominal categorical variables were transformed using one-hot encoding. High-cardinality features (e.g., medications, diagnoses) were grouped or truncated to reduce dimensionality.
* **Feature Selection**: Based on correlation analysis and domain knowledge, features with low variance or poor correlation with the target were removed. Strongly correlated variables such as num\_lab\_procedures, num\_medications, and number\_inpatient were retained.
  1. **Exploratory Data Analysis (EDA)**

To guide preprocessing and model selection, extensive EDA was performed. Key visualizations include:

* **Histogram of Readmission Rates**
* **Correlation Heatmap**
* **Class Imbalance Observation**
  1. **Model Development**

A selection of supervised classification models was trained and evaluated using scikit-learn. The algorithms tested include:

* **Logistic Regression**
* **Random Forest Classifier**
* **XGBoost Classifier**

Each model was wrapped in a pipeline that included data scaling (for algorithms sensitive to feature magnitude), encoding, and cross-validation.

Hyperparameters tuning was accomplished through a for loop that would reiterate over the selected parameters and extract the ones with the highest F1 score (This was selected since the presence of a False-Negative would be extremely dangerous on a medical classification model).

* 1. **Evaluation Metrics**

To comprehensively assess model performance, the following metrics were calculated:

* **Accuracy** – Overall classification correctness.
* **Precision** – Ability to correctly identify readmissions.
* **Recall** – Ability to capture all actual readmissions.
* **F1 Score** – Harmonic mean of precision and recall, suitable for imbalanced data.

Confusion matrices were also plotted for the top models.

* 1. **Model Selection and Deployment**

XGBoost emerged as the best-performing model across multiple metrics, demonstrating superior F1 score and balanced precision-recall performance on the test set.

The final XGBoost model was serialized and integrated into a lightweight web application using **Flask**. The deployment script, app.py, includes endpoints for model prediction based on user input features, making the model accessible in real-time.

1. **Results**

This section presents the findings from training and evaluating the machine learning models described in the methodology. Performance comparisons were made based on test set metrics and visual diagnostics to identify the most effective model for predicting hospital readmission.

* 1. **Exploratory Data Analysis Outcomes**

The exploratory data analysis (EDA) offered key insights that guided subsequent modeling decisions:

A graph of a number of blue rectangular bars

AI-generated content may be incorrect.

* **Histogram of Readmission Rates (Fig. 1)**: This visualization highlighted the fact that the <30 group was severely underrepresented, SMOTE (Synthetic minority oversampling technique) was used to produce synthetic samples for the <30 group, this was crucial in preventing overfitting.

A screenshot of a graph

AI-generated content may be incorrect.

* **Correlation Heatmap (Fig. 2)**: The heatmap revealed several relationships between variables. Notably, num\_medications showed a moderate positive correlation with time\_in\_hospital, which could result in multicollinearity, however the threshold for that is 0.8, which our results were not close to yet.
  1. **Model Training Performance**

Three primary models were trained and evaluated: Logistic Regression, Random Forest, and XGBoost. A consistent data pipeline was used for all models, incorporating feature scaling, encoding, and cross-validation.

1. **Discussion**

The outcomes of this project demonstrate the potential of machine learning models to effectively predict hospital readmissions among diabetic patients. Among the evaluated classifiers, XGBoost slightly outperformed both logistic regression and random forest models, which is critical when dealing with imbalanced datasets and high-stakes decisions, such as those in healthcare.

* 1. **Model Interpretability and Real-World Impact**

XGBoost’s performance can be attributed to its ability to model non-linear relationships and its resilience to noise and missing data. Importantly, its relative interpretability compared to deep learning models makes it suitable for clinical decision support systems, where explainability is essential for gaining trust from healthcare professionals.

In practical terms, a deployed XGBoost model that identifies patients at high risk of readmission could be integrated into hospital information systems. This would enable clinicians to flag vulnerable patients before discharge, recommend additional monitoring, or provide targeted interventions. Such strategies could reduce avoidable readmissions, enhance patient outcomes, and reduce operational costs.

* 1. **Insights from EDA**

Exploratory data analysis provided several non-trivial insights that shaped the modeling process:

* The **age group of 70–80 years** and **emergency admissions** were associated with higher readmission rates (*Fig. 1*). This underscores the importance of age-specific care plans and pre-discharge assessments for acute care patients.
* The **correlation heatmap (Fig. 2)** revealed modest but meaningful relationships between features like num\_medications, time\_in\_hospital, and number\_inpatient, suggesting that patients requiring more intensive treatment or hospitalization history were more likely to be readmitted.

The **imbalance in the dataset**, with only ~11% readmitted cases, made accuracy a misleading performance metric. This justified the focus on recall and F1 score, which are more informative in the presence of minority class underrepresentation.

* 1. **Limitations**

While the model achieved promising results, several limitations should be acknowledged:

* **Class imbalance** remained a challenge. Although stratified validation and F1 scoring helped, some cases were still misclassified. More sophisticated resampling methods or cost-sensitive learning could improve sensitivity in future work.
* **Feature availability** was constrained by the dataset’s structure. Variables such as social determinants of health, post-discharge support, or comorbidity indices were either absent or not even detailed. Including such data could strengthen predictive power.

**5.4 Model Deployment Considerations**

The deployment of the final XGBoost model via a Flask app.py script represents a proof-of-concept for real-world integration. It would allow for scalable use in clinical dashboards, though further work would be needed to ensure data security, and continuous monitoring.

1. **Conclusion**

This project sets out to develop a robust machine learning framework capable of predicting hospital readmission among diabetic patients using clinical and administrative data. Through careful preprocessing, exploratory analysis, and systematic evaluation of multiple classification models, we identified XGBoost as the most effective algorithm for this task. Its performance across key metrics, such as the F1 score highlighted its suitability for imbalanced medical classification problems, where missing a positive case (false negative) could have serious consequences.

The results indicate that certain patient characteristics such as older age, type of admission, and number of prior hospitalizations are predictive of early readmission. These insights were uncovered during EDA and confirmed by model performance, reinforcing the value of combining data visualization with statistical modeling.

Despite its strengths, further optimization is necessary before clinical deployment. Future work could incorporate richer feature sets, apply advanced resampling techniques, or explore ensemble strategies to boost sensitivity without sacrificing precision.

The project concluded in the deployment of the trained XGBoost model using a Flask-based API (app.py), enabling real-time predictions in an accessible and scalable format. This deployment marks an important step toward practical integration of predictive analytics into clinical workflows.

In summary, this study demonstrates how machine learning, when applied, can offer valuable tools for healthcare decision-making. With continued improvement, models like the one developed here have the potential to support proactive decision making, reduce hospital readmissions, and improve the patient outcomes.