**Decision Tree Analysis of Hospital Readmission**

Data source: <https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008>

Reads: current folder + Relevant article

The objective of this project was to analyze hospital readmission data using a decision tree model to identify the key factors influencing whether patients with diabetes are readmitted within a given timeframe. Hospital readmissions present significant financial and operational challenges for healthcare systems, and understanding the variables that contribute to readmission risk is crucial for improving patient care and reducing costs. A decision tree model was chosen due to its interpretability and its ability to handle both numerical and categorical data, making it particularly suited for this type of analysis.

|  |  |
| --- | --- |
| Variable | Description |
| encounter\_id | Unique identifier for each hospital encounter (visit). |
| patient\_nbr | Unique identifier for each patient, which may have multiple encounters. |
| race | Race of the patient (e.g., Caucasian, African American, etc.). |
| gender | Gender of the patient (e.g., Male, Female). |
| age | Age range of the patient (e.g., [0-10), [10-20), etc.). |
| weight | Weight of the patient (often missing in datasets). |
| admission\_type\_id | Type of admission (e.g., emergency, urgent, elective). |
| discharge\_disposition\_id | Disposition after discharge (e.g., home, rehab, expired). |
| admission\_source\_id | Source of admission (e.g., physician referral, emergency room). |
| time\_in\_hospital | Length of the hospital stay in days. |
| citoglipton | Indicator of whether Citoglipton medication was used (Yes/No). |
| insulin | Indicates insulin usage, showing dose changes (e.g., No, Up, Down). |
| glyburide-metformin | Indicates if glyburide-metformin combination was prescribed (Yes/No). |
| glipizide-metformin | Indicates if glipizide-metformin combination was prescribed (Yes/No). |
| glimepiride-pioglitazone | Indicates if glimepiride-pioglitazone combination was prescribed (Yes/No). |
| metformin-rosiglitazone | Indicates if metformin-rosiglitazone combination was prescribed (Yes/No). |
| metformin-pioglitazone | Indicates if metformin-pioglitazone combination was prescribed (Yes/No). |
| change | Indicates if there was a change in the diabetes medication (Ch for change, No for no change). |
| diabetesMed | Indicates whether the patient was on diabetes medication (Yes/No). |
| readmitted | Indicates whether the patient was readmitted after discharge (NO, >30, <30). |

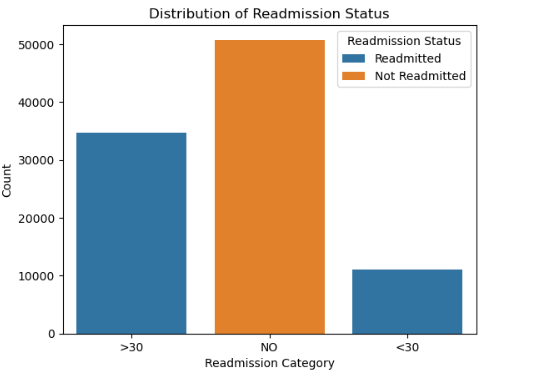
The dataset used in this study contains detailed information on hospital admissions, patient demographics, and clinical attributes. Key variables include patient age, time in the hospital, and medication usage, as well as the target variable, *readmitted*, which indicates whether a patient was readmitted within 30 days, after 30 days, or not readmitted at all. This dataset provides a comprehensive view of potential predictors for hospital readmissions, offering valuable insights into the patterns and factors that contribute to readmission rates.

**Data Preprocessing**

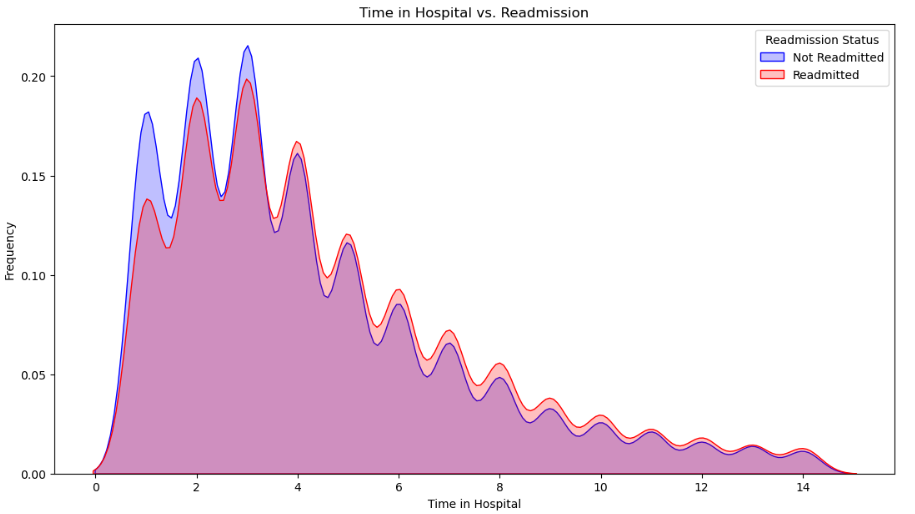
Before training the model, several preprocessing steps were implemented. Missing values were handled by removing records with substantial missing information, and imputation methods were applied where feasible to fill in minor gaps. Additionally, categorical variables, such as gender and age, were converted into numerical formats, as decision trees require numerical input. The data was then split into training and testing sets, with 80% allocated for training and 20% for testing. This split ensured sufficient data for training the model while maintaining a robust test set for model evaluation.

**Exploratory Data Analysis**

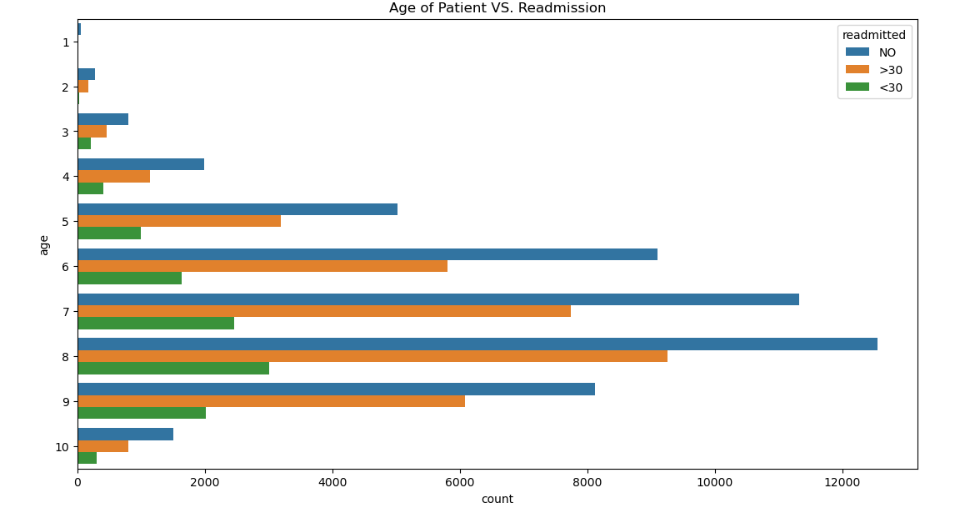
Exploratory data analysis (EDA) provided the following insights into of key variables.



This bar chart shows that the majority of patients fall into the NO category, indicating they were not readmitted after their initial discharge. A smaller, yet significant, portion of patients fall into the >30 category, meaning they were readmitted more than 30 days after discharge. The smallest group is <30, representing patients readmitted within 30 days, which is often a critical measure in healthcare quality. This distribution highlights that while most patients do not return to the hospital, there is still a considerable number of readmissions, particularly after the 30-day mark, which may warrant further investigation.



This chart reveals that shorter hospital stays are more common among both readmitted and non-readmitted patients, with the frequency declining as the length of stay increases. Although both groups follow a similar trend, the peaks in the distribution indicate slightly higher densities of shorter stays among non-readmitted patients. This pattern suggests that while short hospital stays are typical for all patients, there may be subtle differences that could correlate with a higher likelihood of readmission. The similar density patterns between groups, however, indicate that hospital stay duration alone may not be a primary predictor of readmission risk



This chart examines the distribution of patient ages by readmission category using a horizontal bar chart. The age groups are numerically coded, and in nearly every age group, the majority of patients are in the NO readmission category. However, there is a noticeable trend in older age groups (e.g., age categories 7, 8, and 9), which show higher counts of readmission, especially in the >30 category. This pattern suggests that age may be a contributing factor to readmission risk, with older patients more likely to be readmitted, possibly due to chronic health conditions or age-related vulnerabilities. This insight implies a potential need for targeted interventions for older patients to reduce their likelihood of readmission.

**Model Implementation**

The decision tree model was implemented using Python’s scikit-learn library, which provided a straightforward method for training and evaluating decision trees. Hyperparameter tuning was conducted to optimize model performance, with a focus on limiting the maximum depth of the tree to prevent overfitting, thereby ensuring the model remained generalizable to new data. Cross-validation was used during parameter tuning to assess model stability and accuracy across different data splits.

**Model Performance**

The decision tree model was evaluated on a dataset of 19,290 patient records, with the goal of predicting hospital readmissions. The model achieved an overall accuracy of 55%, meaning it correctly classified the readmission status for approximately half of the cases. Precision for the "no readmission" class was 0.56, and for the "readmission" class, it was 0.53, indicating the model was correct around 53-56% of the time for both classes. The recall was 0.57 for non-readmission and 0.52 for readmission, capturing 57% of true non-readmission cases and 52% of true readmission cases. The F1-scores, balancing precision and recall, were 0.57 for non-readmission and 0.53 for readmission.

The confusion matrix showed 5,697 true negatives, 4,852 true positives, 4,324 false positives, and 4,417 false negatives. Although the model demonstrated some predictive ability, the relatively low accuracy and recall suggest room for improvement, particularly in reducing false negatives (missed readmissions) and false positives (incorrect readmissions). Further optimization, such as hyperparameter tuning, feature engineering, or using more advanced algorithms, is expected to enhance the model's performance.

**Strengths and Limitations**

One of the key strengths of the decision tree model is its interpretability. By examining feature importance scores, we were able to identify the variables that the model considered most influential in its predictions. This insight can be valuable for healthcare providers, as it allows them to focus on the factors that most significantly contribute to readmission risk and prioritize interventions for high-risk patients.

However, the model also has several limitations. The dataset exhibited class imbalance, with a larger proportion of patients not readmitted compared to those who were readmitted, which likely impacted the model’s ability to predict short-term readmissions effectively. Additionally, while decision trees are interpretable, their performance could potentially be improved by exploring ensemble methods, such as random forests, which combine multiple trees for greater accuracy.

**Conclusion**

In conclusion, this project demonstrated that decision tree models can effectively analyze hospital readmission data, providing insights into the key factors contributing to readmission risk. Variables such as age, time in the hospital, and medication usage were identified as important predictors of readmission. While the model’s performance was moderate, further improvements can be made by addressing class imbalance, optimizing hyperparameters, and exploring more advanced models. This work highlights the potential of decision trees for healthcare data analysis and the importance of predictive modeling in improving patient care and operational efficiency.