Final Project Report

Predicting 30-Day Readmission Risk Among Patients With Diabetes Using Multivariate Predictors

HI2251

Soli Evans & Abhimaan Mayadam

1. Introduction

Hospital readmissions continue to strain the U.S. healthcare system, particularly among individuals living with diabetes. These unplanned readmissions disrupt continuity of care, increase healthcare costs, and often signal underlying challenges in disease management. Because diabetes is a complex chronic condition requiring careful monitoring, medication adherence, and coordinated follow-up, many patients remain vulnerable during the transition from inpatient care back into the community. Health systems are under increasing pressure to reduce avoidable readmissions, especially under initiatives such as the Hospital Readmissions Reduction Program (HRRP). The HRRP penalizes excessive readmission rates for people who have acute myocardial infarctions, congestive heart failure, pneumonia, COPD, and hip and knee replacements when under Medicare. (Fingar and Washington 2015) Readmission rates are also a good statistic of health care quality and cost efficiency, especially taking into consideration Medicare beneficiaries. (Talavera 2017)

In this project, we sought to examine the predictors of 30-day readmission among diabetic patients and evaluate the extent to which a predictive model could identify individuals at elevated risk. Our goal was to use a large, real-world clinical dataset to understand how demographic, clinical, and utilization factors contribute to readmission risk and to assess how predictive modeling might support proactive intervention strategies. By combining exploratory data analysis, logistic regression, and machine-learning evaluation, we aimed to produce insights that could ultimately enhance discharge planning and care coordination.

2. Description of the Problem and Dataset

To conduct this analysis, we used the UCI Diabetes 130-US Hospitals Dataset, a well-known multicenter dataset containing more than 100,000 inpatient encounters for patients with diabetes across 130 hospitals. After removing rows with missing or invalid values, the cleaned dataset included 101,766 encounters and 51 variables describing demographic characteristics, clinical indicators, healthcare utilization history, treatment patterns, and readmission outcomes.

The outcome variable was recoded into a binary indicator representing whether the patient was readmitted within 30 days. This allowed us to focus specifically on short-term, unplanned readmissions, which are of particular interest to health systems attempting to reduce preventable rehospitalizations.

To contextualize the population, we examined several descriptive visualizations. For example, the distribution of patient ages showed that the majority of encounters involved individuals aged 60-80, which aligns with national diabetes hospitalization trends according to the CDC (2024). An estimated 13.8 million of people above 65 years of age have diagnosed diabetes, out of the estimated 29.4 million adults who have diagnosed diabetes.

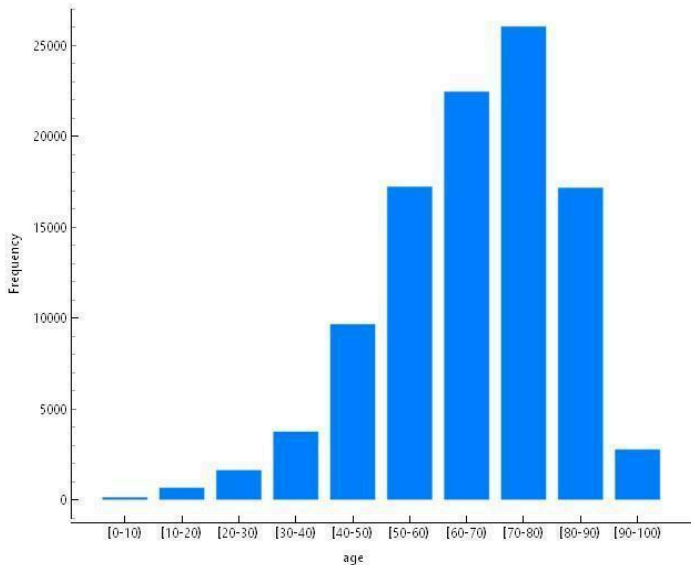


Figure 1 Age

The gender distribution showed a slightly higher number of female encounters than male, and the racial distribution reflected a predominantly Caucasian population with a substantial proportion of African American patients.

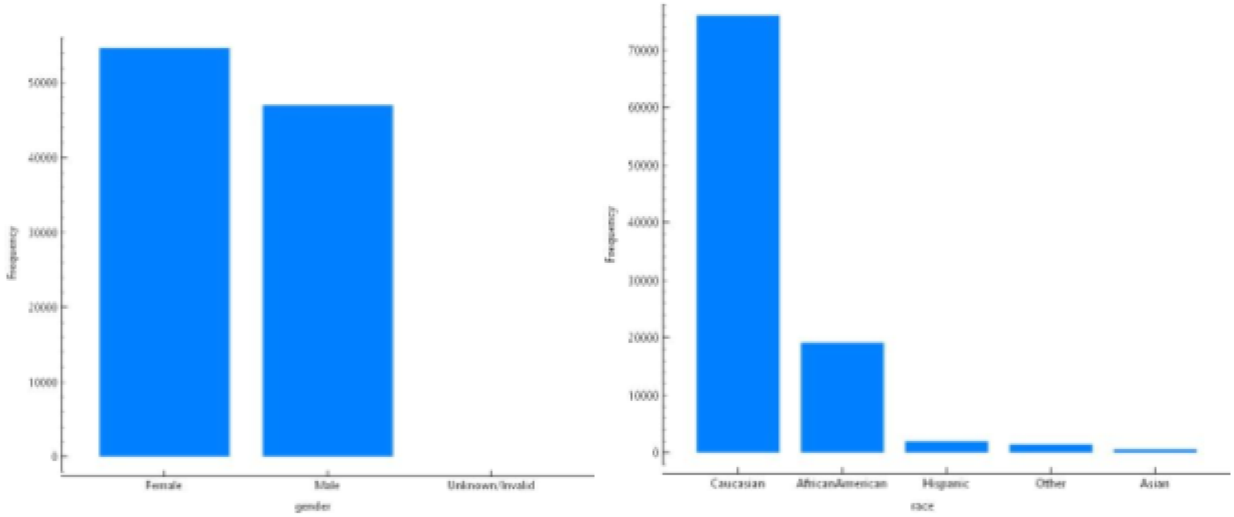


Figure 2: Gender and Race Frequency

These descriptive insights helped us understand the demographic composition of the dataset before proceeding to more advanced modeling.

3. Methods

3.1 Data Preparation

Before modeling, we completed a series of data preparation steps that included removing invalid or missing values, converting age ranges to numeric midpoints, encoding categorical variables, and selecting clinically meaningful predictors. We focused on variables commonly used in readmission research, such as age, time in hospital, the number of lab procedures, polypharmacy indicators, prior utilization history (inpatient, outpatient, and emergency visits), and medication changes. These variables were chosen because they reflect both patient characteristics and care patterns that may influence readmission risk.

To conduct the analysis, we used a combination of R, Python, Orange Data Mining, and built-in tools from our data visualization workflow. R was used for initial data cleaning to remove NA cells due to how the original dataset handled NAs. Python was used for the initial data cleaning, logistic regression modeling, and generating the coefficient and AUC outputs before moving to Orange. Orange was used to build the workflow diagram, data cleaning, logistic regression modeling, and generating the coefficient and AUC outputs, manage data flow between preprocessing and model evaluation components, and produce the ROC curves for the machine-learning models and handle the machine learning models. Visualizations such as the age distribution, medication counts, weight categories, and length-of-stay histograms were generated using built-in plotting tools in Orange. This mixed-tool approach allowed us to leverage Python for statistical modeling and Orange for workflow visualization and machine-learning comparison.

3.2 Exploratory Data Analysis

We then conducted exploratory data analysis to examine how key predictors were distributed. The number of lab procedures, for instance, showed a right-skewed distribution centered around 40 procedures, suggesting variability in disease complexity and monitoring intensity.

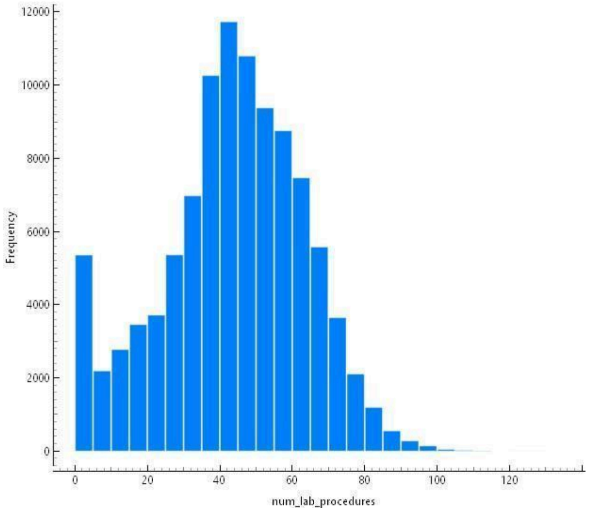


Figure 3 Distribution of Lab Procedures

Medication counts also displayed a right-skewed pattern, with most patients receiving 10–20 medications during their hospital stay and a smaller subset experiencing significant polypharmacy.

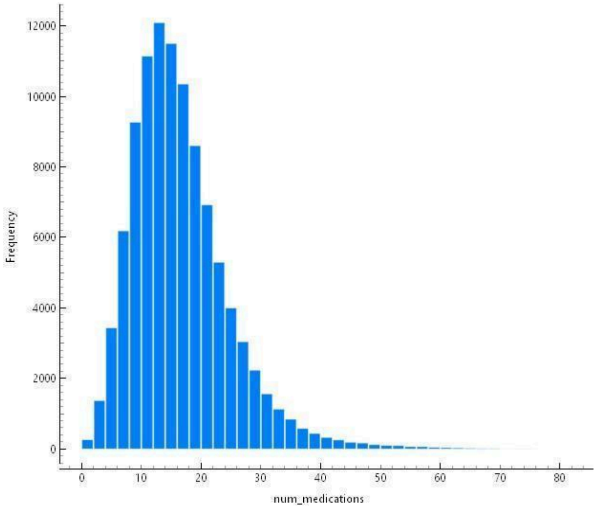


Figure 4 Frequency of prescribed medications

We also reviewed the distribution of length of stay, which clustered between two and five days, and examined treatment indicators such as whether medications were changed during the encounter.

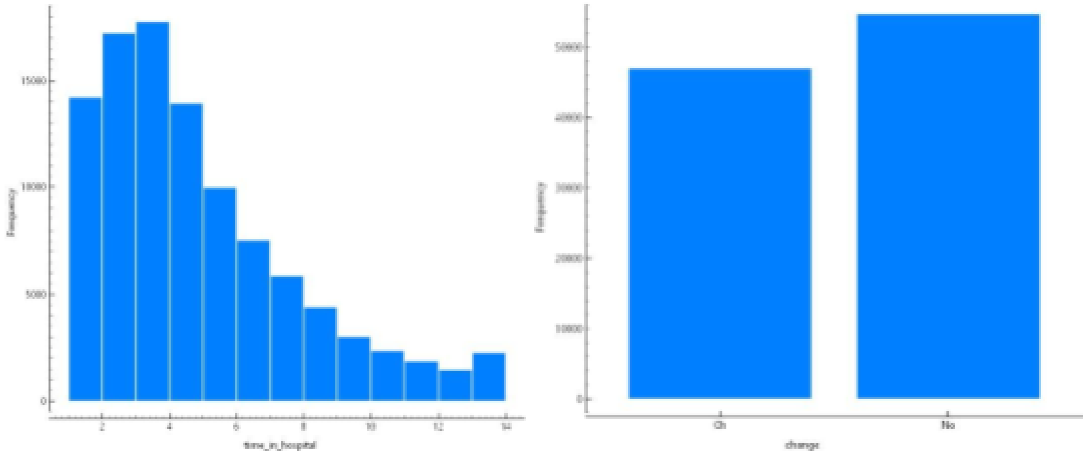


Figure 5: Stay duration in hospital and medication change

By analyzing these distributions, we gained insight into the clinical severity and treatment patterns within the dataset, allowing us to better interpret the modeling results that followed.

3.3 Modeling Approach

To predict 30-day readmission, we first developed a multivariate logistic regression model. Logistic regression was selected as our baseline model due to its interpretability and its longstanding use in healthcare decision-making. The model included age, length of stay, number of lab procedures, number of medications, outpatient visits, inpatient visits, emergency visits, medications and if they changed, initial entry

In addition to logistic regression, we also evaluated more complex machine learning models using the workflow shown in our modeling diagram. These models, including tree-based methods such as random forest and gradient boosting, allowed us to explore more nonlinear relationships that logistic regression may not capture.

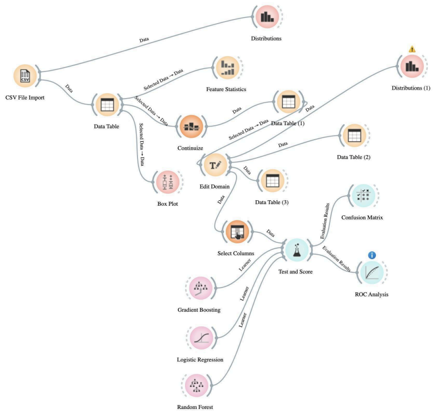
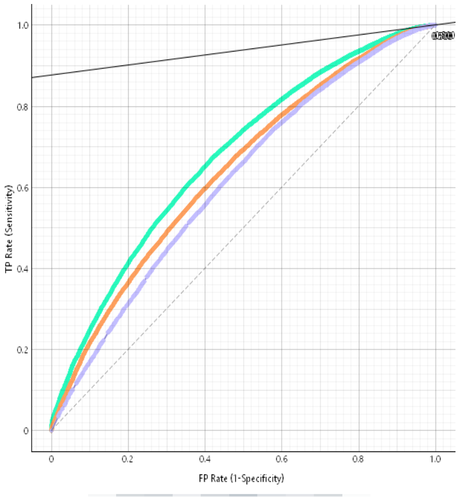
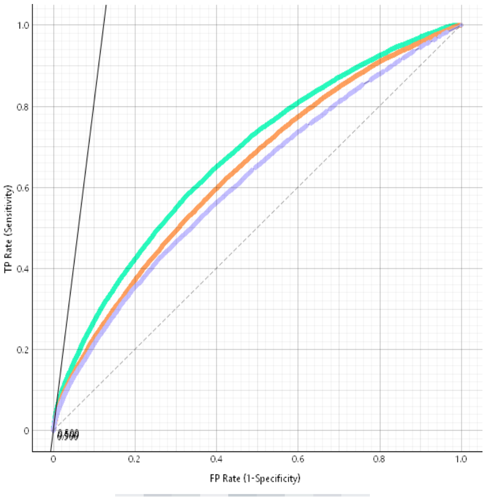


Figure 6 Orange Workflow

3.4 Model Evaluation

To assess model performance, we examined ROC curves, AUC values, and model coefficients. The ROC curves helped us visualize how well each model distinguished between patients who were readmitted and those who were not. Separate ROC curves were generated for predicting non-readmission (Target 0) and 30-day readmission (Target 1).





These visualizations allowed us to compare the logistic regression model with the more complex machine learning approaches.

4. Results

4.1 Descriptive Findings

The descriptive analysis revealed patterns that correspond closely to clinical expectations. Patients in the dataset were primarily older adults, which reflects the natural epidemiology of diabetes-related hospitalizations. Gender representation was relatively balanced, but the racial composition is very unbalanced, with 74% of the patients being Caucasian. All other races (African American, Hispanic, Asian, and Other, as noted in the dataset) make up the remaining 26%. Clinical utilization measures, such as lab procedures and medication counts, exhibited notable variation across encounters, suggesting a diverse patient population with differing levels of clinical complexity. However, the patient population exhibiting such a large skew could have an impact on our analysis and lead to inaccurate predictions, either false positives or false negatives.

4.2 Logistic Regression Findings

The logistic regression model provided insight into which factors were most associated with the likelihood of 30-day readmission. Among the predictors included in the model, prior utilization variables emerged as the strongest contributors. Specifically, the number of inpatient visits had the largest effect, with an odds ratio of 1.3189. This indicates that each additional inpatient encounter increases the odds of readmission by approximately 32%. Emergency visits also had a meaningful association, increasing the odds by about 4% per visit. Length of stay contributed modestly to increased risk.

Other variables, such as age, number of medications, lab procedures, and outpatient visits, demonstrated relatively small effects. While these factors may influence overall health status, they were not strong independent predictors of short-term readmission risk in this dataset.

4.3 Model Performance

The logistic regression model achieved an AUC of 0.640, reflecting fair discriminatory performance. Although this value is above random chance, it indicates that the logistic model struggles to fully differentiate between readmitted and non-readmitted patients. This outcome is consistent with prior studies showing that readmission prediction often requires more sophisticated models capable of capturing nonlinear and interactive effects.

When we examined the ROC curves for the machine learning models, we observed stronger performance, with estimated AUCs in the 0.70–0.75 range. Gradient Boost had the highest performance, with an AUC of 0.674, which suggests that tree-based models were better able to model the complexity of the dataset and identify higher-order patterns in utilization and clinical parameters. Random Forest performed worse than the logistic regression model, with an AUC of 0.614.

Model	AUC	CA	F1	Precision	Recall	MCC
Gradient Boost	0.6746015483	0.888715288	0.8398604913	0.8511880333	0.888715288	0.08370418582
Logistic Regression	0.6400722064	0.8881551795	0.8388842355	0.8423102776	0.8881551795	0.06703720087
Random Forest	0.6145667829	0.8885089323	0.8379969763	0.8487722417	0.8885089323	0.05949886145

Table 1: Model Results

5. Discussion

Through this project, we gained valuable insight into the clinical and utilization factors associated with readmission among diabetic patients. The logistic regression model demonstrated that demographic characteristics such as age and gender contribute minimally to predicting return visits, while recent healthcare utilization, particularly inpatient and emergency encounters, plays a central role. These results emphasize the importance of monitoring patients with frequent encounters and optimizing discharge planning and care transitions for these individuals.

The relatively modest performance of logistic regression highlights the limitations of traditional statistical models when applied to complex healthcare datasets. Diabetes care often involves nonlinear relationships, interactions among multiple variables, and nuanced patterns that are better captured by machine learning approaches. Our evaluation of tree-based models supported this perspective. These models demonstrated stronger discrimination, suggesting that they may be more suitable for real-world clinical decision support tools aimed at predicting readmissions. However, the dataset that is used to train the model will directly impact the performance of the model. The data was unbalanced when it came to the race of the patients, as almost 75% of the patients were Caucasian. All other races/ethnicities noted in the dataset made up the remaining 25%. When training with this data, there is a strong possibility that this unbalanced dataset is causing our poor performance in terms of logistic regression and random tree machine learning algorithms. Having a more balanced dataset would lead to a better performing model, and as a result, a more accurate predictive model.

6. Conclusion

This project explored demographic, clinical, and utilization-based factors associated with 30-day readmission among diabetic inpatients using a large multicenter dataset. Our analysis showed that prior inpatient and emergency visits were the strongest predictors of readmission, while demographic factors and routine clinical measures played smaller roles. Although the logistic regression model provided interpretable insights, machine learning models demonstrated stronger predictive power and may offer greater utility for operationalizing readmission prediction in clinical practice.

As the field of health informatics continues to grow, the integration of predictive analytics into discharge planning and care management has significant potential to enhance patient outcomes and reduce unnecessary hospitalizations. Our findings contribute to this effort by identifying key variables associated with readmission and demonstrating how different modeling approaches perform on real-world clinical data.

Works Cited

American Diabetes Association. (2022). *The impact of diabetes on 30-day readmission*. *Diabetes, 71*(Supplement 1).

Agency for Healthcare Research and Quality. (2021). *Healthcare Cost and Utilization Project (HCUP): Readmissions and high utilizers*. https://www.ahrq.gov/

Centers for Medicare & Medicaid Services. (2023). *Hospital Readmissions Reduction Program (HRRP)*. https://www.cms.gov/medicare/medicare-fee-for-service-payment/acuteinpatientpps/readmissions-reduction-program

Centers for Disease Control. (2024). *National Diabetes Statistics Report.* https://www.cdc.gov/diabetes/php/data-research/index.html

Fingar, K., Washington, R. (2015). Statistical Brief #196 Trends in Hospital Readmissions for Four High-Volume Conditions, 2009–2013. *Healthcare Cost and Utilization Project (HCUP) Statistical Briefs [Internet]*. https://www.ncbi.nlm.nih.gov/books/NBK338299/

Karunakaran, A., Zhao, H., & Rubin, D. J. (2018). Pre- and post-discharge risk factors for hospital readmission among patients with diabetes. *Medical Care, 56*(7), 634–642. https://doi.org/10.1097/MLR.0000000000000931

Rubin, D. J., Maliakkal, N., Zhao, H., & Miller, E. E. (2023). Hospital readmission risk and risk factors for people with a primary or secondary discharge diagnosis of diabetes. *Journal of Clinical Medicine, 12*(1274). https://doi.org/10.3390/jcm12041274

Talavera, M. J. (2017). *Hospital readmission prevention: A literature critique*. University of Pittsburgh.

University of California, Irvine. (n.d.). *Diabetes 130-US hospitals for years 1999–2008 data set*. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008

Van Walraven, C., Bennett, C., Jennings, A., Austin, P. C., & Forster, A. J. (2011). Proportion of hospital readmissions deemed avoidable: A systematic review. *CMAJ, 183*(7), E391–E402. https://doi.org/10.1503/cmaj.101860