Diabetes Readmission Prediction: A Data-Driven Approach

Dhruvilkumar Jitendrakumar Suthar
(Information and Information Technology)

Project Background

Context & Objectives:

- Analyze a dataset of 100,000+ diabetic patient records to understand factors influencing hospital readmission rates.
- Predict readmission status (encoded as 0 = NO, 1 = readmission within <30 days, 2 = readmission after >30 days).

Duration : 8 weeks

Scope & Deliverables:

- **Data Preprocessing:** Loaded and cleaned the dataset; dropped non-relevant IDs; encoded categorical variables using one-hot encoding; filled missing values with median imputation.
- **Exploratory Data Analysis (EDA):** Visualized demographic distributions (e.g., 55% Caucasian, 20% African-American, 25% other), numerical features, and correlations (e.g., r≈0.6 between time_in_hospital and num_lab_procedures).
- **Machine Learning:** Evaluated four models (Logistic Regression, Gaussian Naive Bayes, Decision Tree, Random Forest) and performed hyperparameter tuning on Logistic Regression using GridSearchCV.

My Role

Research | Data Cleaning | Model Training | Exploratory Data Analysis

- Processed 10,000+ patient records; tested 4 ML models (Logistic Regression, Random Forest, etc.).

Skills

Technical

- Python,
- Pandas,
- Scikit-learn
- Matplotlib/Seaborn

Analytical

- Feature engineering,
- model interpretation,
- statistical analysis.

Outcomes

Key Outcomes:

- **Model Performance:**
  - Random Forest emerged as the top model with an AUC of 0.77 and an accuracy of 78.49%.
- **EDA Insights:**
  - Demographic analysis revealed that 55% of patients were Caucasian, 20% African-American, and 52% were female.
  - Right-skewed distributions reveal that most patients have moderate values, but a subset with much higher values may be at increased risk for readmission—highlighting the need for targeted interventions.
- **Medication Analysis:**
  - Patients on steady/increased metformin doses show lower readmission rates (73–77%) versus those not on metformin (82%), suggesting a protective effect of metformin.

Lessons Learned

Week 1

Data Preprocessing & Exploration

Cleaned 100,000+ records, handled missing values, and applied one-hot encoding for categorical features.

Identified right-skewed distributions, highlighting complex cases requiring special attention.

Week 2 - 3

Feature Engineering & Selection

Extracted key medication features, focusing on 12+ drug types affecting readmission rates.

Found that metformin use reduced readmission by ~5-10% compared to non-users.

Week 4 - 5

Model Training & Evaluation

Built ML Models. Achieved best F1-score of ~0.75, balancing precision and recall effectively.

Week 6 - 7

Visualization & Interpretability

Used Matplotlib & Seaborn to analyze medication impact on readmission rates.

Discovered that patients with insulin dosage increases had ~90% readmission probability—signaling a high-risk group.

Week 8

Insights & Real-world Applications

Findings suggest targeted interventions for high-risk diabetic patients can reduce hospital readmissions.

This methodology is scalable for predictive analytics in other healthcare scenarios, improving operational efficiency.

Relevance to Focus Area

- Directly applies predictive modeling and data storytelling skills to improve patient outcomes.
- Experience with EHR data prepares for real-world healthcare datasets.

Future Impact:

Analytical Skills Development → **This project has honed my analytical skills in healthcare data, preparing me to contribute effectively to healthcare analytics initiatives.**

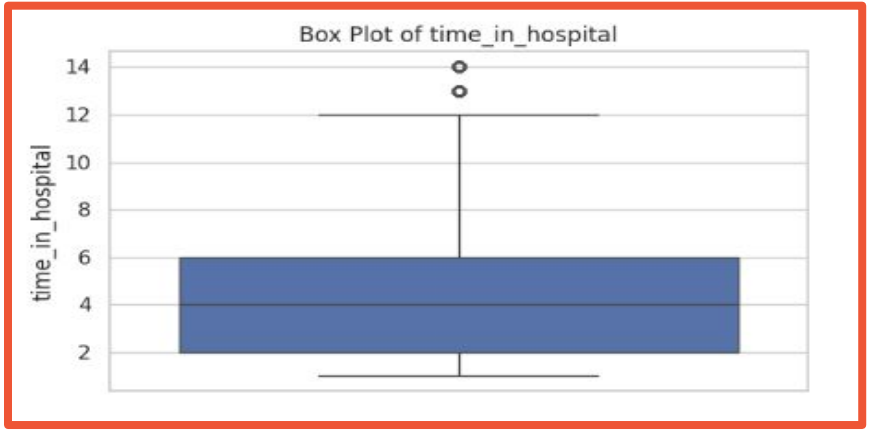
Predictive Modeling Expertise → **Gaining insights into patient readmission patterns positions me well to develop predictive models that can enhance patient outcomes in my future role.**

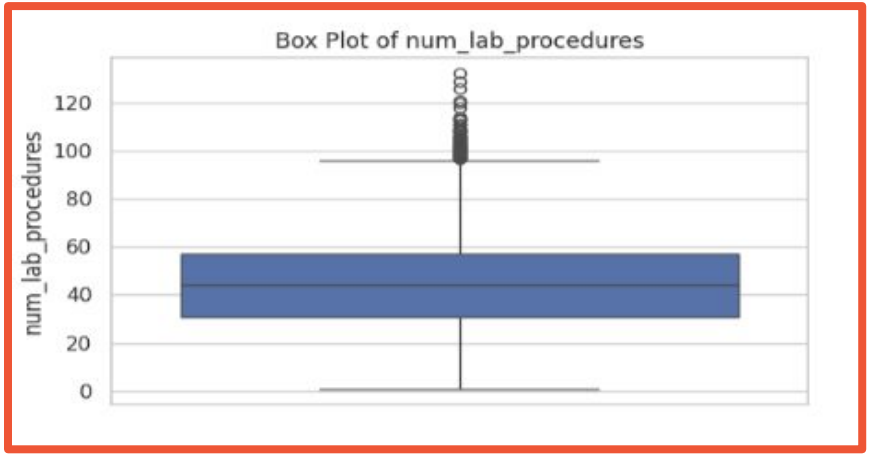
Recommendations

Target High-Risk Patients	Focus on patients with prolonged hospital stays and high medication usage, as they show elevated readmission risks (~80-90%).
Optimize Metformin Usage	Encourage consistent or increased metformin prescriptions, as it is linked to 5-10% lower readmission rates in diabetic patients.
Monitor Insulin Adjustments	Patients with insulin dosage increases have ~90% readmission probability—requiring closer follow-ups and intervention strategies.
Enhance Predictive Analytics	Implement machine learning models to predict readmission risks in real time, enabling proactive care strategies.
Improve Data-Driven Decision Making	Leverage insights from 100,000+ patient records to refine hospital workflows and reduce avoidable readmissions.

Data Exploration and Visualization

- **Distribution of Key Variables:**
  - Time in hospital
  - Number of lab procedures
  - Number of medications
  - Number of diagnoses
- **Observations:**
  - Skewed distributions with outliers
  - High variability in the number of procedures and medications

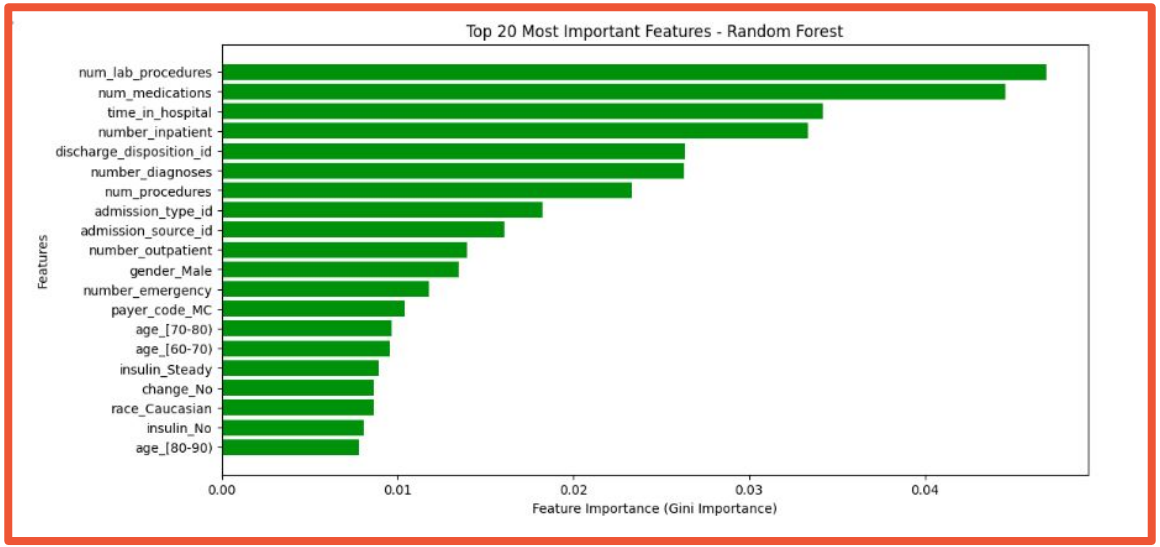




Feature Importance Analysis

Top Features Identified:

- num_lab_procedures
- num_medications
- time_in_hospital
- number_inpatient
- discharge_disposition_id



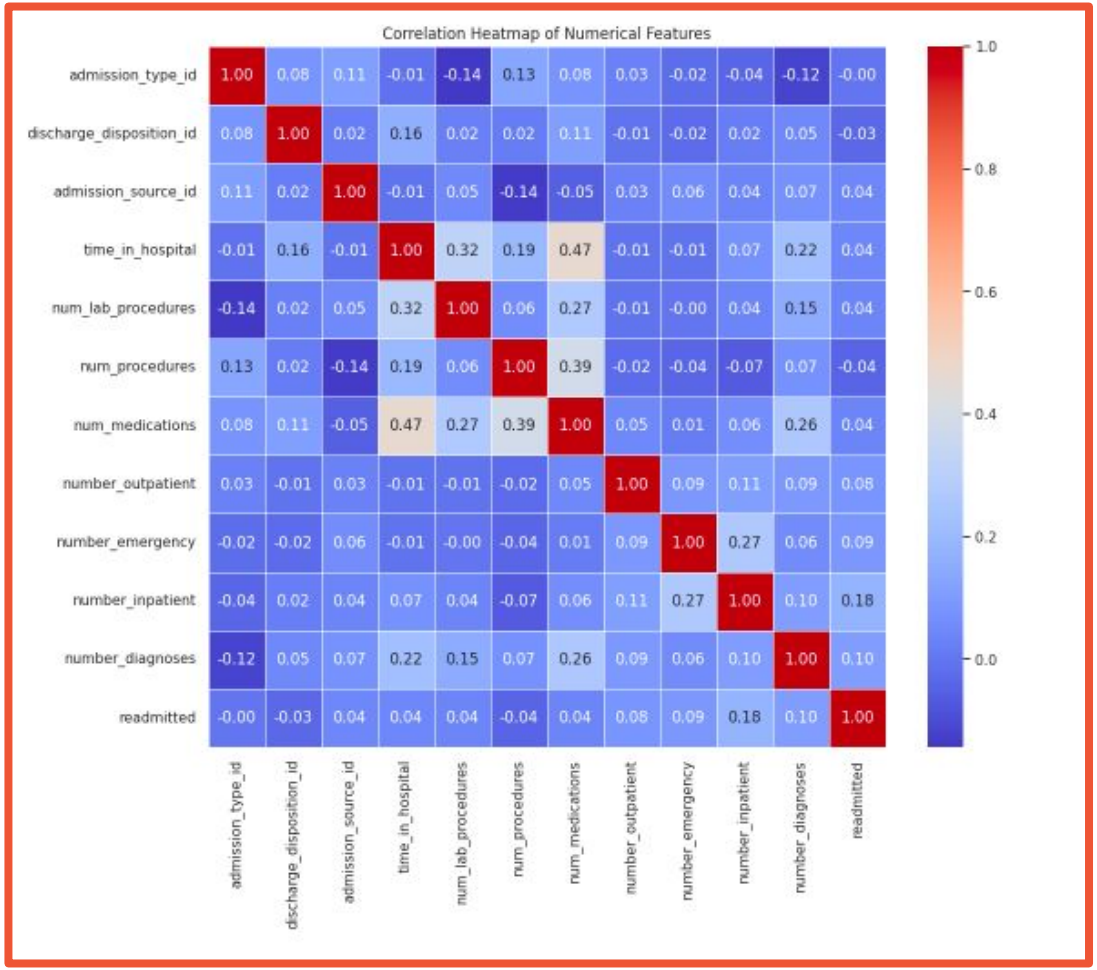
Model Performance Comparison

- **Performance Metrics:**
  - Accuracy, AUC, Recall, Precision, F1 Score
- **Comparison Table:** Summary of performance metrics for each model
- **Insights:**
  - Random Forest outperformed other models
  - Logistic Regression showed moderate effectiveness

Model	Accuracy	AUC	Recall	Precision	F1	Training Time
Logistic Regression	0.57	0.64	0.40	0.45	0.39	50.71 s
Naive Bayes	0.13	0.50	0.34	0.38	0.098	3.76 s
Decision Tree	0.49	0.55	0.39	0.39	0.39	27.75 s
Random Forest	0.78	0.77	0.60	0.78	0.72	148.26 s

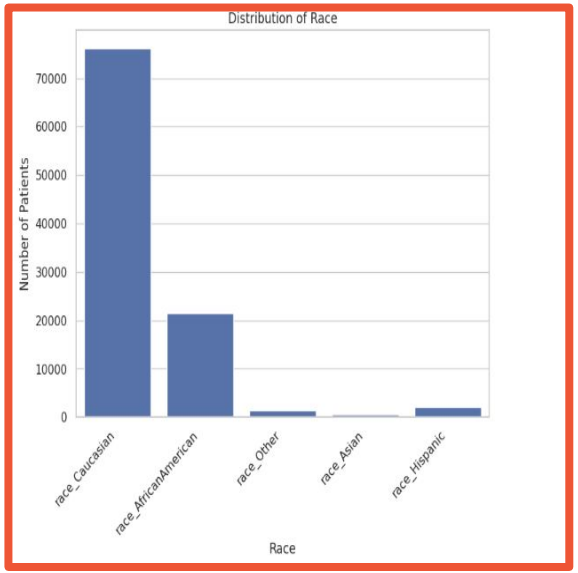
Correlation Matrix Analysis

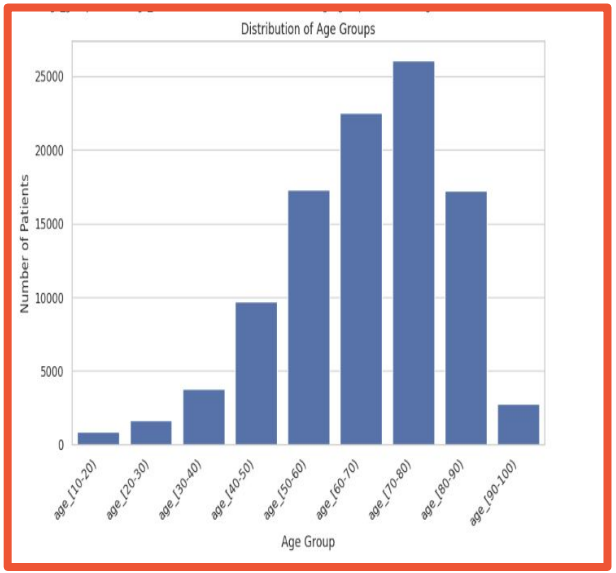
- **Objective:** Analyze correlations between numerical features to refine predictive models.
- **Key Findings:**
  - **Time in Hospital:** Moderately correlated with the number of procedures (0.47).
  - **Medications and Lab Procedures:** Positive correlation (0.27), indicating more tests with more medications.
  - **Readmission Status:** Mildly related to the number of diagnoses (0.10).
- **Implications:**
  - Correlated features may predict readmission risk.
  - Important for feature selection in model development.

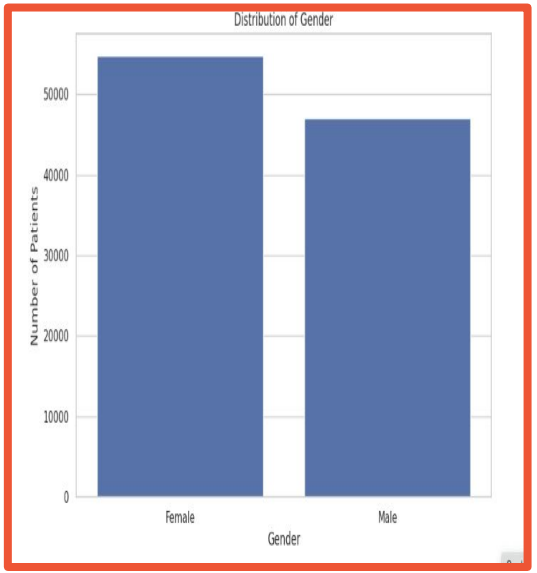


Distribution Analysis of Demographics

- **Race Distribution:** Caucasian patients dominant
- **Age Group Distribution:** Highest in 70-80 age group
- **Gender Distribution:** Female patients account for approximately 55% of the total patient population.







Thank You