Technische Universität Braunschweig

Medizinische Hochschule Hannover

PLRI 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 preprocessing

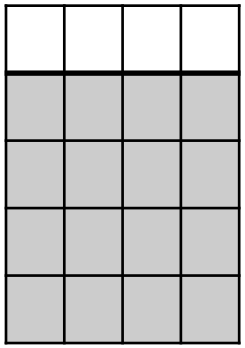
Train-test Split

Train the Model

Evaluate the Model

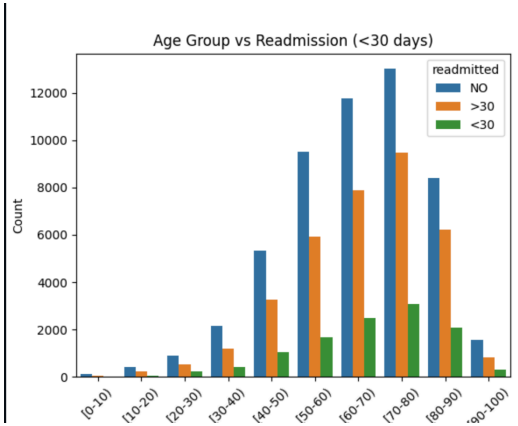
Exploratory Data Analysis

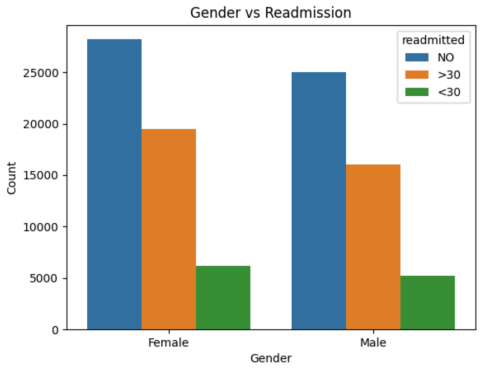
Data

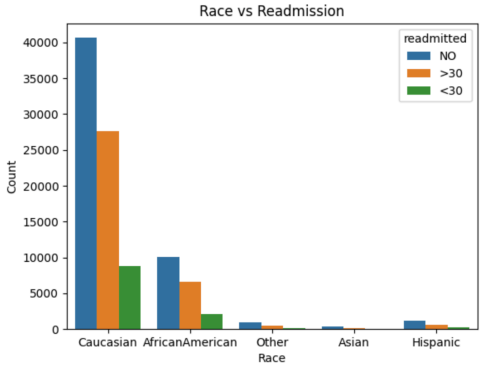


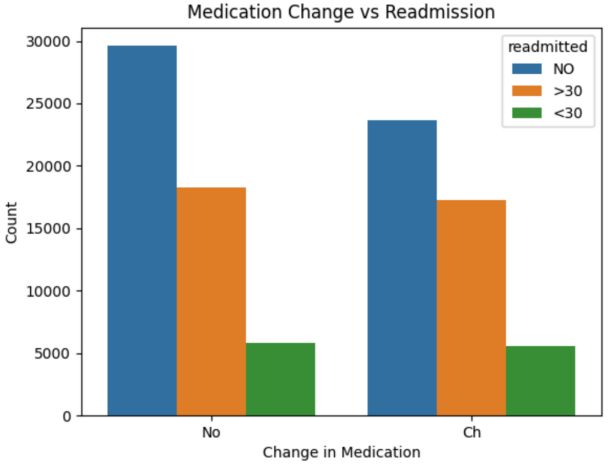
Visualising the data

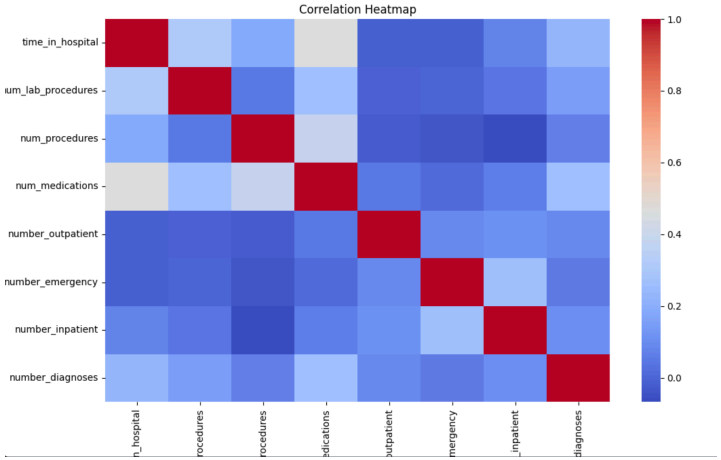
EDA Insights from the project











PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

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:

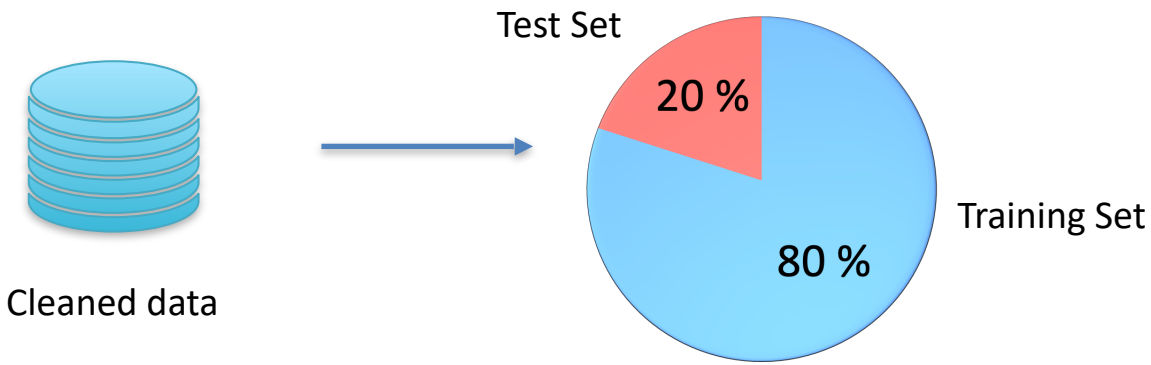
- Read the cleaned dataset.

Define Features & Target:

- X → all relevant features (demographics, labs, meds).
- y → 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.

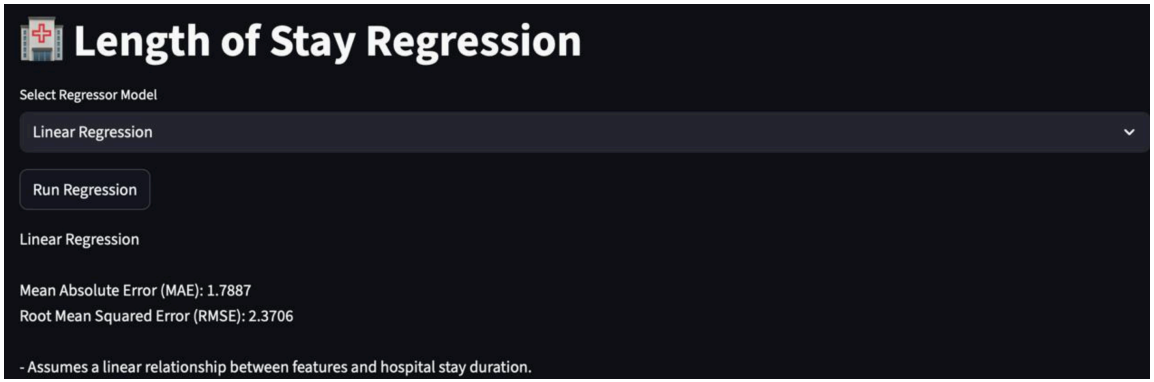
Prediction of Readmission and Length of stay

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

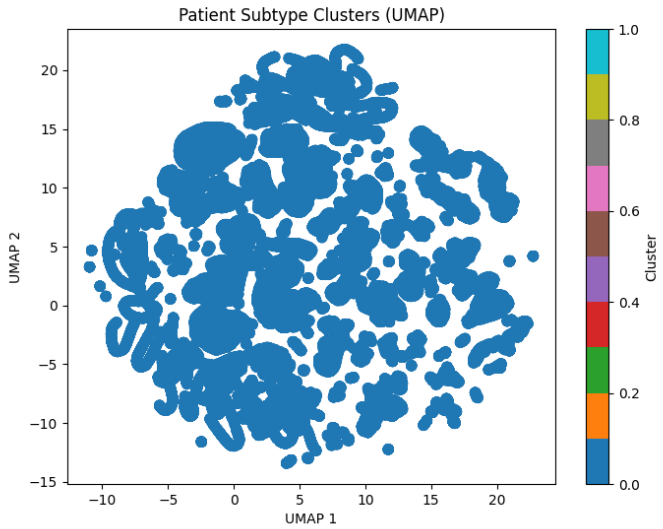
Why?

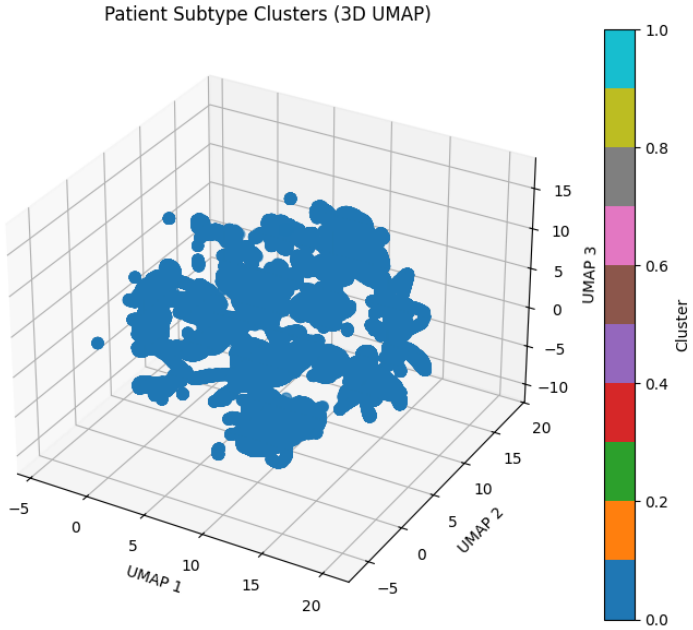
- Helps hospitals plan beds & staff.
- Supports better patient care.
- Can reduce unnecessary costs.

Prediction of Readmission and Length of stay



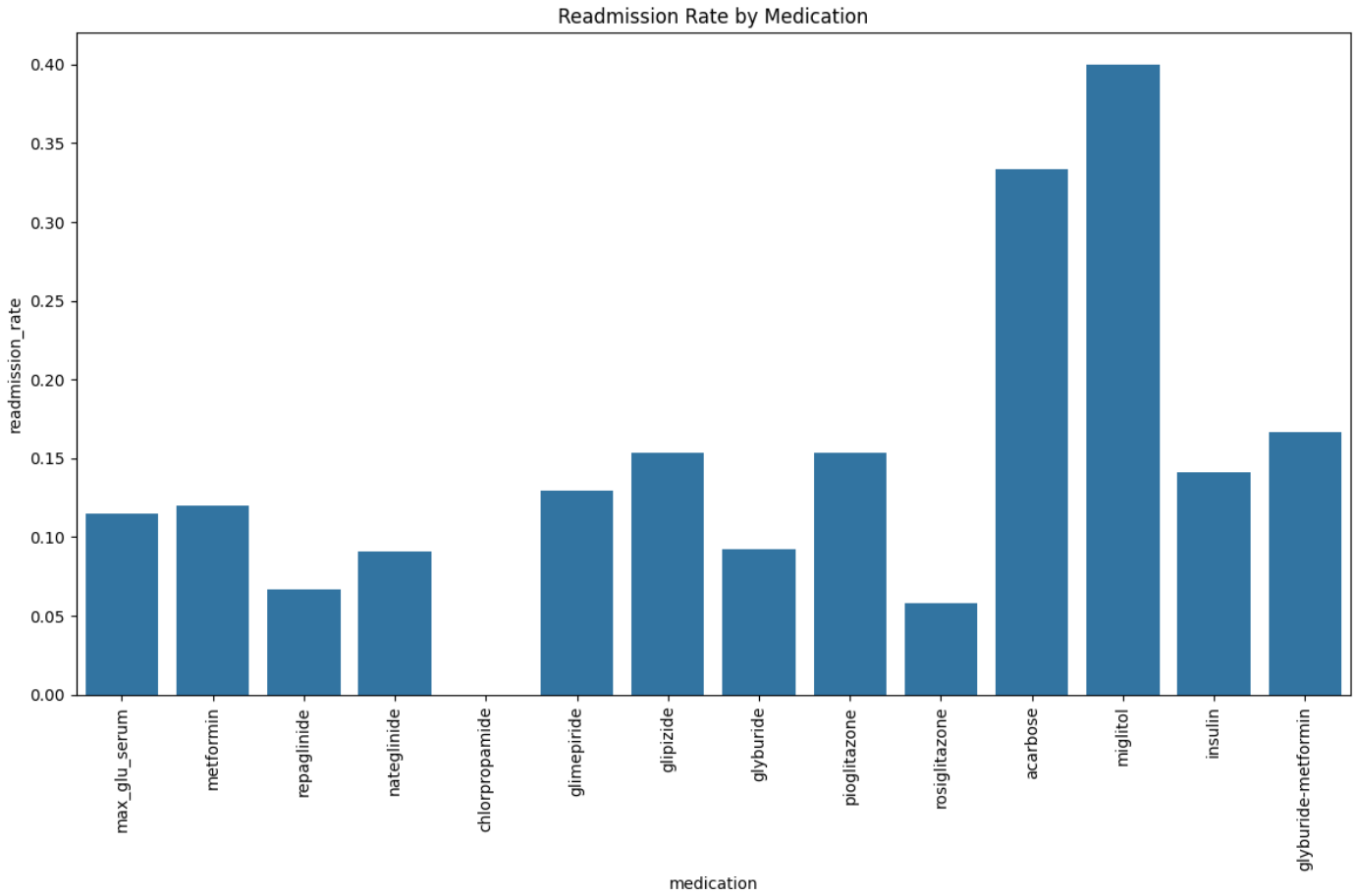
Insights - Patient Grouping



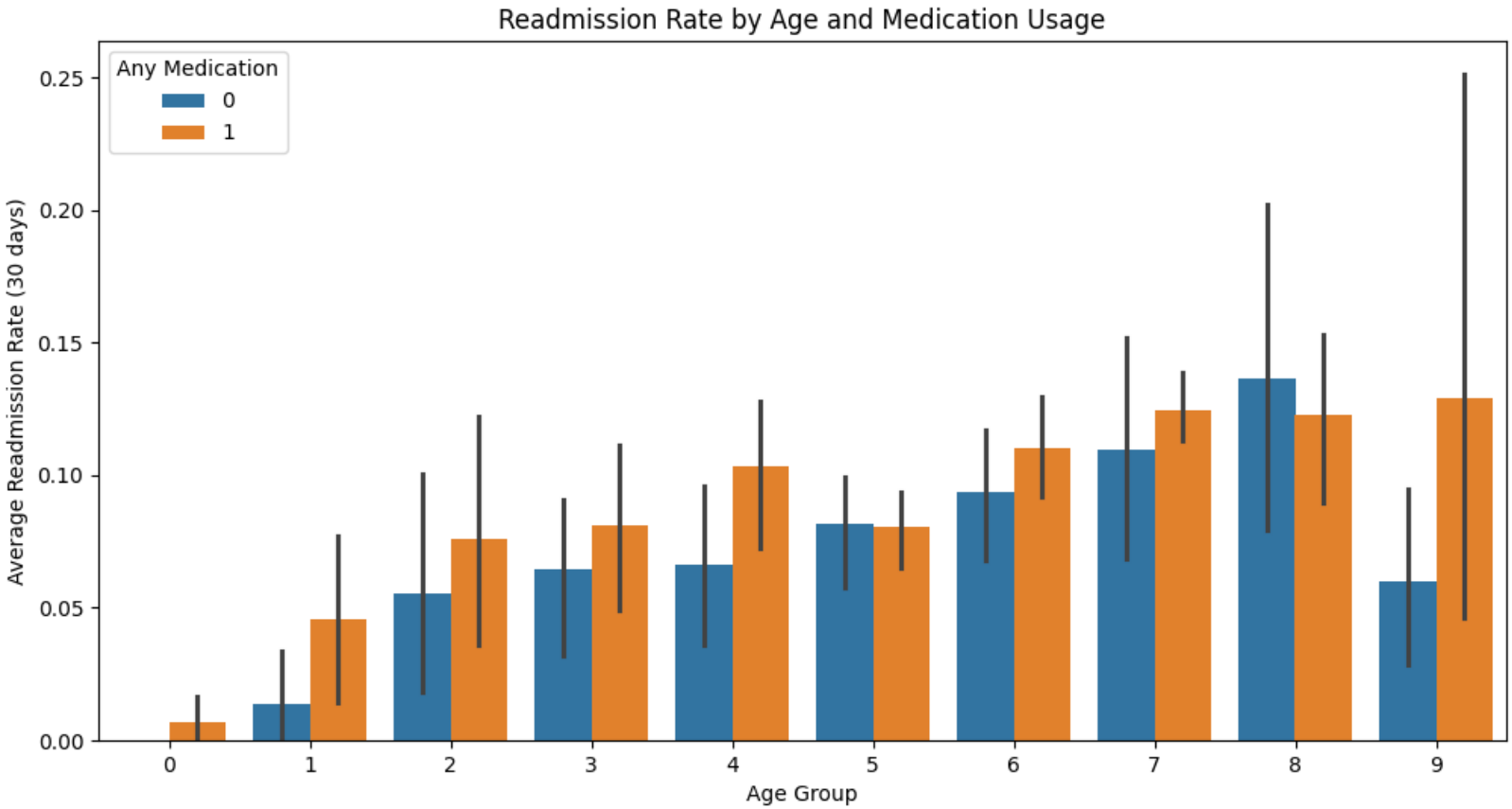


PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

Medication Impact

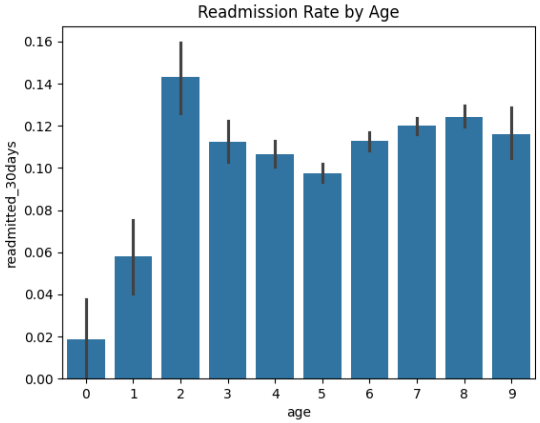


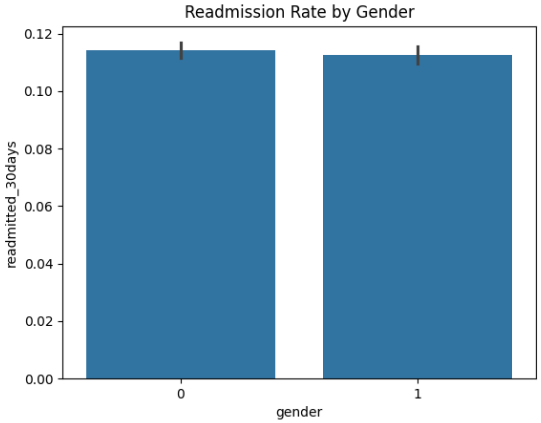
Analyse Risk factors

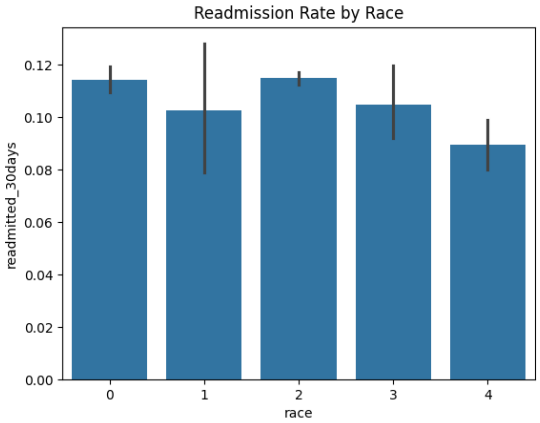


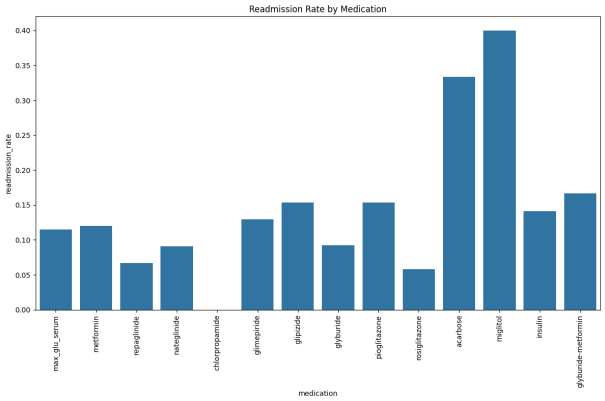
PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

Demographic & Treatment Differences

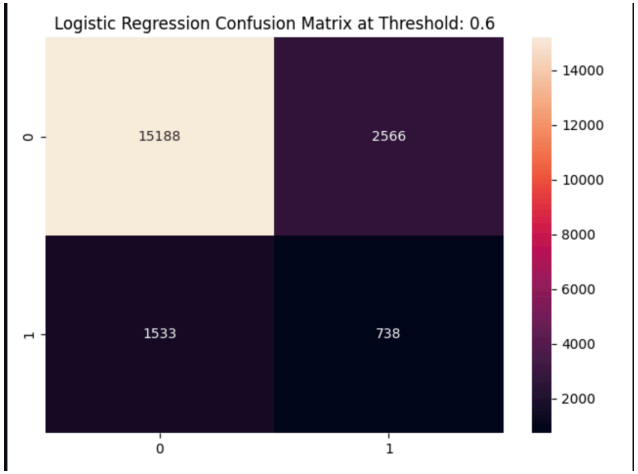


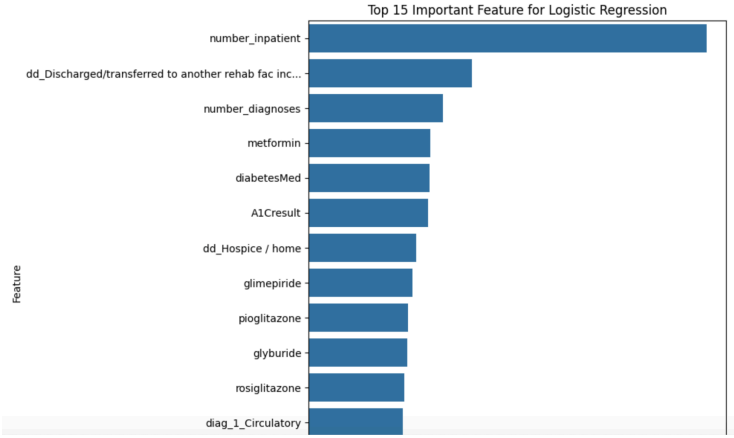


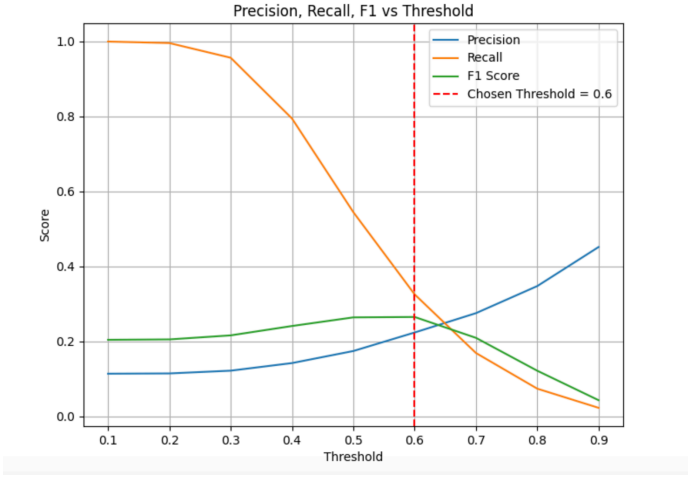




Model Training - Logistic Regression

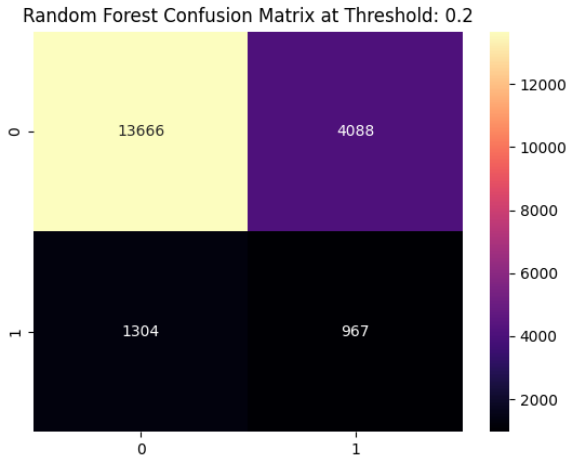


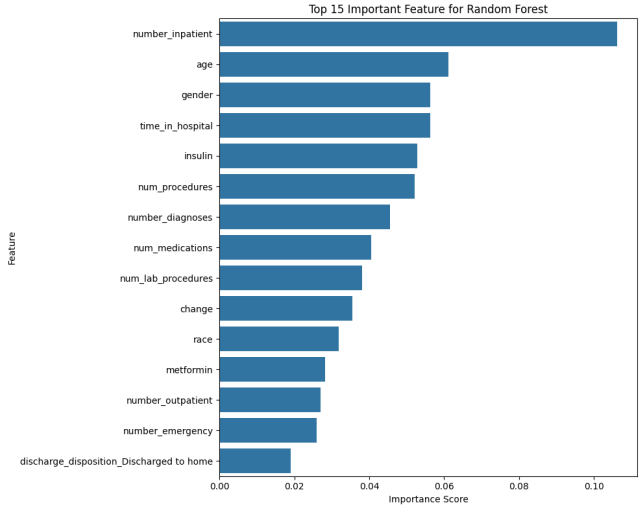


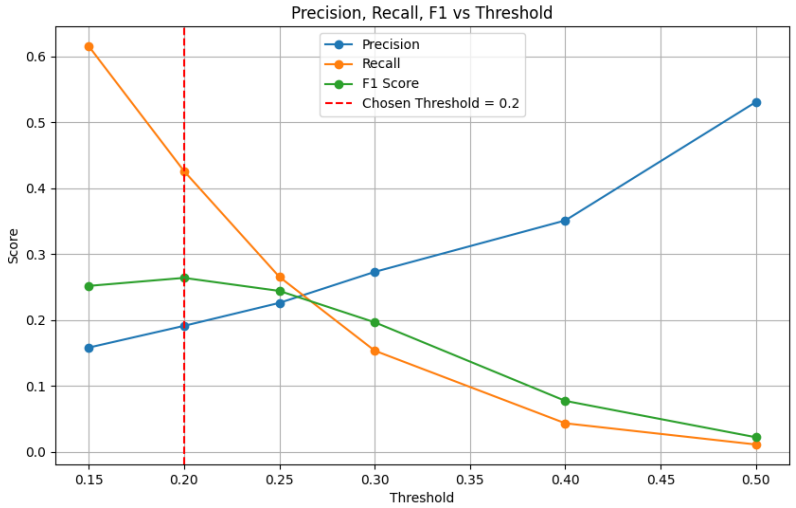


PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

Model Training -Random Forest

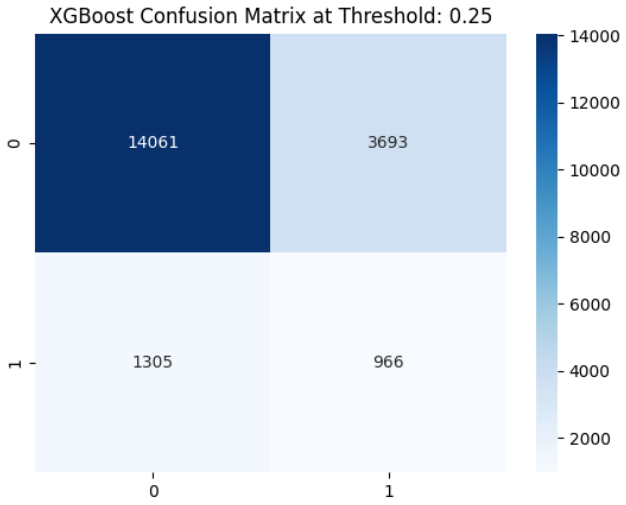


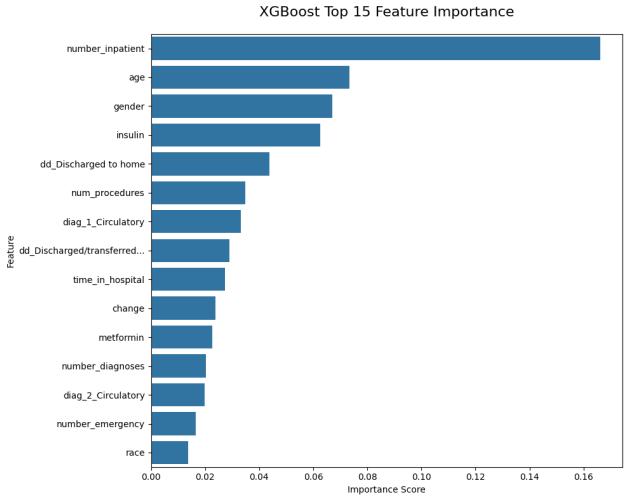


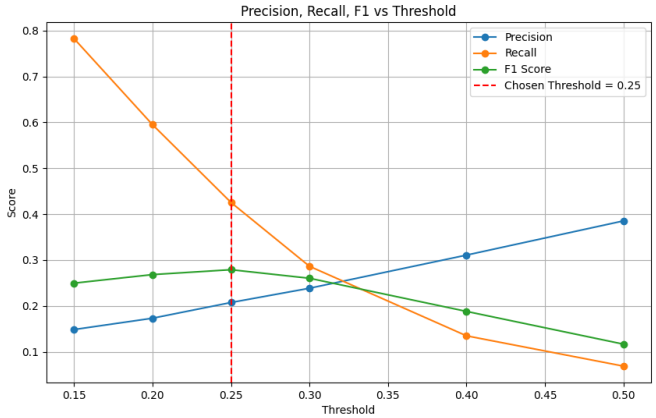


PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

Model Training - XGBoost

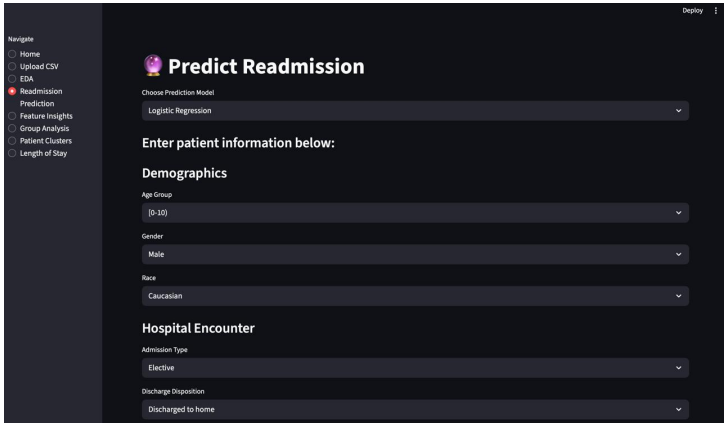


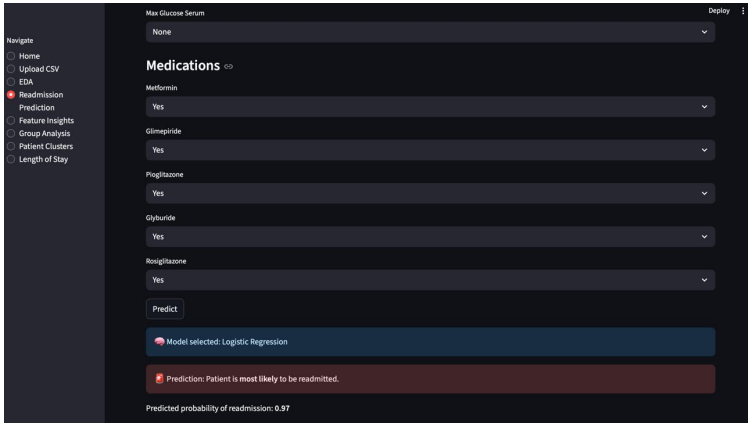




PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

GUI Snapshots





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