



Technische
Universität
Braunschweig

MHH
Medizinische Hochschule
Hannover



PETER L.
REICHERTZ INSTITUT
FÜR MEDIZINISCHE
INFORMATIK



Diabetes Readmission Prediction

By

Alisha Sarkar, Rabiya Farheen, Rajanigandha Dutt, Ria Mohan, Athira Sadhasivan

Objective

To build predictive models that determine whether a diabetic patient will be **readmitted within 30 days** of discharge.

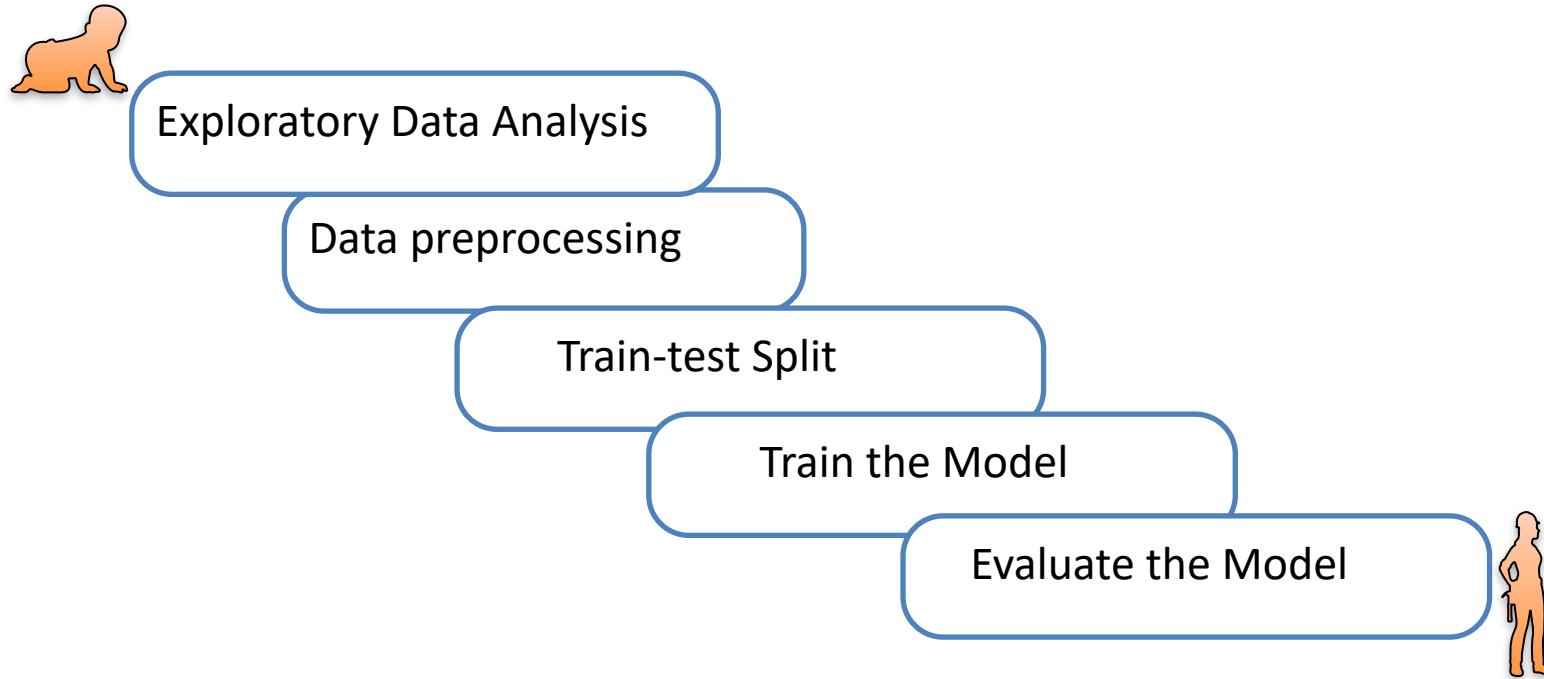
Dataset

- We use the **Diabetes 130-US Hospitals** dataset from the **UCI ML Repository**.
- It contains **101,766 hospital records** of diabetic patients.
- Includes features like **age, gender, lab results, medications, diagnosis, and readmission status**.

Tasks

- Exploratory Data Analysis (EDA)
- Predict Readmission & Length of Stay
- Analyse Risk Factors
- Medication Impact
- Patient Grouping
- Demographic & Treatment Differences

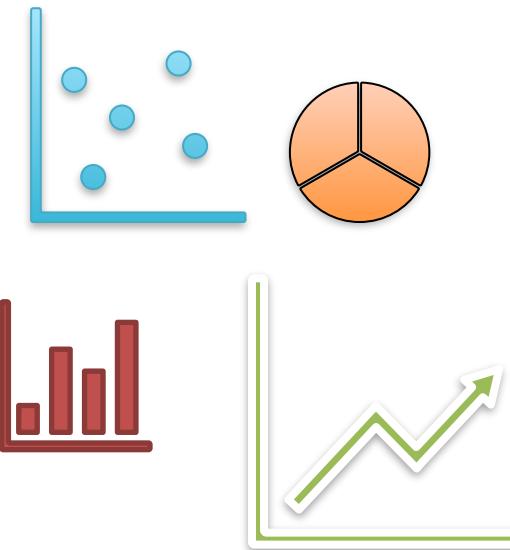
Proposed Methodology



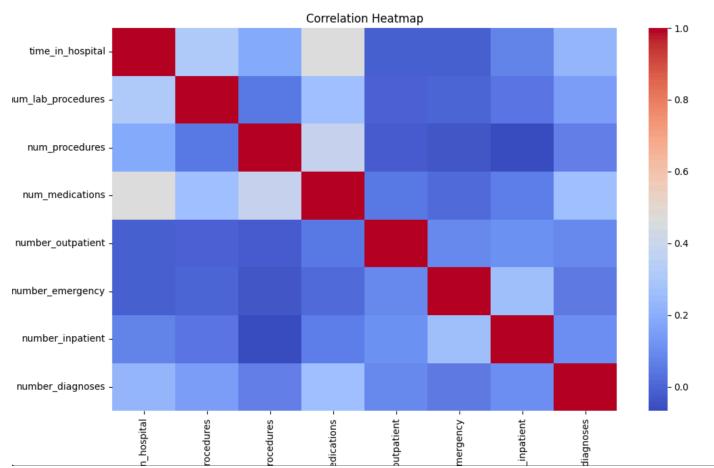
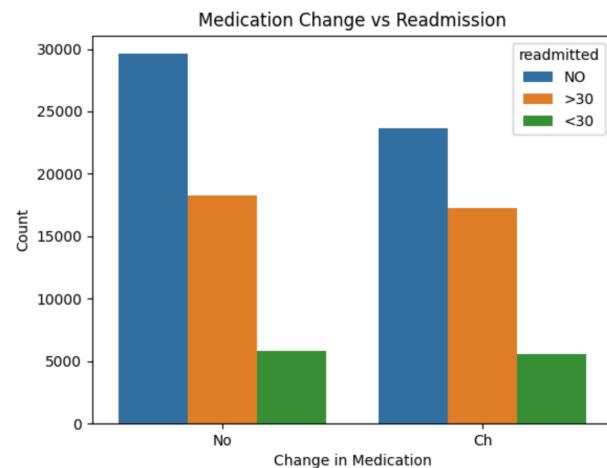
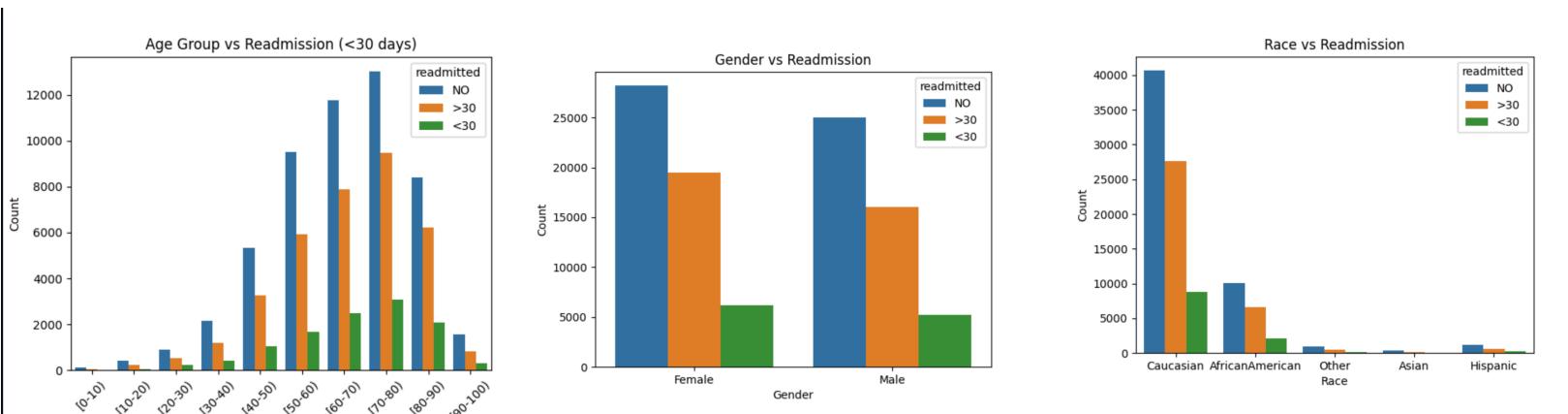
Exploratory Data Analysis

Data

Visualising the data



EDA Insights from the project



Data Preprocessing

Transform raw patient data into a clean, structured format ready for modeling.

Raw Data →

Clean →

Feature Mapping & filter→

Encoding and Transformation →

Save

Data Split for Model Training

Load Clean Data:

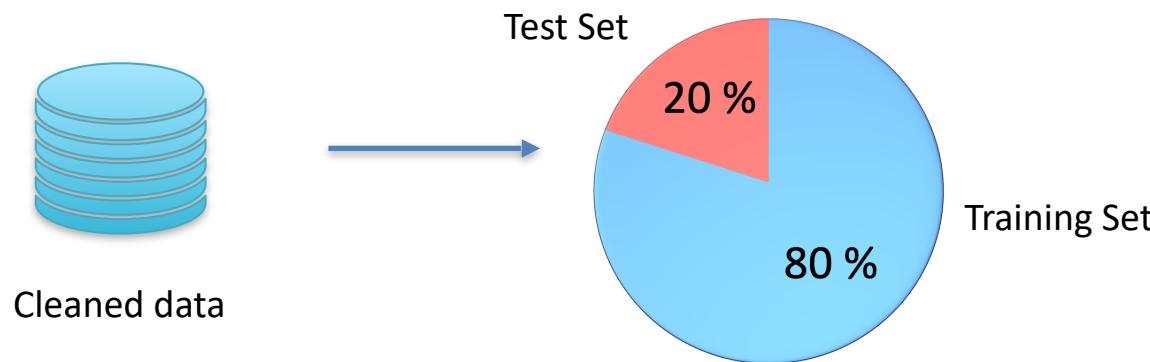
- Read the cleaned dataset.

Define Features & Target:

- $X \rightarrow$ all relevant features (demographics, labs, meds).
- $y \rightarrow$ target variable: **readmitted_30days**.

Train-Test Split:

- 80% training, 20% testing.



Model Training - Logistic Regression

What is it?

- A **supervised classification algorithm** used to predict a **binary outcome** (e.g., readmitted: **Yes/No**).

How does it work?

- Estimates the **probability** that an observation belongs to a class using a **logistic (sigmoid) function**.
- Outputs probabilities → applies a **threshold** to classify into **0 or 1**.

Why use it?

- Simple, **interpretable**, and works well as a **baseline model**.
- Coefficients show **feature impact** on readmission risk.

Model Training - Random Forest

What: An ensemble of many decision trees.

How: Builds trees on random data & features, combines votes.

Why: High accuracy, reduces overfitting, shows feature importance.

Model Training - XGBoost

What: Extreme Gradient Boosting

How: Builds decision trees **sequentially**, each new tree **corrects errors** of previous ones.

Why use it: High predictive **accuracy**, handles missing data well, faster and more efficient.

Prediction of Readmission and Length of stay

Goal: Predict if the patient is readmitted and for how many days a diabetic patient is likely to stay in the hospital.

Why?

- Helps hospitals plan beds & staff.
- Supports better patient care.
- Can reduce unnecessary costs.

Prediction of Readmission and Length of stay

Length of Stay Regression

Select Regressor Model

Linear Regression

Run Regression

Linear Regression

Mean Absolute Error (MAE): 1.7887
Root Mean Squared Error (RMSE): 2.3706

- Assumes a linear relationship between features and hospital stay duration.

Length of Stay Regression

Select Regressor Model

Random Forest

Run Regression

Random Forest Regression

Mean Absolute Error (MAE): 1.7112
Root Mean Squared Error (RMSE): 2.2889

- Uses an ensemble of decision trees for prediction.

Length of Stay Regression

Select Regressor Model

XGBoost

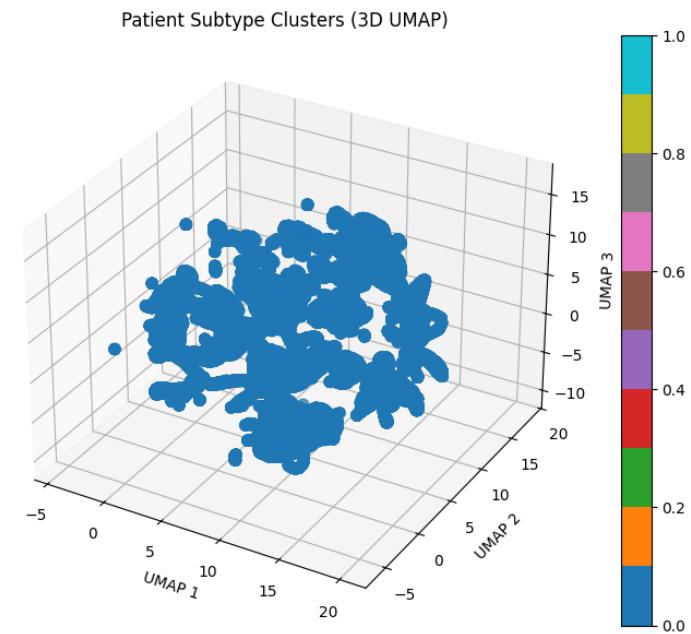
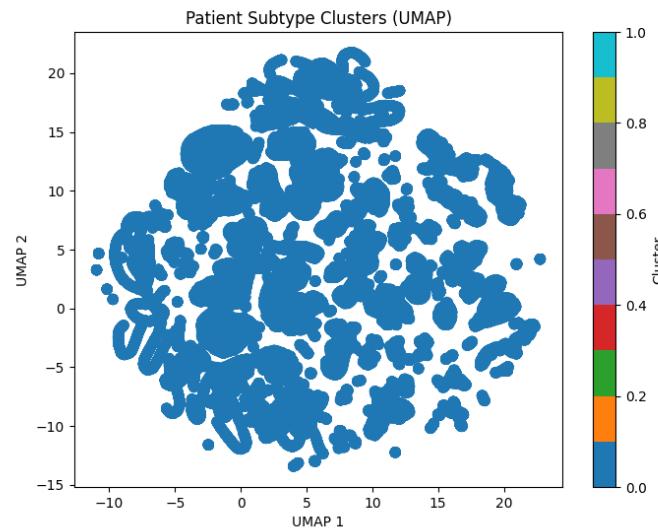
Run Regression

XGBoost Regression

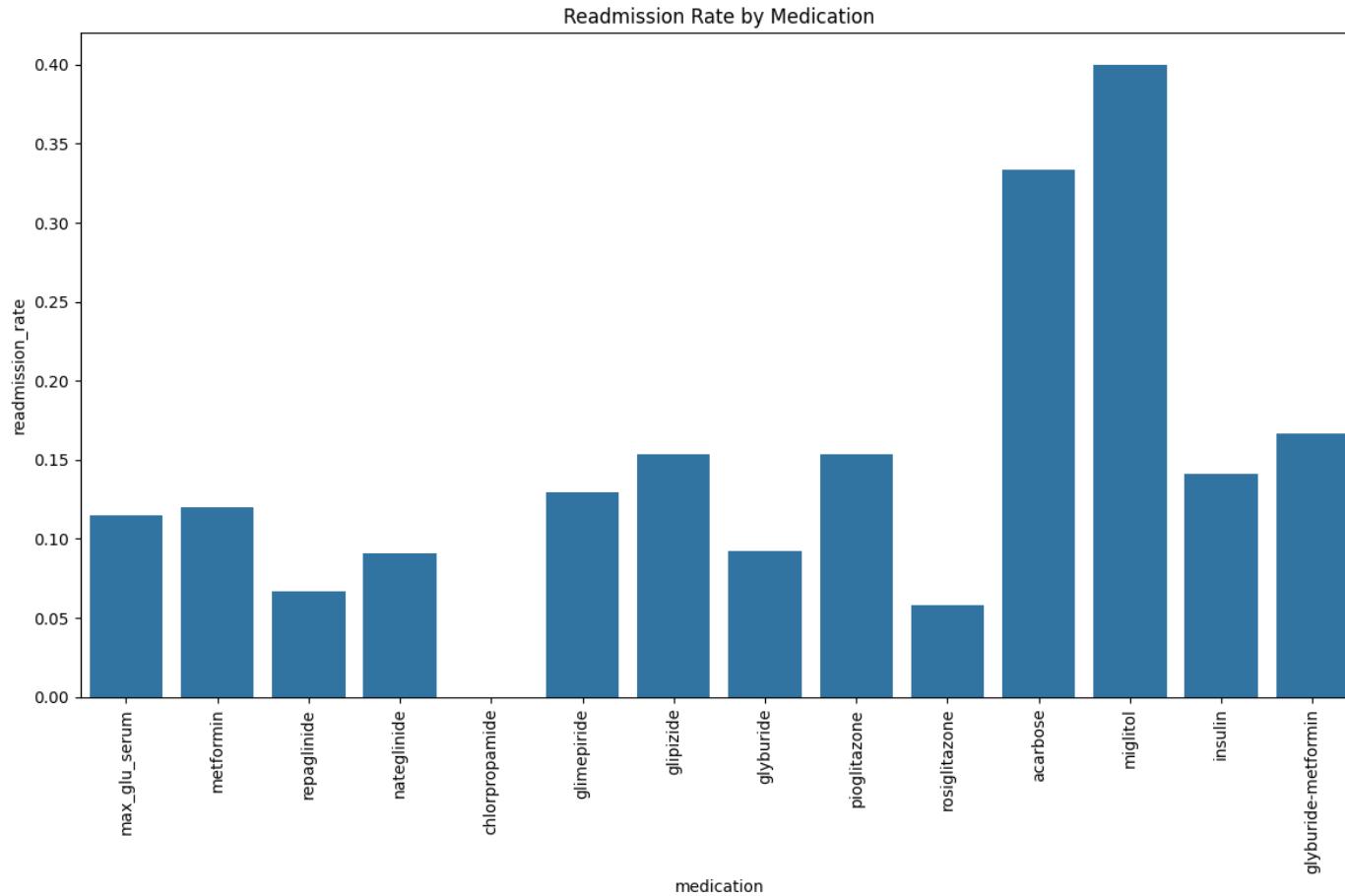
Best Hyperparameters: {'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 200, 'subsample': 0.8}
Mean Absolute Error (MAE): 1.6547
Root Mean Squared Error (RMSE): 2.2260

- Uses gradient-boosted trees optimized via GridSearchCV.

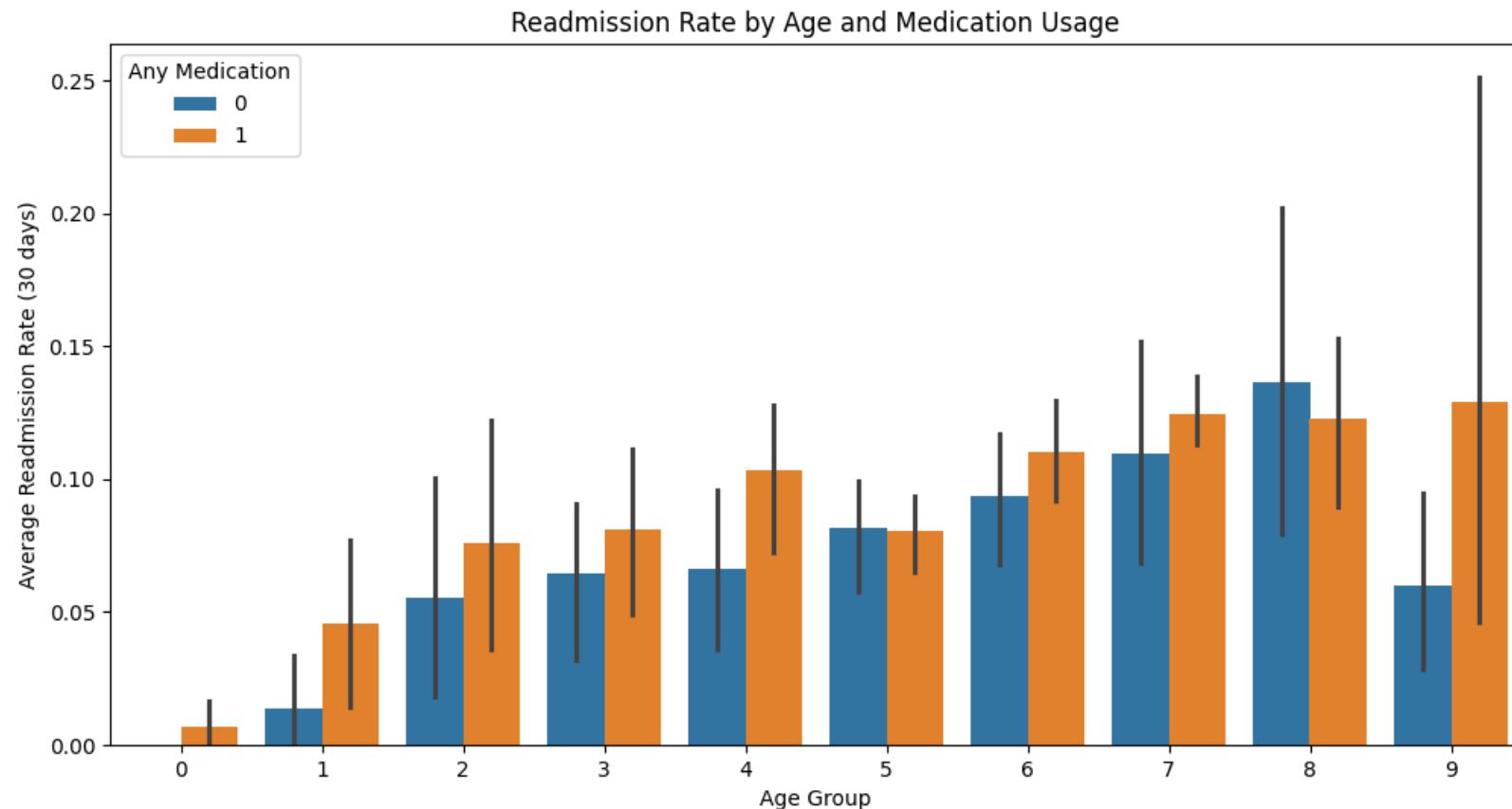
Insights - Patient Grouping



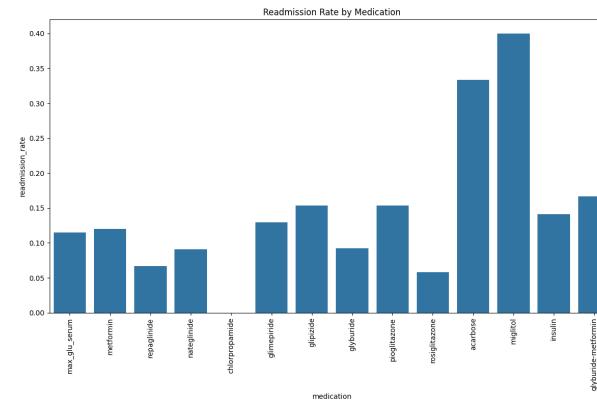
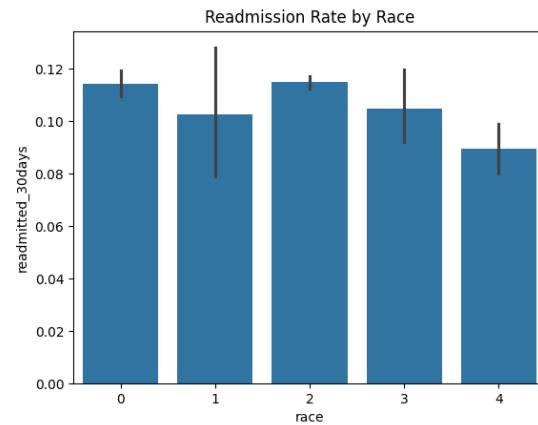
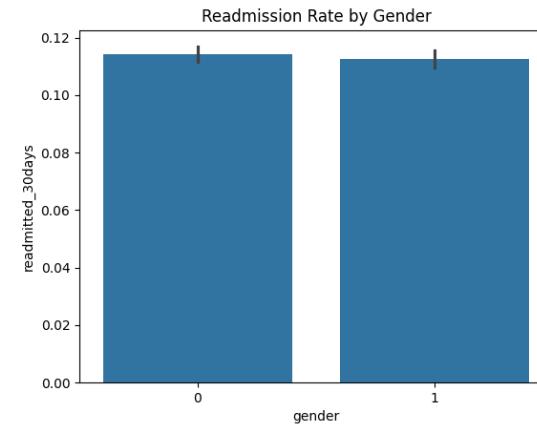
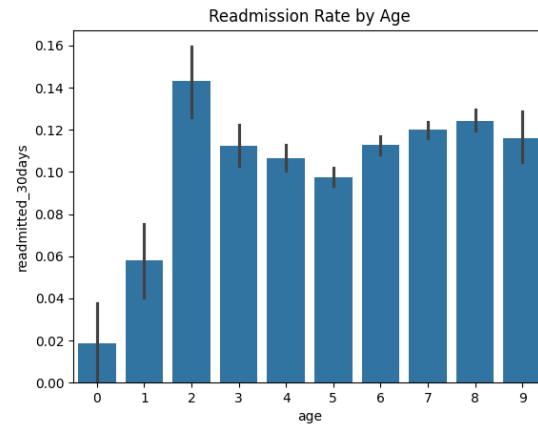
Medication Impact



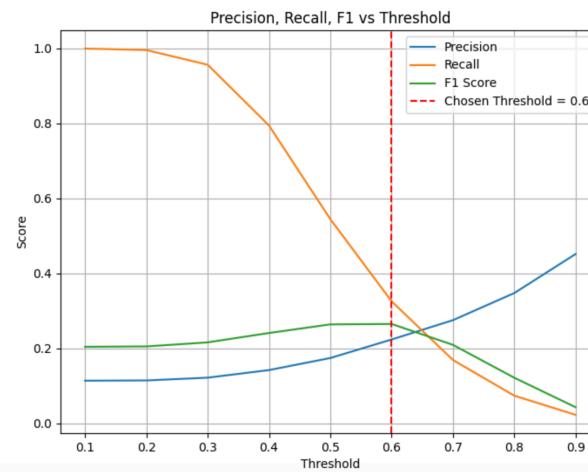
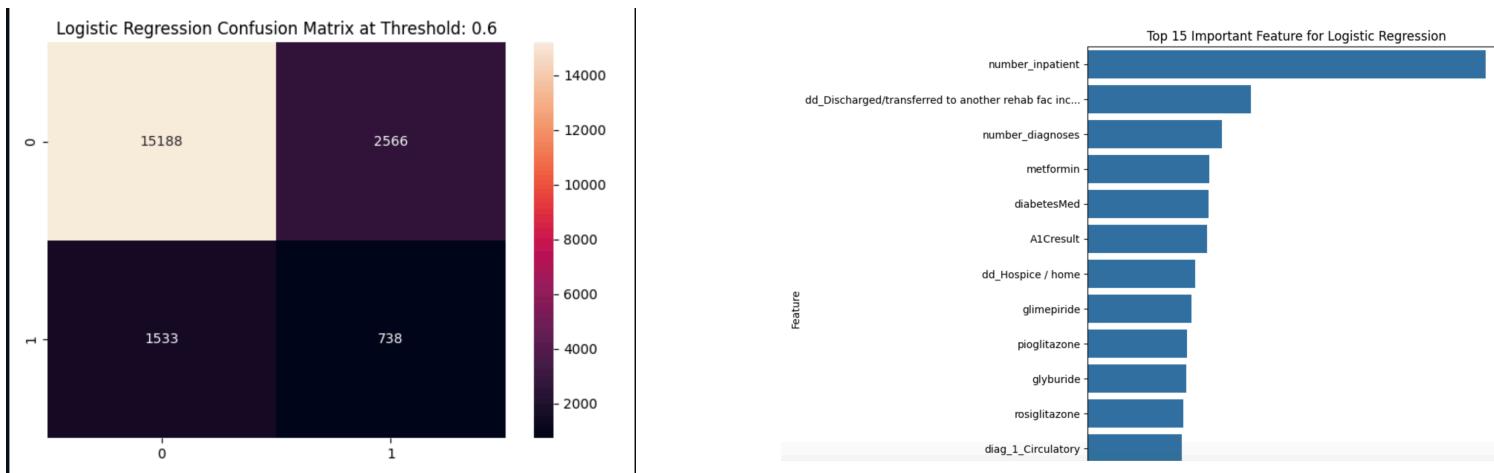
Analyse Risk factors



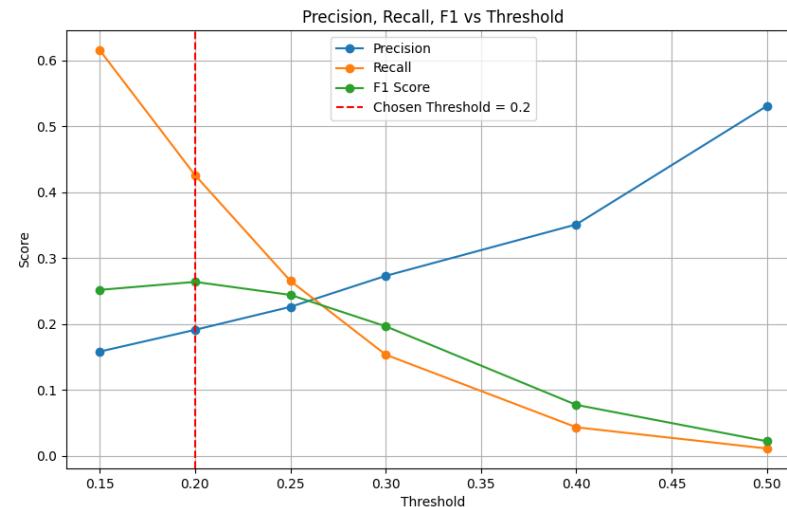
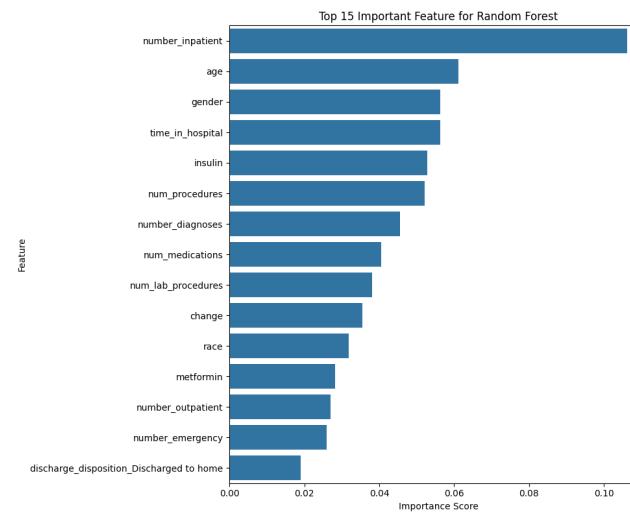
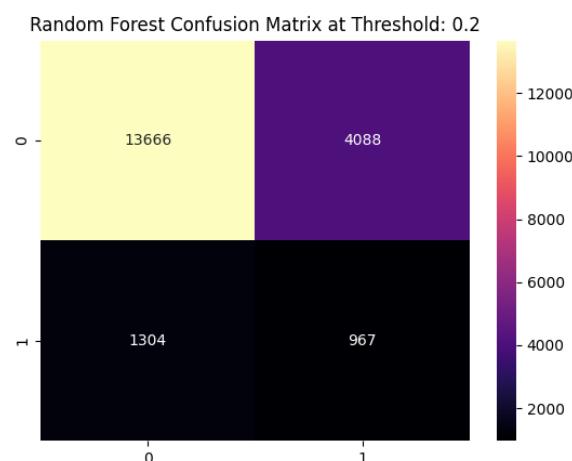
Demographic & Treatment Differences



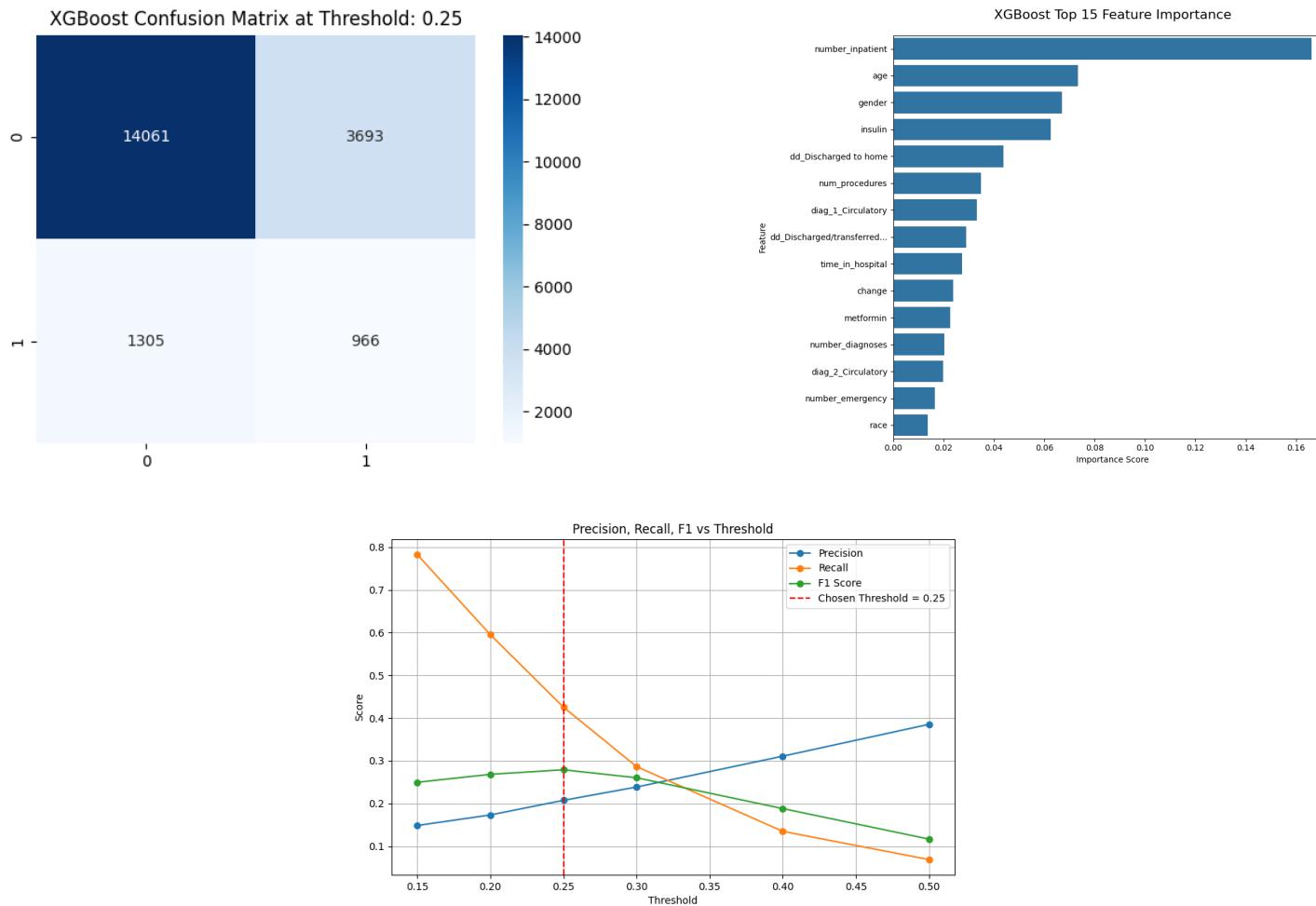
Model Training - Logistic Regression



Model Training -Random Forest



Model Training - XGBoost



GUI Snapshots

The screenshot shows the 'Predict Readmission' interface. On the left, a sidebar titled 'Navigate' includes links for Home, Upload CSV, EDA, Readmission Prediction, Feature Insights, Group Analysis, Patient Clusters, and Length of Stay. The main area has a title 'Predict Readmission' with a subtitle 'Choose Prediction Model' set to 'Logistic Regression'. Below this, a section titled 'Enter patient information below:' contains two groups: 'Demographics' (Age Group [0-10], Gender [Male], Race [Caucasian]) and 'Hospital Encounter' (Admission Type [Elective], Discharge Disposition [Discharged home]).

This screenshot shows the results of a prediction. The 'Medications' section lists several drugs with their status: Metformin (Yes), Glimepiride (Yes), Pioglitazone (Yes), Glyburide (Yes), and Rosiglitazone (Yes). A 'Predict' button is visible. Below it, a message states 'Model selected: Logistic Regression' and 'Prediction: Patient is most likely to be readmitted.' with a predicted probability of 0.97.

The screenshot shows the 'Length of Stay Regression' interface. It features a logo with a cross and a plus sign. The main title is 'Length of Stay Regression'. Under 'Select Regressor Model', 'XGBoost' is selected. A 'Run Regression' button is present. The results section displays the following text:
Best Hyperparameters: {'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 200, 'subsample': 0.8}
Mean Absolute Error (MAE): 1.6547
Root Mean Squared Error (RMSE): 2.2260
- Uses gradient-boosted trees optimized via GridSearchCV.

Challenges

Data Challenges (EDA)

- Dealing with **missing values** in patient records.
- Handling **imbalanced classes** (very few readmitted vs. many not readmitted).
- Understanding & cleaning **categorical variables** (e.g., discharge status, admission type).

Feature Engineering

- Selecting meaningful variables among many hospital & treatment factors.
- Encoding complex categorical features correctly.
- Reducing high cardinality (e.g., diagnosis codes).

Model Training

- Avoiding **overfitting** with small/imbalanced data.
- Tuning hyperparameters for models like **XGBoost** and **Random Forest**.
- Choosing **thresholds** to balance **precision & recall**.
- Ensuring **interpretability** (doctors need to trust the model's decisions).

References

- [1] Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., & Kripalani, S. (2011). Risk prediction models for hospital readmission: A systematic review. *JAMA*, 306(15), 1688–1698. <https://doi.org/10.1001/jama.2011.1515>
- [2] Hasan, M. M., Haque, M. M., Rahman, M. M., & Moniruzzaman, M. (2019). Predictive modeling of hospital readmissions using metaheuristics and data mining. *Expert Systems with Applications*, 127, 170–186. <https://doi.org/10.1016/j.eswa.2019.03.051>
- [3] LucienCastle. *Diabetes Patient Readmission Prediction*. GitHub repository. Available at: <https://github.com/LucienCastle/diabetes-patient-readmission-prediction>
- [4] Laurenemilyto. *Predicting Hospital Readmission*. GitHub repository. Available at: <https://github.com/laurenemilyto/predicting-hospital-readmission>

Thank You



PETER L.
REICHERTZ INSTITUT
FÜR MEDIZINISCHE
INFORMATIK