Smart Stays: Predicting Hospital Stay Duration in Newfoundland and Labrador

Participant Name: Maria Hennebury

Date: July 22, 2024

Abstract

Newfoundland and Labrador's healthcare system faces significant challenges, including overcrowded emergency departments, a severe shortage of doctors, and increasing burnout among medical professionals. This project leverages machine learning techniques to predict hospital stay durations for patients in the region. By analyzing key factors such as patient demographics, morbidity trends, and hospital resources, the study aims to enhance healthcare delivery efficiency and improve patient outcomes. Utilizing models like the Random Forest Regressor, the project demonstrated substantial improvements in prediction accuracy, with the Final Refined Model achieving a Mean Squared Error (MSE) of 0.088 and an R-squared of 0.964. This approach provides actionable insights for optimizing resource allocation and management within the healthcare system, supporting better decision-making for healthcare providers in addressing the region's pressing issues.

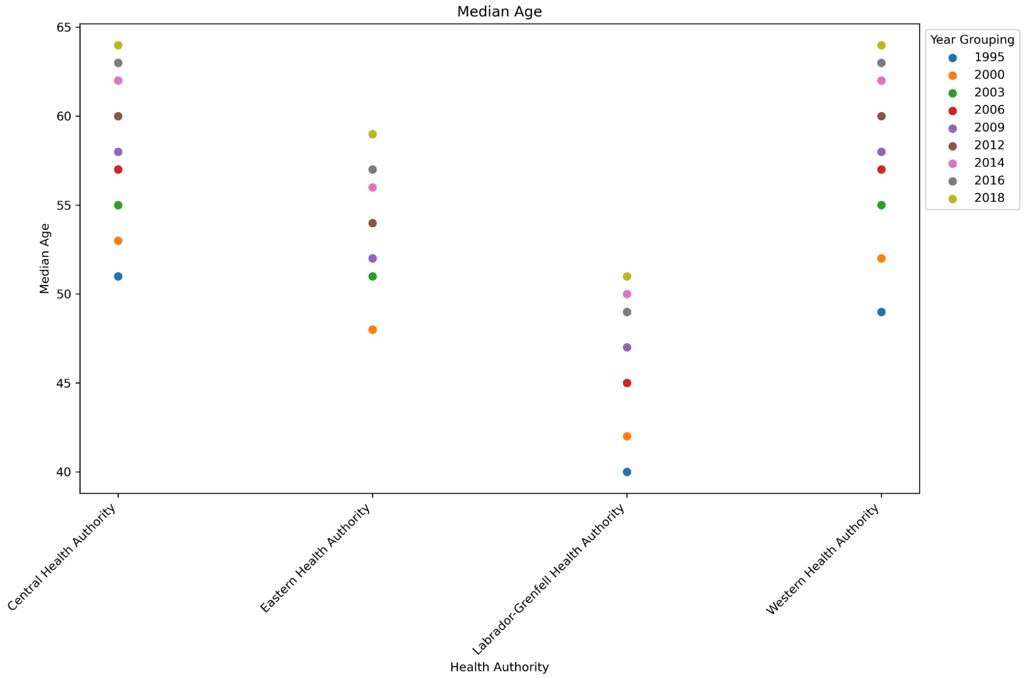
Introduction

This study tackles the critical challenge of predicting hospital stay durations in Newfoundland and Labrador, where the healthcare system is grappling with issues such as overcrowded emergency departments, a shortage of medical professionals, and increasing burnout among healthcare providers. The primary objective of this research is to apply machine learning models, specifically the Random Forest Regressor, to identify and predict factors influencing hospital stay durations. By doing so, the study aims to optimize resource allocation, improve patient care, and enhance the overall efficiency of healthcare delivery. Accurate predictions of hospital stays are crucial for effective management of hospital resources, enabling better planning and potentially alleviating some of the systemic pressures faced by the healthcare system in the region.

Background

The healthcare system in Newfoundland and Labrador is confronted with rising morbidity rates and variable hospital stay durations, which pose significant challenges for effective resource management and quality patient care. Accurate predictions of hospital stay lengths are essential for optimizing resource allocation and improving patient outcomes. The "Health Accord NL" report, released in February 2022, outlines a comprehensive framework for transforming healthcare in the province. The report emphasizes the need for enhancements in primary care access, mental health services, and the addressing of social determinants of health. It advocates for service integration, innovative care models, and the strategic use of technology to drive improvements.

Notably, the report also highlights that life expectancy in Newfoundland and Labrador lags behind the Canadian average, further underscoring the urgency for effective healthcare strategies. By aligning with the report's recommendations, this study aims to contribute to the ongoing efforts to improve healthcare delivery through machine learning. The project seeks to address some of the systemic issues identified, such as resource allocation and patient care management, by providing actionable insights into hospital stay durations. Compared to other studies, this research offers a region-specific approach, adding unique value to the understanding of healthcare delivery in Newfoundland and Labrador.

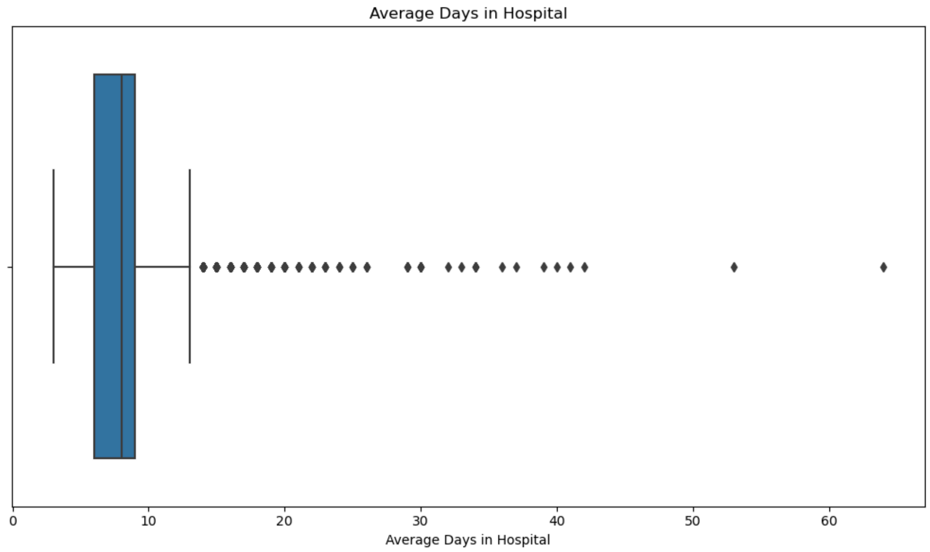


Data Collection and Preparation

This project utilized the Hospital Morbidity dataset for Newfoundland and Labrador, which includes various variables such as patient demographics, diseases treated, and temporal factors (available at OpenData.gov.nl.ca). The dataset comprises 13,979 entries and 57 columns, with key statistics indicating a mean patient age of approximately 54 years and an average hospital stay duration of about 8 days. Critical categorical variables influencing hospital stay predictions include geography, gender, and specific diseases.

To ensure the integrity of the analysis, a copy of the original DataFrame was maintained. Non-numeric values in the target column were converted to NaN, and rows with NaN values were removed to ensure completeness and reliability of the data. Outliers were detected and filtered using the Interquartile Range (IQR) method to improve model accuracy and prevent skewed results. Challenges encountered during data preparation included handling missing values and outliers, which were addressed through imputation and filtering techniques.

The feature matrix (X) and target vector (y) were prepared by converting non-numeric feature columns to numeric formats and removing rows with missing values. Selected features for the model included patient demographics, morbidity trends, and hospital resources, chosen for their relevance to predicting hospital stay durations. Numerical and categorical features were processed through distinct pipelines: numerical features were imputed and scaled to ensure consistency and comparability, while categorical features were imputed and one-hot encoded to facilitate model training. A ColumnTransformer was employed to integrate these preprocessing steps, and a model pipeline was constructed to fit a RandomForestRegressor. The model's performance was evaluated using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), with cross-validation applied to assess the model's robustness and generalizability.

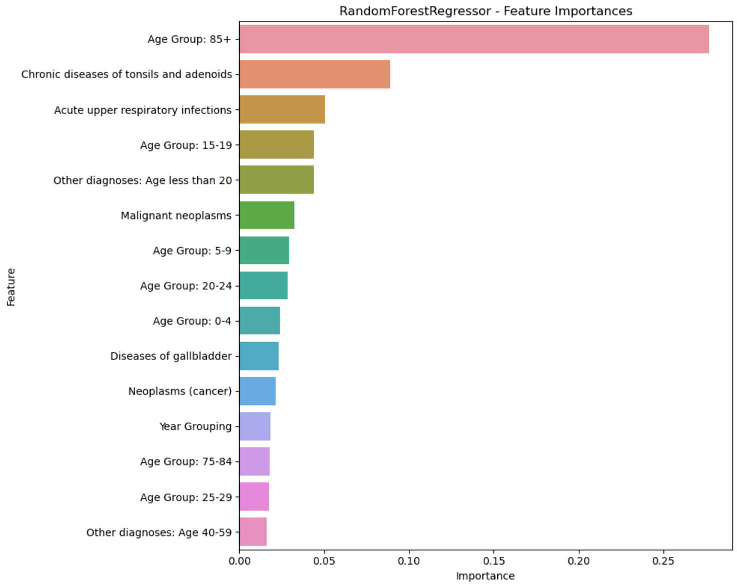


Results

The Initial Model achieved a Mean Squared Error (MSE) of 0.637, an R-squared value of 0.736, and a Mean Absolute Error (MAE) of 0.576. These metrics provided a baseline for evaluating the model's predictive performance. The Refined Model exhibited a slight increase in MSE to 0.699 and a decrease in R-squared to 0.710, with an MAE of 0.609. These changes suggest that while the refinement process led to some adjustments in performance, it did not significantly enhance the model's predictive capability compared to the Initial Model.

The Final Refined Model demonstrated substantial improvement, achieving an MSE of 0.088, an R-squared of 0.964, and an MAE of 0.210. This indicates a much better fit to the data and more accurate predictions of hospital stay durations. The improved metrics highlight the effectiveness of the final model in capturing the underlying patterns in the data.

Cross-validation results for the Final Refined Model revealed an MSE of 1.140 with a standard deviation of 0.296. While these results show the model's robustness and generalizability, the variability in performance suggests that there are some inconsistencies in predictions across different subsets of the data. Despite this variability, the model overall demonstrated strong performance and reliability.



Discussion

I had anticipated that fractures and asthma would be prominent in terms of importance, but it appears that the 85+ age group was notably significant. It would be interesting to explore what percentage of these cases are related to end-of-life care. Additionally, tonsil and adenoid issues, along with upper respiratory conditions, were high on the importance list, which aligns with their frequent need for medical intervention. Cancer also ranked high, reflecting its significant impact on hospital stays due to its complex treatment requirements.

This study underscores the critical role of age groups and specific diseases in predicting hospital stay durations. The model offers valuable insights for optimizing resource allocation and enhancing patient care. By accurately predicting lengths of stay, hospitals can better manage staff and resources, potentially reducing costs and improving operational efficiency. The model also provides guidance for policy decisions related to workforce dynamics and healthcare trends.

Geographic specifics could have provided additional valuable insights into regional variations in hospital stay durations, potentially revealing more nuanced patterns. However, due to limitations in my coding skills and time constraints, I was unable to fully incorporate geographic data into the model. Future research should aim to include geographic factors to explore regional differences in hospital stays and enhance the overall analysis.

Conclusion

In conclusion, I am incredibly satisfied with the outcome of this project. Starting with just a concept and minimal coding experience, to ultimately producing executable code, has been a deeply rewarding journey. Despite facing challenges, particularly with high-performance computing (HPC) due to time constraints, I successfully ran the project on SIKU. However, there is potential for scaling up HPC use in future research to handle larger datasets and more complex models. The project effectively showcases the potential of machine learning in predicting hospital stay durations, providing valuable insights for healthcare providers in Newfoundland and Labrador. Looking ahead, further research could focus on enhancing the model and integrating a broader range of data, including larger-scale HPC implementations, to advance healthcare analytics even further and address the broader challenges faced by the healthcare system.

References

- GitHub Repository: https://github.com/MariaHennebury/ISP.git
- Dataset: https://opendata.gov.nl.ca/public/opendata/page/?page-id=datasetdetails&id=69
- Health Accord NL Report: https://www.healthaccordnl.ca/wp-content/uploads/2022/02/HANL_Report_Document_Web_modFeb28-2022