Data Exploration and System Management Project on

Patient Readmission prediction using H20 AutoML

Submitted By: Natasha Shereen Benita **Program:** Master's in Artificial Intelligence **Enrollment No:** 5002516

BTU Cottbus-Senftenberg

Brandenburgische Technische Universität Cottbus - Senftenberg

UNDER THE GUIDANCE OF

Prof. Dr. Ireneusz Jablonski

1.Introduction

Patient readmission prediction is a critical challenge in healthcare, impacting hospital resource allocation, patient care, and overall medical costs.

In this study, **H2O AutoML** is employed to automate the modeling process and identify patterns in patient data to predict **critical discharge**. The study aims to determine the effectiveness of this approach in predicting critical patient outcomes and reducing hospital readmission rates.

The best-performing model, **Gradient Boosting Machine (GBM)**, achieved an **accuracy of 99%** and an **AUC of 1**, demonstrating exceptional predictive performance.

2.Literature Review

Prior studies have used **logistic regression, decision trees, and deep learning** to predict hospital readmission, but traditional models struggle with complex medical data. More recently, **machine learning techniques like GBM, XGBoost, and Random Forests** have improved predictive performance, particularly in handling large healthcare datasets.

AutoML frameworks, such as **H2O AutoML**, simplify the process of model selection and tuning, making predictive analytics more accessible in healthcare settings. Research has also shown that **discharge outcomes** are strong indicators of readmission risk. By focusing on **critical discharge prediction**, this project aligns with existing studies advocating for early identification of high-risk patients to enhance post-discharge care and reduce readmission rates.

3.Material

3.1 Dataset

The dataset used for this study is the **eICU Collaborative Research Database Demo** from **PhysioNet**. The eICU database is a **publicly available, de-identified critical care dataset**

containing information from multiple intensive care units (ICUs) across the United States. It includes patient demographics, vital signs, lab results, medications, diagnoses, and discharge details.

For this project, **key tables such as patient, admissionDx, diagnosis, lab,treatment** were used to extract relevant features for predicting **critical discharge status**, which was used as the target variable.

3.2 Libraries

The following Python libraries were utilized for data preprocessing, feature engineering, model training, and evaluation:

- **Pandas & NumPy** – For data loading, manipulation, and preprocessing
- **Dask** – For handling large datasets efficiently
- **Matplotlib & Seaborn** – For data visualization
- **Scikit-learn** – For feature scaling, model evaluation, and preprocessing
- **H2O AutoML** – For automated machine learning model training and selection
- **Polars & PyArrow** – For optimized dataframe operations

4.Methodology

4.1 Data Preprocessing

To prepare the data for predictive modeling, several preprocessing steps were undertaken to ensure data quality, consistency, and optimal memory usage. The steps are outlined below:

4.1.1 Dataset Loading and Inspection:

The datasets (admissionDX, patient, lab, diagnosis, and treatment) were loaded into Pandas dataframes. The medication table was excluded due to its high memory usage and an excessive number of unique drug names, which made merging infeasible. Initial inspections were performed to examine the structure, column names, and data types of each dataset.

4.1.2 Column Selection:

Only relevant columns were retained from each dataset to reduce unnecessary memory usage:

a. admissionDX: Patient ID, admission diagnosis name, and admission diagnosis text.
b. patient: Patient ID, age, gender, ethnicity, hospital admission source, discharge location, and discharge status.
c. lab: Patient ID, lab test name, and lab results.
d. diagnosis: Patient ID, diagnosis string, and ICD-9 code.
e. treatment: Patient ID and treatment description.

4.1.3 Removing Duplicates:

Duplicate rows were dropped across all datasets to avoid redundancy and improve efficiency.

4.1.4 Subset Sampling:

A subset of 1,000 random patient IDs was selected to reduce the size of the dataset for faster processing and analysis. Each dataset was filtered to include only these patient IDs, ensuring consistent patient representation across all tables.

4.1.5 Data Type Optimization:

f. Numeric columns (e.g., age, labresult) were converted to float32 to save memory.
g. Categorical columns (e.g., gender, ethnicity, hospitaladmitsource) were converted to category type to further optimize memory usage.
h. Other columns such as patientunitstayid were cast to int32 where applicable.

4.1.6 Merging Datasets:

The datasets were merged sequentially on the patientunitstayid column using a left join. This ensured that no patient data from the base dataset (admissionDX) was lost during merging.

4.1.7 Handling Missing Values:

Missing values were observed in some columns, such as age, hospitaladmitsource, and labresult. These were retained for further analysis to decide on appropriate imputation or handling methods based on their significance in modeling.

4.1.8 Saving the Preprocessed Dataset:

The final preprocessed dataset, containing 1,000 rows and 14 columns, was saved as a CSV file (preprocessed.csv) for further analysis.

4.2 Feature Engineering

Feature engineering was conducted to enhance the dataset and make it suitable for predictive modeling:

4.2.1 Handling Missing Values

a. **Numeric Columns**: Missing values in age and labresult were filled with their mean.

b. **Categorical Columns**: Missing values in columns like ethnicity, hospitaladmitsource, and icd9code were filled with the mode.

4.2.2 New Features

c. **Age Groups**: Patients were categorized into Child, Young Adult, Adult, and Senior based on age.

d. **Text Length**: admitdxtext_length captured the length of admission diagnosis descriptions.

e. **Gender Encoding**: A binary is_female column replaced the gender column.

f. **Combined Diagnosis**: diagnosis_combined merged admitdxname and diagnosisstring for more comprehensive diagnostic data.

e. **Critical Discharge**: The hospitaldischargestatus column was transformed into a binary feature, critical_discharge, where values such as Expired or Critical were assigned 1, and others were assigned 0. This helps model critical outcomes.

4.2.3 Encoding Categorical Features

Label Encoding was applied to columns such as ethnicity and hospitaladmitsource to convert them into numerical form.

4.2.4 Feature Reduction

Redundant columns (gender, admitdxtext, admitdxname, labname,diagnosis_combined) were dropped to reduce noise and prevent multicollinearity.

4.2.5 Imbalanced Data

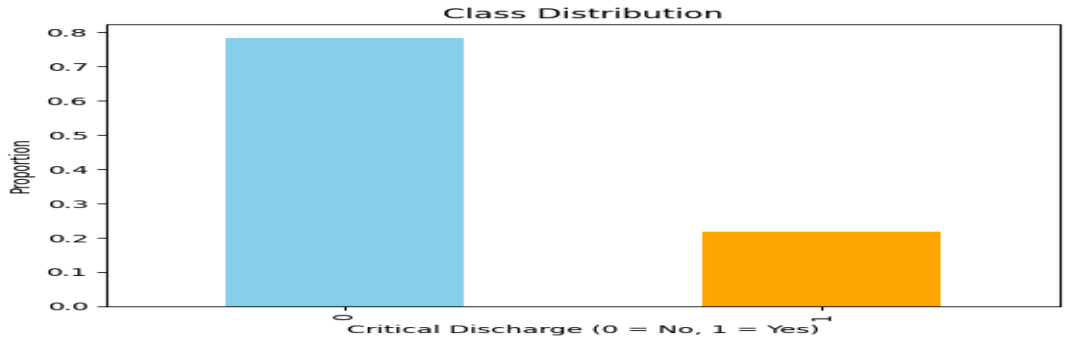
The target variable, critical_discharge, showed class imbalance (78.3% in class 0-no, 21.7% in class 1-yes)

The final dataset was saved as feature_engineered.csv.

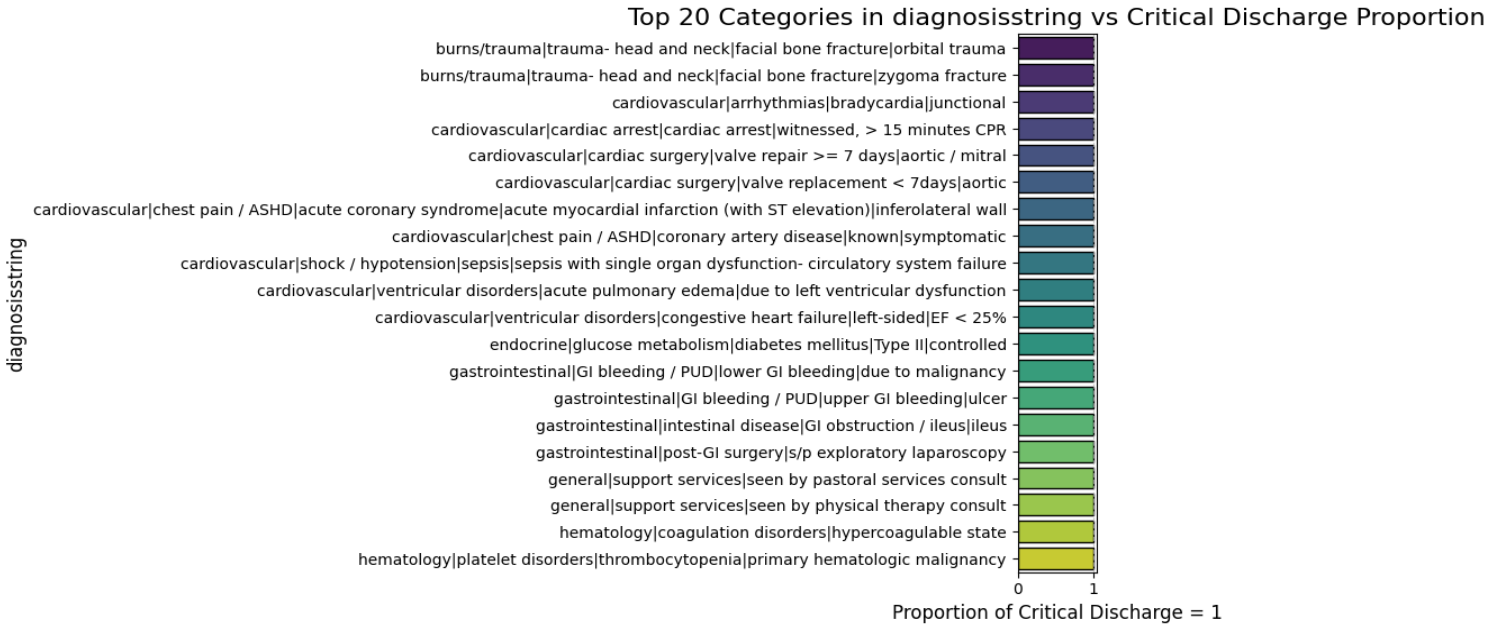
4.3 Exploratory Data Analysis

4.3.1 Class Balance

Critical Discharge (0 vs. 1): Imbalanced distribution (78.3% non-critical, 21.7% critical).

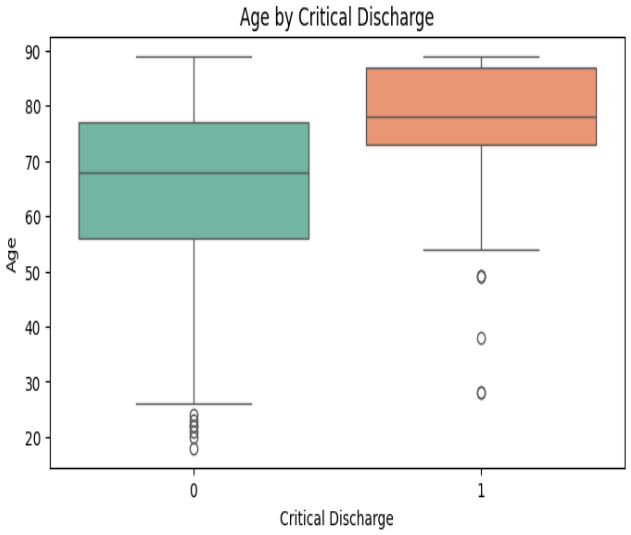


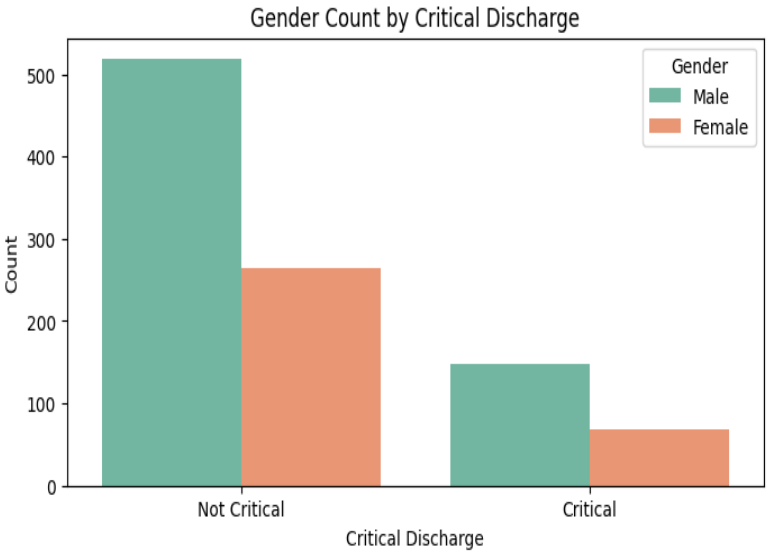
4.3.2 Categories in diagnosisstring vs Critical Discharge Proportion

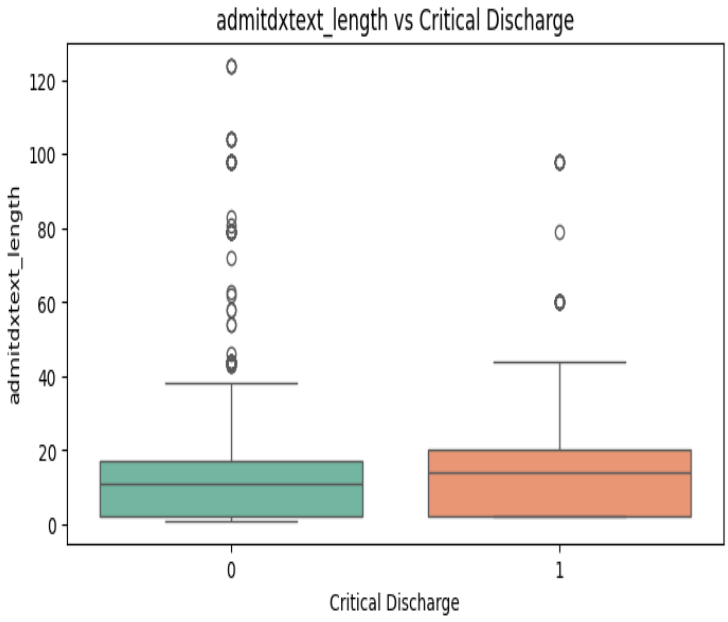


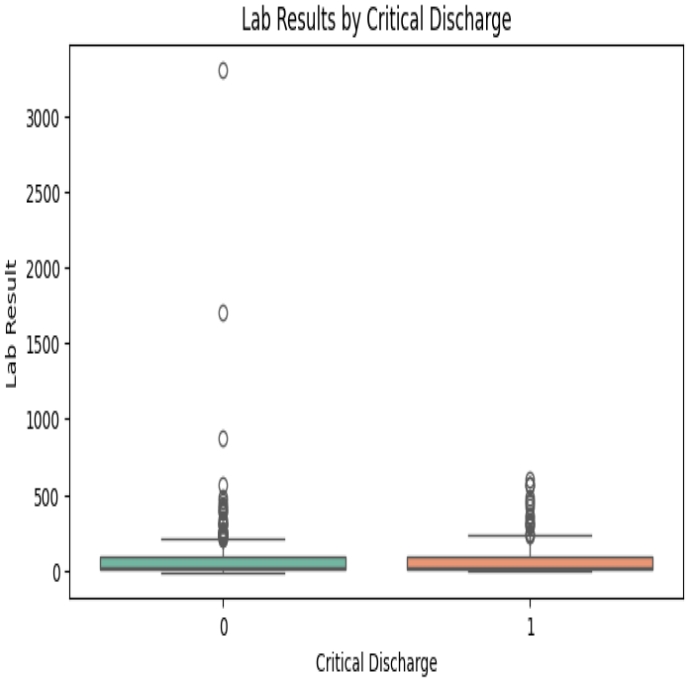
A proportion of 1 indicates that all patients in that category had a critical discharge, while a proportion close to 0 means almost none did. The categories at the top of the chart (e.g., burns/trauma-related diagnoses) have the highest association with critical discharges. This suggests that patients in these categories are more likely to experience readmission. Diagnoses like cardiovascular arrest, acute coronary syndrome, and sepsis are strongly linked with critical discharges, which aligns with their severity and life-threatening nature.

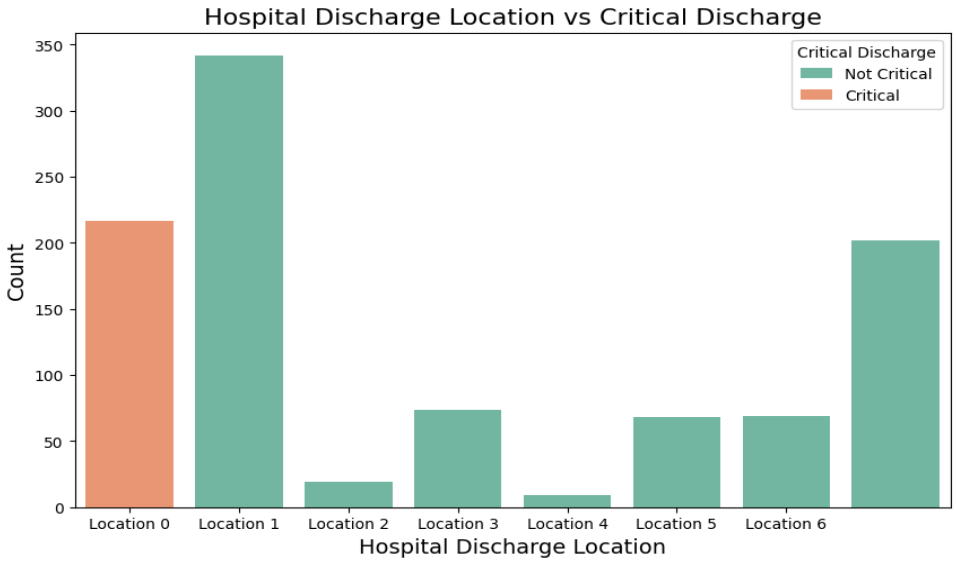
4.3.3 Other Bivariate Analysis











A) **Lab Results vs Critical discharge**

This visualization suggests that the overall distribution for both groups are quite similar, indicating no drastic differences between the lab results of critical and non-critical discharges.

Lab results alone may not strongly differentiate between critical and non-critical discharges since the distributions overlap considerably.

Outliers could represent specific conditions or errors in lab data

B) **Age vs Critical discharge**

Patients who are older often have more chronic conditions, frailty, or complications, which can increase the likelihood of readmission after discharge, especially if discharged in a critical condition.

The boxplot shows that critical discharge is strongly associated with older patients, suggesting age may be an indirect predictor of readmission risk. Since older patients dominate the critical discharge group, it reinforces the notion that older, critically discharged individuals are key targets for readmission prediction models.

C) **Gender vs Critical Discharge**

It does not say much. Just shows there are more men in this subset of this dataset and that corresponds to the class imbalance in critical_discharge. It can also be said that men receive more treatment from the hospitals.

D)**Admittextlength vs critical discharge**

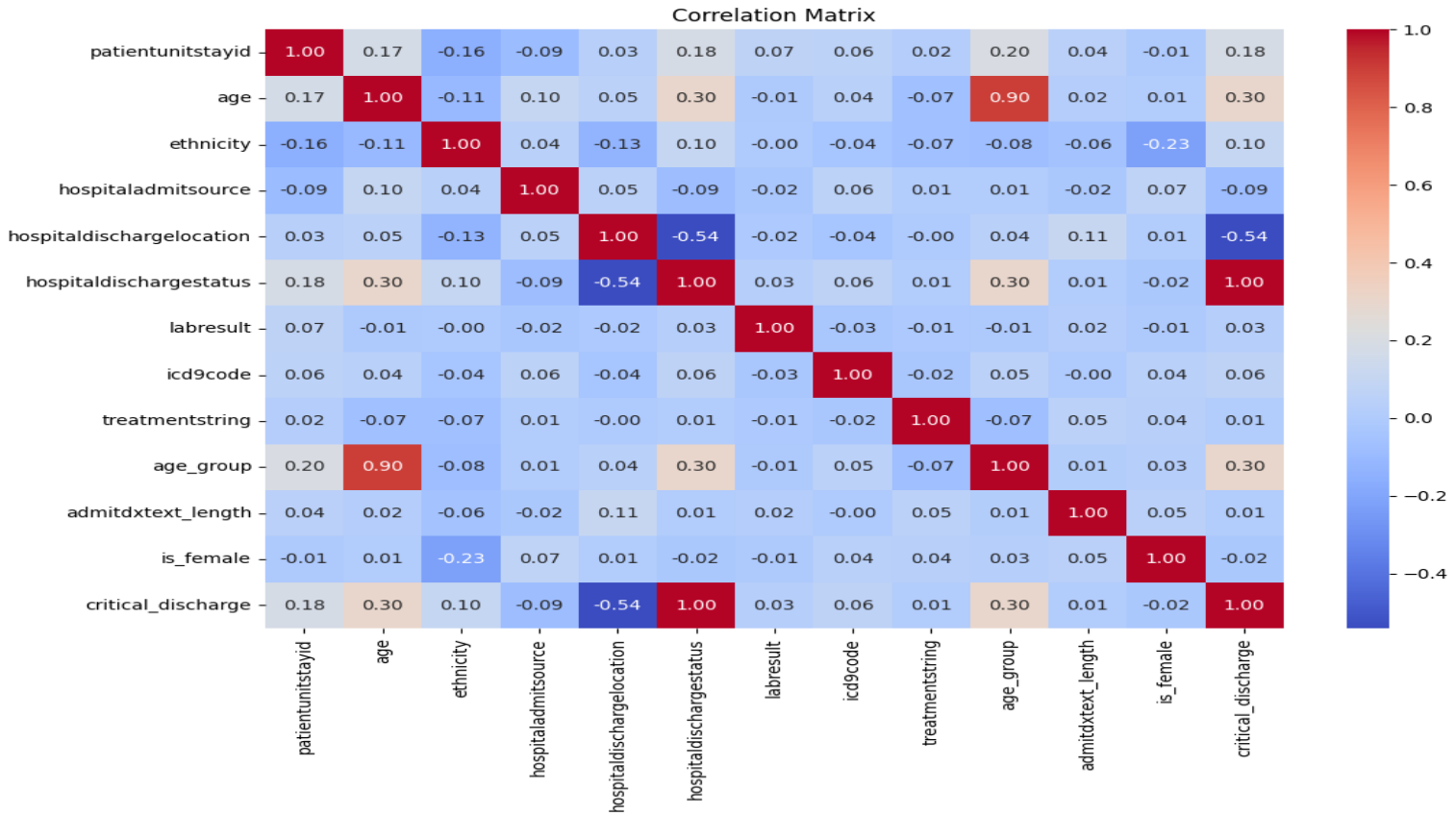
Both groups have a similar median length of admitting diagnosis descriptions, suggesting that on average, the level of detail in the admitting diagnosis does not differ significantly between critically discharged and non-critically discharged patients.

E) **Hospital Discharge Location vs Critical Discharge**

0:Death 1:Home 2:Nursing Home 3:Other 4:Other External 5:Other hospital 6:Rehabilitation

7:Skilled nursing facility . Patients discharged to facilities like nursing homes, skilled nursing, or other hospitals often show higher readmission rates.

4.3.4 Correlation Matrix



5.Results

5.1 Model Training:

1. Import training and testing data (train_data.csv and test_data.csv).
2. Set the target feature (critical_discharge) for patient readmission prediction.
3. Initialize H2O Auto ML and use it to train up to 20 models (max_models=20)

4. Set balance_classes=True so it will automatically handle class imbalance by using class weights in the model to give higher importance to the minority class.

5.2 AML Leaderboard

Best is Gradient Boosting Machine

model_id	rmse	mse	mae	rmsle	mean_residual_deviance
GBM_grid_1_AutoML_1_20250120_172333_model_17	1.39721e-07	1.9522e-14	7.27513e-08	7.05875e-08	1.9522e-14
GBM_grid_1_AutoML_1_20250120_172333_model_19	1.39721e-07	1.9522e-14	7.27513e-08	7.05875e-08	1.9522e-14
GBM_grid_1_AutoML_1_20250120_172333_model_2	1.39721e-07	1.9522e-14	7.27513e-08	7.05875e-08	1.9522e-14
GBM_grid_1_AutoML_1_20250120_172333_model_13	1.39721e-07	1.9522e-14	7.27513e-08	7.05875e-08	1.9522e-14
GBM_grid_1_AutoML_1_20250120_172333_model_1	3.30671e-07	1.09343e-13	8.27802e-08	3.11002e-07	1.09343e-13
GBM_grid_1_AutoML_1_20250120_172333_model_16	0.000634837	4.03018e-07	0.000277034	0.000408196	4.03018e-07
StackedEnsemble_AllModels_1_AutoML_1_20250120_172333	0.00273976	7.50627e-06	0.00156764	0.00172258	7.50627e-06
GLM_1_AutoML_1_20250120_172333	0.0034585	1.19612e-05	0.00285516	0.00282503	1.19612e-05
StackedEnsemble_BestOfFamily_1_AutoML_1_20250120_172333	0.00374225	1.40044e-05	0.00224006	0.00249188	1.40044e-05
GBM_2_AutoML_1_20250120_172333	0.0110038	0.000121084	0.00605146	0.00810294	0.000121084

[30 rows x 6 columns]

5.3 Evaluation Metrics

Optimal Threshold: 0.9931157914848616

```
AUC: 1.0
Logloss: 3.459647956244093e-05
Confusion Matrix:
Confusion Matrix (Act/Pred) for max f1 @ threshold = 0.9931157914848616
       0    1    Error    Rate
-----  ---  ---  -------  -----------
0      157  0    0        (0.0/157.0)
1      0    43   0        (0.0/43.0)
Total  157  43   0        (0.0/200.0)
```

```
Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      1.00       157
           1       1.00      0.98      0.99        43

    accuracy                           0.99       200
   macro avg       1.00      0.99      0.99       200
weighted avg       1.00      0.99      0.99       200
```

6.Conclusion

In this study, we explored the application of H2O AutoML to predict patient readmission cases based on the provided dataset. The automated approach enabled us to quickly identify the most suitable models and evaluate their performance. The top-performing model achieved an AUC of 1.0 and a log loss of 3.459 on the test set, indicating a reliable predictive capacity. Additionally, an optimal threshold to maximize the F1-score, enhancing the balance between precision and recall for the model was identified

6.1 Future Improvements

Data related to expired people can be removed when working with the whole dataset as it will form a significant group. However, for this study it was not excluded since the data suggests otherwise. From the data we can see that some patients are alive and are not critically discharged even though they are categorized under deadly diseases while some others die from minor infections. If removed from this subset valuable information related to other variables and features will be lost. Overall, the numbers correspond to the class imbalance in the target column.

6.2 Bibliography

1. eICU Collaborative Research Database Demo:https://physionet.org/content/eicu-crd-demo/2.0.1/
2. H2o AutoML Documentation: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html#citation

3. H2O AutoML: Scalable Automatic Machine Learning:https://www.automl.org/wp-content/uploads/2020/07/AutoML_2020_paper_61.pdf

4: A Systematic Literature Review of Predicting Patient Discharges Using Machine Learning:https://link.springer.com/article/10.1007/s10729-024-09682-7

5. Machine-learning-based hospital discharge predictions can support multidisciplinary rounds and decrease hospital length-of-stay :https://innovations.bmj.com/content/7/2/414