Predicting Hospital Readmissions for Diabetic Patients

A Machine Learning Approach Using the Diabetes 130-US Hospitals Dataset

Ehab Henein

Overview

Diabetes is a chronic condition that disrupts how our bodies convert food into energy , leading to severe complications such as heart disease, kidney failure, and stroke , making it a significant public health issue . One of the most challenging aspect s of managing diabetes is the frequent hospital admissions due to related complications.

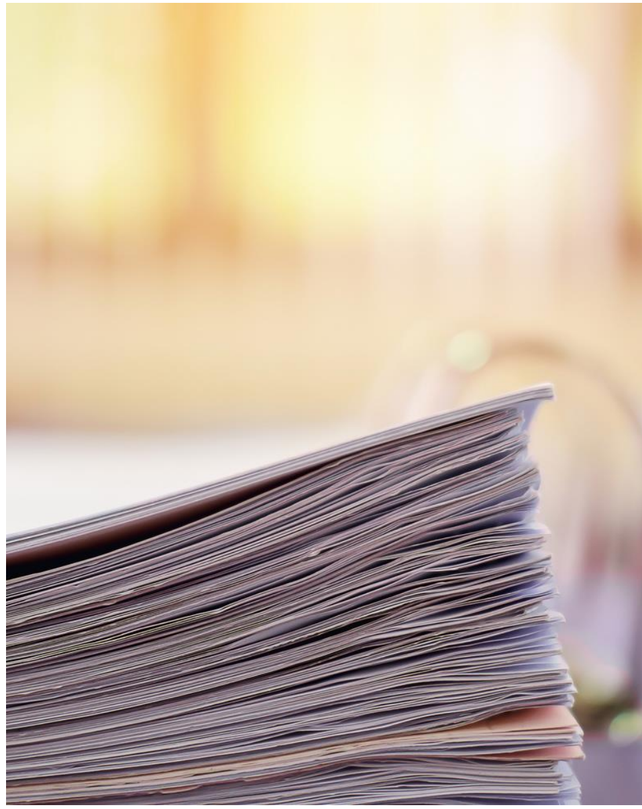
Predicting hospital readmissions is critical as it allows healthcare providers to improve patient outcomes through timely interventions, reduce costs, and manage hospital resources more effectively.

Significance of Hospital Readmissions

- Costly and often preventable readmissions
- Prediction allows for proactive care and better resource management
- High-risk identification targets interventions effectively

Data Selection

- **Source: UCI Machine Learning Repository**
- **Dataset: 130 US hospitals, 100,000+ observations**
- **Objective: Identify risk factors for diabetes readmissions**
- **Target variable: 'Readmitted'**



Data Processing

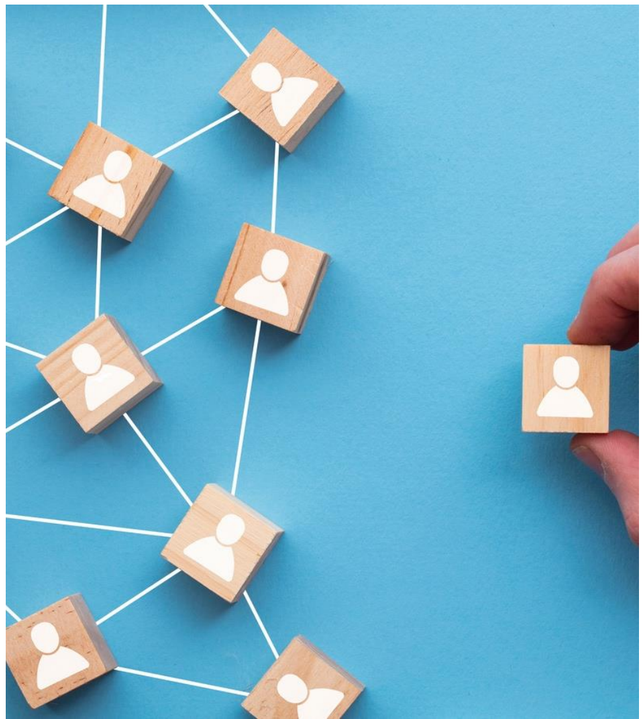
The data for this project was sourced from the UCI Machine Learning Repository and contains over 100,000 observations from 130 hospitals in the United States between 1999 and 2008. This dataset includes various features, such as patient demographics, diagnoses, treatments, and hospital readmission status.

1. Transforming the target variable and the diabetes medication variable into binary to make it suitable for binary classification.

2. Encoding diagnosis codes, age, and admission type.

3. Continuous features are standardized using a scaler, which normalizes the values around a mean of zero and a standard deviation of one.

4. Binning and encoding numeric count features.



Feature Selection

- **Key variables: Patient demographics, hospital metrics, health conditions**
- **Additional indicators: Treatment and admission type**
- **Focus on factors with likely predictive power for readmissions**

Modeling Approach

- **Models used: Logistic Regression and Random Forest Classification**
- **Train-test split: 80/20 ratio for accuracy assessment**
- **Comparison of model performance**

Model
Logistic Regression

- **Probability-based classification for binary outcomes**
- **Encoding ensures the model handles categorical variables**
- **Suitable for predicting readmission status**

Logistic Regression uses multiple iterations to calculate the probability of an independent variable correlating with the dependent variable. After running our Logistic Regression Model, the output of our model gave us an Accuracy of 55% . This indicates that 5 5 % of the predictions were correct across all classes. While the model performs slightly better than random guessing, there is room for improvement. To improve, the model may need further tuning, feature engineering, or adjustments to focus on enhancing recall for the readmitted class.

- **Ensemble model with decision trees for accuracy**
- **Random state set to 42 for consistent results**
- **Evaluates feature importance in predicting readmission**

Model
Random Forest Classification

In a random forest, decision trees are built using a small portion of the training set, and the features of the information within the subset are evaluated. These features are then used for classification to determine how each independent variable affects the dependent variable. Much like our Logistic Regression model, shows moderate performance, with better results for non-readmitted cases (class 0) than for readmitted cases (class 1). Given the low recall for class 1, the model may benefit from strategies like rebalancing the classes, adjusting model parameters, or adding relevant features to improve its sensitivity to readmitted cases.

Logistic Regression Results

- **Overall accuracy of the model: 57%**
- **Predominantly performs better with non-readmitted cases (Class 0)**
- **Class 0 (non-readmitted):**
  - Precision: 58% (proportion of true non-readmissions among predicted non-readmissions)
  - Recall: 69% (proportion of actual non-readmissions correctly identified)
  - F1-Score: 0.63 (balance between precision and recall for non-readmitted cases)

- **Class 1 (readmitted):**
  - Precision: 54% (proportion of true readmissions among predicted readmissions)
  - Recall: 43% (proportion of actual readmissions correctly identified)
  - F1-Score: 0.48 (weaker performance in identifying readmitted cases)

- **Observations**:
  - The model does better in predicting non-readmissions, as shown by the higher precision and recall for Class 0.
  - Lower recall for readmitted cases (Class 1) indicates challenges in identifying true readmissions.
  - Overall, the F1-score suggests improvement, especially in detecting readmissions.

Random Forest Results

- **Overall accuracy of the model: 55%**
- **Predominantly performs better with non-readmitted cases (Class 0)**
- **Class 0 (non-readmitted):**
  - Precision: 57% (proportion of true non-readmissions among predicted non-readmissions)
  - Recall: 60% (proportion of actual non-readmissions correctly identified)
  - F1-Score: 59 (balance between precision and recall for non-readmitted cases)

- **Class 1 (readmitted):**
  - Precision: 51% (proportion of true readmissions among predicted readmissions)
  - Recall: 47% (proportion of actual readmissions correctly identified)
  - F1-Score: 0.49 (reflecting challenges in accurately predicting readmitted cases)

- **Observations:**
  - Random Forest has slightly better precision and recall for non-readmissions (Class 0) than readmissions.
  - Recall for readmitted cases (Class 1) is relatively low, indicating that the model struggles with identifying true readmissions.
  - The overall F1-score for Class 1 suggests limited effectiveness in predicting readmitted cases accurately.
  - Consistent with Logistic Regression, the model is biased toward non-readmitted cases.

Interpretation of Results

- **Overall Performance:**

  Both models (Logistic Regression and Random Forest) show moderate accuracy, around 55-57%.
- **Class 0 (non-readmitted cases):**

  - Both models perform better at predicting non-readmitted cases.
  - Higher precision, recall, and F1-scores than Class 1, suggesting model bias towards non-readmissions.

- **Class 1 (readmitted cases):**

  - Lower precision and recall for readmitted cases in both models, indicating challenges in accurately identifying readmissions.
  - Logistic Regression has a slightly higher F1-score for Class 1 than Random Forest, but both models struggle with low recall.

- **Average Metrics:**

  - Macro and weighted averages align closely with overall accuracy, reflecting moderate model performance.
  - Indicates the models may benefit from addressing class imbalance.

- **Improvement suggestions:**
  - Rebalancing the dataset to mitigate class imbalance.
  - Feature engineering to capture more relevant predictors for readmissions.
  - Hyperparameter tuning to enhance recall for readmitted cases.

Ethical Implications

Ensure Fairness and Privacy

- Data anonymization toprotect privacy as per HIPAA Law.
- Informed consent from patients on their data usage.
- Diverse demographic representation
- Majority of data from Caucasians

Ensure Better Patient Care

- Reduce hospital readmissions without compromising patient care.
- Requires regular evaluation of the model's performance for consistency or the need for adjustments.

Conclusion

- Predicting readmissions is essential for diabetic patient care and cost reduction
- High-risk factors identified: diagnosis count, medication count, and diabetes medication usage
- Lesser impact from variables like age and lab procedures
- Implications for proactive, targeted care for high-risk patients
- Potential for reducing healthcare costs and improving patient outcomes

Next Steps & Recommendations

- **Potential improvements to model accuracy:**
  - Rebalancing classes to address model bias
  - Additional feature engineering to capture complex interactions
  - Hyperparameter tuning for Random Forest

- **Broader implications:**
  - Insights for healthcare policy targeting readmission reduction
  - Implementing proactive strategies for high-risk patients

Thank You!