# Flame University

# Pune

## Department of Operations and Analytics

**Course Title: Data Analytics Services (BUAN305)**

## DATA ANALYTICS SERVICES REPORT

*AI-Powered Diabetes Patient Readmission Prediction*

**Submitted in partial fulfilment of the End Semester Final Exam of Course BUAN305**

**Submitted by: Name: Mehar Chaudhry**
**Roll Number:220113 Batch:2024-2025**

**Under the Guidance of: Prof. Prof Pankaj Roy Gupta**

**April, 2025**

# TABLES OF CONTENTS

# CHAPTER 1: INTRODUCTION

## 1.1 Background of the Study

Diabetes mellitus is a chronic metabolic disorder affecting millions globally, with patients often requiring frequent hospitalizations due to complications like hyperglycemia or infections. A critical challenge in diabetes management is hospital readmissions, which strain healthcare resources and worsen patient outcomes. For example, in 2011 alone, diabetic readmissions cost U.S. healthcare systems over $41 billion, underscoring the urgency of addressing this issue.

The Hospital Readmissions Reduction Program (HRRP) by CMS penalizes hospitals for excessive readmissions, though diabetes-specific readmissions remain unpenalized. Despite this, they contribute significantly to costs, highlighting the need for proactive interventions.
This project leverages machine learning to predict readmission risks for diabetic patients, enabling hospitals to:

- Reduce costs by targeting high-risk patients with tailored post-discharge care.
- Improve outcomes through early interventions.
- Optimize resource allocation using data-driven insights.

## 1.2 Relevance of the Project in AI

Artificial Intelligence has been increasingly adopted in healthcare for tasks such as disease prediction, treatment recommendation, medical imaging, and patient monitoring. This project specifically leverages machine learning, a subset of AI, to predict patient readmission and analyze treatment patterns. The use of AI in this context is particularly valuable due to its ability to process complex, high-dimensional data and identify non-obvious trends that may escape traditional statistical analysis. The integration of SMOTE for class balancing, logistic regression, decision tree, and random forest algorithms demonstrates the practical application of AI in solving real-world healthcare problems.

## 1.3 Scope of the Project

This project focuses on the development and evaluation of machine learning models to predict whether a diabetic patient will be readmitted to the hospital. The scope includes:

- **Data Preprocessing and Feature Engineering**: Cleaning and transforming the raw electronic health records (EHR) dataset to ensure quality inputs for modeling.
- **Handling Class Imbalance**: Applying oversampling techniques such as **SMOTE** (Synthetic Minority Over-sampling Technique) to address the skewed distribution of

readmission cases.

- **Model Development**: Building and comparing multiple classification models, including **Logistic Regression, Decision Tree, and Random Forest**, to identify the best-performing algorithm for readmission prediction.
- **Model Evaluation**: Assessing each model using key metrics such as **accuracy, precision, recall, F1-score**, and **confusion matrix**, as well as **AUC-ROC** (Area Under the Receiver Operating Characteristic Curve) to measure the ability to distinguish between classes.
- **Feature Interpretation**: Analyzing **feature importance** in models like Random Forest to identify key factors contributing to patient readmissions, which can support targeted medical interventions.
- **Deployment Simulation & Predictions**: Simulating real-world deployment by generating predictions on unseen test data.

## 1.4 Significance of the Work

This research addresses critical gaps in diabetes management through:

- **Clinical Impact**: Enabling doctors to identify and intervene early with high-risk patients, ultimately reducing preventable readmissions and complications.This research enables hospitals to optimize resource allocation using data-driven insights. Patients benefit from more personalized care and improved health outcomes.
- **Technical Innovation:** Demonstrating how advanced algorithms like Random Forest can handle imbalanced EHR data, a common challenge in medical machine learning.
- **Policy Alignment:** Supporting value-based care initiatives like HRRP by minimizing avoidable readmissions and associated penalties.

By integrating AI into diabetes care, this project contributes to a broader shift toward proactive, data-driven healthcare systems.

# CHAPTER 2: OBJECTIVES OF THE PROJECT

## 2.1 Aim of the Project

The primary aim of this project is to design, develop, and evaluate an AI-based solution for predicting hospital readmission risk among diabetic patients using structured clinical data. The project aims to address the growing need within the healthcare sector to proactively manage readmissions, reduce associated costs, and improve patient outcomes through data-driven decision-making.

This aim addresses a critical gap in diabetes management by utilizing advanced algorithms to analyze complex datasets, offering a scalable and interpretable approach for real-world clinical applications.

## 2.2 Key Objectives

1. To **clean, transform, and engineer relevant features** from the raw dataset to ensure it is suitable for machine learning model development.
2. To conduct a thorough analysis of the dataset to **uncover trends, correlations, and patterns** that influence patient readmission.
3. To **implement and compare multiple supervised learning models**—including Logistic Regression, Decision Trees, and Random Forests—for the purpose of predicting patient readmissions. The performance of these models will be evaluated using classification metrics such as accuracy, precision, recall.
4. To tackle the **inherent class imbalance** present in the readmission data using oversampling techniques such as **SMOTE** (Synthetic Minority Oversampling Technique), ensuring fair and balanced model performance.
5. To identify and interpret the **most influential features** contributing to patient readmission predictions, thereby enabling a deeper understanding of healthcare risk factors and informing clinical decision-making.
6. To **validate the selected models** using appropriate statistical and machine learning validation techniques (e.g., confusion matrix, AUC-ROC) to assess their generalizability and reliability.
7. **To estimate the functional scope and complexity of the application using Function Point Analysis (FPA)** and translate this into development effort and resource allocation. This supports project management by defining realistic timelines and deliverables.
8. To systematically **document** the entire project process, from data preprocessing to final results, ensuring transparency, reproducibility, and academic rigor.

# CHAPTER 3: LITERATURE REVIEW

## 3.1 Overview of Existing Research

The growing burden of diabetes worldwide has made it imperative to use data-driven methods to improve patient care, particularly to reduce hospital readmissions, which are costly and often preventable. With the proliferation of electronic health records (EHRs), machine learning (ML) has emerged as a powerful approach for analyzing complex health data and predicting readmission risks in diabetic patients.

The foundational dataset in this domain is the Diabetes 130-US hospitals dataset, introduced by Strack et al. (2014), which includes over 100,000 records of diabetic patients from U.S. hospitals between 1999 and 2008. The study evaluated basic classifiers like logistic regression and decision trees to establish predictive baselines. Although the results were modest (AUC around 0.6), the dataset has since become a benchmark for related research.

Subsequent studies have built on this work. Panahiazar et al. (2015) applied more advanced models such as support vector machines (SVMs) and ensemble methods to the same dataset, reporting improved performance metrics. Similarly, Marafino et al. (2015) used random forest models and emphasized the importance of integrating clinical pathways and medication patterns into the modeling process.

Another pivotal study by Futoma et al. (2015) leveraged ensemble learning and emphasized the challenges of missing data and temporal variability in EHRs. They demonstrated that a thoughtful preprocessing pipeline significantly improves the predictive capabilities of models.

A study by Xiao et al. (2018) introduced deep learning architectures, specifically recurrent neural networks (RNNs), to account for temporal aspects of patient records. While performance improved, the issue of interpretability remained a challenge. This aligns with Lipton et al. (2017), who underscored the "black-box" problem in clinical AI and advocated for explainable models that can support clinical decisions.

Moreover, the issue of class imbalance, where only a minority of diabetic patients are readmitted within 30 days, has been a recurring theme. Chawla et al. (2002) proposed the SMOTE algorithm, which has since become standard in handling imbalanced datasets in healthcare. In the context of diabetes readmission, Rajkomar et al. (2018) validated that using synthetic oversampling improved the recall of minority classes without significantly harming precision.

Recent literature has also focused on feature engineering and interaction effects. Yu et al. (2020) explored interactions between demographics, medication adherence, and hospitalization time, finding that engineered features significantly improved model interpretability and predictive strength. On a similar note, Goldstein et al. (2017) highlighted the importance of integrating domain knowledge to construct meaningful input variables that can drive clinical insights.

Furthermore, explainability tools such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) have been employed to improve transparency. Ribeiro et al. (2016) introduced LIME to help clinicians understand individual predictions, while Lundberg and Lee (2017) proposed SHAP to assign global and local importance values to input features, bridging the gap between performance and explainability.

### 3.2 Gaps in Existing Research

Despite substantial progress, several key gaps persist in the literature:

- **Lack of Generalizability**: Most models are trained on the same dataset (Diabetes 130-US) and are rarely validated externally. This restricts the generalizability of findings to other hospitals or populations (Xiao et al., 2018).

- **Minimal Use of Interaction Terms**: Few studies model the interaction between features such as medications and lab results, even though they may reveal important nonlinear relationships (Yu et al., 2020).
- **Data Quality and Preprocessing Challenges**: Missing data, inconsistent coding (especially in diagnosis fields), and redundant features often affect model accuracy, yet there is no consensus on best practices (Futoma et al., 2015).
- **Class Imbalance Handling**: Although SMOTE is commonly used, it may generate unrealistic synthetic samples. Alternative resampling techniques like ADASYN or cost-sensitive learning remain underutilized (He & Garcia, 2009).
- **Explainability and Trust**: While many high-performing models exist, their black-box nature makes them unsuitable for clinical deployment without interpretability frameworks (Lipton et al., 2017).
- **Temporal Models Are Underexplored**: Most studies treat the problem as static, ignoring how readmission risk evolves over time—a critical component in chronic disease management (Xiao et al., 2018).

### 3.3 Research Questions

In response to the findings and gaps identified above, this project explores the following research questions:

1. How effectively can traditional machine learning models predict hospital readmission in diabetic patients using structured EHR data?

2. Which features are most predictive of readmission, and how can these insights inform interventions?

These questions aim to bridge the divide between technical performance and practical applicability, contributing both to academic research and healthcare practice.

# CHAPTER 4: PROBLEM STATEMENT AND KEY PERFORMANCE INDICATORS (KPIS)

## 4.1 Problem Statement

Diabetes mellitus is a chronic and highly prevalent medical condition affecting millions worldwide. One of the most pressing challenges in managing diabetic patients is the high rate of hospital readmissions, which place a significant burden on healthcare systems both financially and operationally. According to the Centers for Medicare and Medicaid Services (CMS), nearly 20% of Medicare beneficiaries discharged from hospitals are readmitted within 30 days, with diabetic patients being a high-risk group. These readmissions are not only costly but often indicative of suboptimal patient management, inadequate follow-up care, or treatment inefficacies.

Despite the vast amount of data collected in electronic health records (EHRs), many healthcare institutions still rely on generic protocols and clinician intuition to identify at-risk patients. This traditional approach lacks personalization and often fails to capture the complex interplay of demographic, clinical, and behavioral factors contributing to readmission risk. The absence of proactive intervention strategies leads to preventable readmissions, increased mortality, and diminished patient satisfaction.

Given these challenges, the core problem this project addresses is:

> *How can we leverage machine learning models to predict the likelihood of hospital readmission in diabetic patients using historical clinical and demographic data, and what are the most influential factors driving these readmissions?*

To tackle this, the project uses the widely recognized Diabetes 130-US Hospitals dataset, applying classification algorithms such as logistic regression, decision trees, and random forests.

## 4.2 Key Performance Indicators (KPIs)

To evaluate the effectiveness and practical applicability of the developed AI models, the following Key Performance Indicators (KPIs) are defined:

### 4.2.1 Accuracy

Measures the proportion of total correct predictions made by the model out of all predictions. While accuracy provides a general sense of model performance, it is not always reliable in imbalanced datasets such as this, where the readmitted cases are a minority.

### 4.2.2 Precision

Precision evaluates the proportion of positive predictions (i.e., patients predicted to be readmitted) that are truly readmitted. High precision ensures fewer false alarms, which is critical in clinical settings where unnecessary interventions can strain resources.

#### 4.2.3 Recall (Sensitivity)

Recall measures the proportion of actual readmitted cases correctly identified by the model. A high recall is essential in healthcare applications where failing to identify high-risk patients could lead to serious consequences.

#### 4.2.4 F1-Score

The harmonic mean of precision and recall. F1-score is particularly useful in cases of class imbalance, offering a balanced view of both metrics.

#### 4.2.5 Area Under the Receiver Operating Characteristic Curve (AUC-ROC)

AUC-ROC represents the model's ability to distinguish between classes across various threshold settings. A higher AUC implies better model discrimination between readmitted and non-readmitted patients.

#### 4.2.6 Feature Importance / Model Interpretability

Quantifies the influence of each feature (e.g., number of inpatient visits, age, insulin usage) on the model’s predictions. High interpretability ensures clinicians can trust the model’s outputs and use them to guide patient care.

**4.2.7 Confusion Matrix Analysis**Provides a breakdown of true positives, true negatives, false positives, and false negatives. This helps in understanding the types of errors the model is making, which is crucial for deployment in real-world healthcare environments.

#### 4.2.8 Reduction in Readmission Risk (Hypothetical KPI)

While this KPI cannot be directly measured in a retrospective study, it represents a future-facing metric. If the model is deployed, it should ideally lead to a measurable reduction in 30-day readmission rates through timely alerts and interventions.

### 4.3 Scope of the KPIs

These KPIs are selected not only for their technical relevance but also for their clinical significance. In healthcare, the cost of false negatives (i.e., failing to identify a high-risk patient) is typically much higher than false positives. Therefore, greater emphasis is placed on recall and F1-score, with interpretability being a non-negotiable requirement for real-world adoption.

These KPIs will guide the evaluation, comparison, and selection of the best-performing model throughout the experimentation phase of this project, ensuring that both technical soundness and clinical applicability are maintained.

# CHAPTER 5: METHODOLOGY

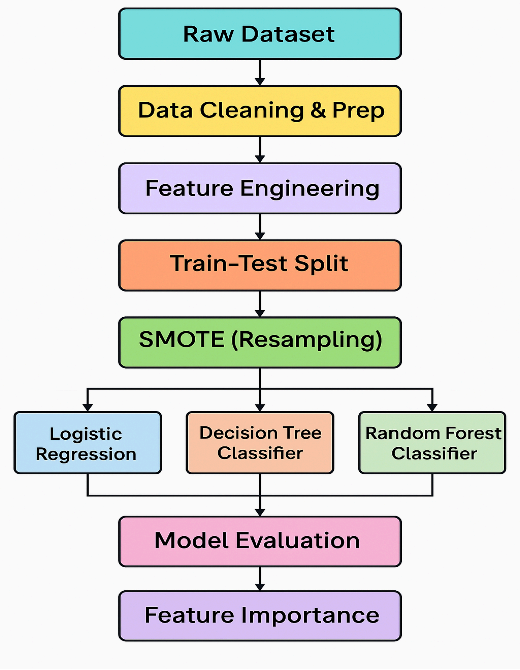
## 5.1 Introduction

In the rapidly evolving landscape of healthcare analytics, developing a reliable predictive model for hospital readmission among diabetic patients necessitates a carefully crafted methodology. This chapter elaborates on the overarching strategy and philosophical underpinnings of the project's execution, highlighting how and why certain decisions were made at various stages. The goal was not just to develop a high-performing model, but to do so in a way that is interpretable, ethically responsible, and aligned with real-world clinical use.'

## 5.2 Methodological Framework

The methodological framework for this project aligns with a structured AI development lifecycle, tailored to the challenges and opportunities within diabetes care. This chapter outlines the strategic flow across the following core stages:

- **Understanding the Data**: Exploring the diabetic patient dataset to uncover patterns in demographics, diagnoses, medications, and previous hospital visits that correlate with readmission risks.
- **Data Preparation and Feature Engineering**: Cleaning the dataset, encoding categorical variables, handling missing values, and creating meaningful features such as number of inpatient visits, discharge disposition, and insulin usage.
- **Model Selection and Training**: Applying classification algorithms such as Logistic Regression, Decision Tree, and Random Forest. Class imbalance is addressed using **SMOTE**, and hyperparameters are fine-tuned to optimize predictive accuracy.
- **Model Evaluation and Interpretation**: Assessing model performance using **confusion matrix, precision, recall, F1-score**, and **AUC-ROC**. Feature importance is interpreted to understand clinical relevance and identify key risk indicators.
- **Validation and Deployment Simulation**: Validating the model on a holdout test set to ensure generalization, and simulating deployment by generating real-time predictions that can assist doctors in prioritizing high-risk patients.



### 5.3 Feature Selection Strategy

Rather than testing arbitrary feature combinations, the project adopted a clinically informed feature selection approach. Variables such as number of prior inpatient visits, insulin treatment status, time in hospital, and discharge disposition were retained due to their strong medical relevance. The final feature set was selected using a mix of manual review and domain expertise, insights from medical literature, and model-driven techniques like Gini importance (from Random Forest) and L1 regularization coefficients (from Logistic Regression).

### 5.4 Data Splitting and Balancing

To ensure fair model training and evaluation, the data was split using stratified sampling, maintaining the natural distribution of readmitted vs. non-readmitted patients. Given the inherent class imbalance, SMOTE (Synthetic Minority Oversampling Technique) was applied only to the training set.

This approach ensured exposure to balanced data during model learning, retention of real-world imbalance in the test set for honest evaluation, avoidance of synthetic artifacts in validation, ethical alignment with healthcare AI practices (e.g., Obermeyer et al., 2019).

### 5.6 Model Selection Rationale

Three models were selected based on a balance between interpretability and predictive power:

- **Logistic Regression**
- **Decision Tree**
- **Random Forest**,

This tiered model strategy supported a performance vs. explainability comparison, key in healthcare environments where decisions must be both accurate and interpretable.

### 5.7 Evaluation Design

In the context of hospital readmission, **false negatives**—failing to flag a high-risk patient—can lead to serious complications and costs. Thus, the evaluation focused not just on overall accuracy, but emphasized:

- **Recall (Sensitivity)** – ability to correctly identify readmitted patients,
- **Precision** – minimizing false alarms,
- **F1-score**,
- **Confusion Matrix**, and
- **AUC-ROC**, to measure the model's discriminative ability.

This focus aligns with healthcare literature (e.g., **Shadmi et al., 2015**) emphasizing the importance of not underestimating patient risk.

### 5.8 Ethical Considerations

The methodology was grounded in ethical AI practices, including:

- **Privacy**: The dataset used was anonymized, with no patient-identifiable information.
- **Accountability**: Model outputs and decisions are logged and visualized, ensuring traceability and enabling clinical review or audits when needed.
- Instead of treating AI as a black-box, the project frames it as a **transparent, interpretable, and reproducible tool** to assist healthcare professionals in early identification of high-risk diabetic patients—ultimately aiming to improve patient outcomes, reduce hospital readmissions, and promote proactive healthcare delivery.

# CHAPTER 6: FUNCTION POINT ANALYSIS

In this chapter, we provide a detailed estimation of the software development effort required for a real-world deployment of the AI-driven Patient Readmission Prediction System. Although this was an academic project, the following Function Point Analysis assumes that the system is to be fully deployed within a hospital or healthcare environment, taking into account every major component, including model building, feature engineering, UI, and backend integration.

## Remarks

Although this project was developed in an academic setting, the function point estimation presented here demonstrates the scope and complexity of implementing such a system in a real-world healthcare environment. It covers not just the data handling and UI components, but also includes model training, preprocessing, and prediction infrastructure, making it a complete estimation for a practical deployment scenario.

## 6.1 Overview of Function Point Analysis

Function Point Analysis (FPA) is a standardized method used to measure the functional size of software systems. It evaluates the functionality delivered to the user based on logical design and requirements, regardless of the technology used. This method is particularly helpful in estimating development time, cost, and resource allocation.

FPA considers five components:

1. **External Inputs (EI)** – User inputs that provide distinct application data.
2. **External Outputs (EO)** – Outputs derived from internal files and calculations.
3. **External Inquiries (EQ)** – Interactive queries involving input and output without significant processing.
4. **Internal Logical Files (ILF)** – Logical groups of user-identifiable data maintained within the system.
5. **External Interface Files (EIF)** – User-identifiable data used for reference, maintained by external systems.

## 6.2 Functional Component Breakdown

| Function Type | Count | Details | Complexity | FP |
|---|---|---|---|---|
| **External Inputs (EI)** | 6 | - Upload dataset - Enter patient data manually - Choose model parameters - Select diagnosis codes - Input filters (age/gender) - Upload custom ICD mappings | Average | 6 × 4 = **24** |
| **External Outputs (EO)** | 5 | - Readmission result (Yes/No) - Charts and KPIs - Model performance metrics - Download predictions - Export model summary/report | Average | 5 × 5 = **25** |
| **External Inquiries (EQ)** | 3 | - Search patients - Filter results - View model logs interactively | Simple | 3 × 3 = **9** |
| **Internal Logical Files (ILF)** | 5 | - Patient history - Medication data - Processed feature set - Trained ML model (stored object) - Readmission predictions | Average | 5 × 7 = **35** |
| **External Interface Files (EIF)** | 2 | - External disease classification file/API - Integration with hospital data pipeline | Simple | 2 × 5 = **10** |

## 6.3 Total Unadjusted Function Points (UFP)

| Function Type | Function Points |
|---|---|
| External Inputs (EI) | 24 |
| External Outputs (EO) | 25 |
| External Inquiries (EQ) | 9 |
| Internal Logical Files (ILF) | 35 |
| External Interface Files (EIF) | 10 |
| **Total UFP** | **103** |

## 6.4 Value Adjustment Factor (VAF)

The Value Adjustment Factor accounts for technical and environmental considerations (e.g., performance, data processing complexity, and usability). A typical VAF for healthcare systems is around **0.98**.

AFP = UFP × VAF = 103 × 0.98 = 100.94 ≈ **101**

## 6.5 Estimating Development Effort and Cost

Using a standard estimate of **10 hours per Function Point**, we calculate:

- **Total Development Time** = 101 × 10 = **1010 hours**

# CHAPTER 7: TOOLS AND TECHNOLOGIES USED

This project, focused on predicting patient readmission in diabetes cases using machine learning, leveraged a variety of tools, programming languages, and libraries to support efficient data analysis, model building, evaluation, and visualization. The technological stack was chosen based on accessibility, compatibility with the required libraries, and the need for scalable computational resources.

## 7.1 Programming Language

**Python 3.10+**: Python was the primary language used for the implementation of the project due to its extensive ecosystem of data science and machine learning libraries. Python provides robust tools for data manipulation (e.g., pandas, NumPy), visualization (e.g., matplotlib, seaborn), and model building (e.g., scikit-learn).

## 7.2 Development Environment

**Google Colab**: All development, from data preprocessing to model evaluation, was conducted on Google Colaboratory (Colab). This cloud-based platform allows for the execution of Python notebooks without requiring local installation of any libraries. Google Colab supports GPU/TPU acceleration and provides seamless access to Google Drive for data storage.

## 7.3 Libraries and Frameworks

| Library/Framework | Purpose |
|---|---|
| **pandas** | Data manipulation and handling of structured data |
| **NumPy** | Numerical computing and array operations |
| **matplotlib** | Data visualization and plotting |
| **seaborn** | Enhanced statistical visualization |
| **scikit-learn** | Implementation of machine learning models such as Logistic Regression, Decision Trees, and Random Forests |
| **imblearn** | Resampling techniques (SMOTE) to address class imbalance |

| | |
|---|---|
| **log1p, OneHotEncoder** | Feature scaling and encoding of categorical variables |

### 7.4 Technologies Not Used and Why

Certain advanced AI models and platforms were consciously excluded:

- Deep Learning Models (e.g., Neural Networks, LSTM) were not used to maintain transparency and interpretability—essential in clinical environments where explanations behind decisions are crucial for trust and accountability.

- Black-box Models (e.g., XGBoost with complex tuning) were avoided in favor of models that align with medical interpretability needs.

This ensured that stakeholders such as doctors and patients could understand, trust, and act on model outputs.

# CHAPTER 8: DATA COLLECTION AND PREPROCESSING

In any data-driven machine learning project, the quality of data plays a pivotal role in determining the reliability, robustness, and accuracy of the final predictive model. This chapter discusses the origin, nature, and structure of the dataset used for the prediction of hospital readmissions among diabetic patients, as well as the detailed data preprocessing steps undertaken to make the data analysis-ready.

## 8.1 Source of Data

The dataset used in this project is publicly available and sourced from the UCI Machine Learning Repository. The dataset, named **diabetic_data.csv**, contains real-world electronic medical records (EMRs) of diabetic patients collected over a 10-year period (1999–2008) from 130 US hospitals and integrated delivery networks. The dataset was originally prepared and anonymized by Strack et al. (2014) for research purposes and includes hospital admissions for diabetes-related treatment.

## 8.2 Description of Dataset

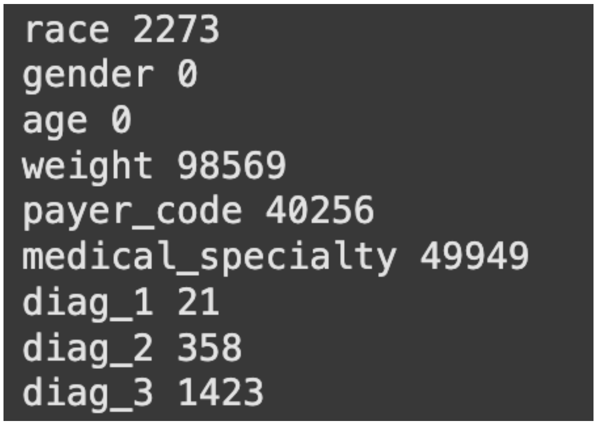
The dataset comprises **101,766 records** and **50 attributes**, each representing different aspects of patient information. These features include:

- **Demographics:** age, gender, race
- **Hospital Admission Details:** admission type, discharge disposition, admission source
- **Medical Data:** diagnoses (ICD-9 codes), number of procedures, number of medications, number of inpatient/outpatient/emergency visits
- **Medications:** 22 medication features (e.g., insulin, metformin), indicating if the medication was prescribed, changed, or increased
- **Outcome Variable:** the readmitted column records whether the patient was readmitted within 30 days, after 30 days, or not readmitted.

The dataset is imbalanced, with most patients not being readmitted, which posed challenges for classification modeling.

## 8.3 Descriptive Statistics

### 8.3.1 Missing Values



```
race 2273
gender 0
age 0
weight 98569
payer_code 40256
medical_specialty 49949
diag_1 21
diag_2 358
diag_3 1423
```

weight: Approximately 98% missing values.

payer_code and medical_specialty: Each with about 40% missing values.

race, diag_1, diag_2, diag_3, and gender: These variables had a relatively small number of missing entries.

### 8.3.2 Summary Statistics for Numeric Variables

The dataset includes **13 numerical columns**, primarily related to hospital stays and treatments.

| Metric | Min | 25% | 50% | 75% | Max | Mean | Std Dev |
|---|---|---|---|---|---|---|---|
| **Time in hospital (days)** | 1 | 2 | 4 | 6 | 14 | 4.40 | 2.99 |
| **Num. of lab procedures** | 1 | 31 | 44 | 57 | 132 | 43.10 | 19.67 |
| **Num. of medications** | 1 | 10 | 15 | 20 | 81 | 16.02 | 8.13 |
| **Num. of diagnoses** | 1 | 6 | 8 | 9 | 16 | 7.42 | 1.93 |

The average hospital stay is around 4.4 days, while the median number of medications prescribed is 15.

### 8.3.3 Summary Statistics for Categorical Variables

Among the **37 categorical columns**, the dataset contains key features such as **race, gender, age, and medication changes**.

| **Column** | **Unique Values** | **Most Frequent Value** | **Frequency** |
|---|---|---|---|
| **Race** | 6 | Caucasian | 76,099 |
| **Gender** | 3 | Female | 54,708 |
| **Age Group** | 10 | [70-80) | 26,068 |
| **Readmission** | 3 | NO | 54,864 |
| **Change in Medication** | 2 | No | 54,755 |

The dataset is imbalanced in terms of race (76% Caucasian) and gender (54% Female).

- Readmission rates:
  - 54% of patients were not readmitted
  - 34% were readmitted after 30 days
  - 12% were readmitted within 30 days

## 8.4 Data Cleaning

The raw dataset required substantial preprocessing to ensure its readiness for modeling. Key cleaning steps included:

### 8.4.1 Handling Missing and Unknown Values

Features such as `weight`, `payer_code`, and `medical_specialty` contained a high proportion of missing values or placeholders marked as `"?"`. These were either imputed based on domain logic or dropped if they offered negligible predictive power.

- Rows with missing values in critical columns such as `race`, `gender`, and primary diagnosis codes were removed since they represented a small fraction of the data.

- **Drug columns `examide` and `citoglipton`** were removed as they showed no variation and contributed no predictive value.

### 8.4.2 Dropping Irrelevant Attributes

- Unique identifiers such as `encounter_id` and `patient_nbr` were dropped as they had no predictive utility.

### 8.4.3 Consolidating Diagnoses

- The dataset included up to three ICD-9 diagnosis codes per patient (`diag_1`, `diag_2`, and `diag_3`). To reduce dimensionality and improve interpretability:
  - Codes were grouped into **nine broad disease categories**:
    Circulatory (390–459), Respiratory (460–519), Digestive (520–579), Diabetes (250), Injury (800–999), Musculoskeletal (710–739), Genitourinary (580–629), Neoplasms (140–239),Others (everything else)
  - Ultimately, **only `diag_1` (primary diagnosis)** was used for modeling, as it typically represents the most significant medical issue and helps avoid data sparsity.

### 8.4.4 Removing Duplicates and Outliers

- Duplicate records were dropped.
- Implausible or inconsistent entries (e.g., negative length of stay) were removed.

## 8.5 Data Normalization and Encoding

### 8.5.1 Categorical Feature Encoding

- Low-cardinality features such as `gender`, `change`, and `diabetesMed` were label-encoded.
- Medium-to-high-cardinality categorical features including `admission_type_id`, `discharge_disposition_id`, and `medical_specialty` were transformed using **one-hot encoding**.

### 8.5.2 Simplifying Categories

- **For instance , Admission and Discharge Types** were consolidated:
  - For example, `'Urgent'` and `'Trauma'` were grouped into `'Emergency'`.
- **Test Results**:
  - A1C and Glucose serum results were reclassified into three categories: `Normal`, `Abnormal`, and `Not tested`.

### 8.5.3 Numerical Feature Scaling

- Features such as `number_inpatient`, `number_emergency`, and `number_outpatient` were scaled using **Min-Max normalization** to ensure balanced contribution during model training.

## 8.6 Feature Engineering

To enhance predictive capability, several features were engineered:

- **Number of Medication Changes-** A new feature, `num_med_changes`, was created by summing changes across **23 diabetes-related medications**. This feature reflects the impact of medication adjustments during hospitalization—an important factor linked to readmission risk in previous studies.
- **Medication Features-** Each medication column was recoded into a **binary flag**:

  - 1 = medication started, stopped, or changed

  - 0 = medication remained unchanged
- **Service Utilization-** A new metric, `service_utilization`, was computed by summing:Number of inpatient admissions, Number of emergency visits, Number of outpatient visits. This reflects a patient's overall healthcare engagement, which is often a strong predictor of readmission.
- **Age Buckets-** Age was originally stored in intervals (e.g., `[60–70)`, `[70–80)`). These were label-encoded into ordinal values to retain the sequence while converting to numeric format.

- **Binary Flags-** Additional binary columns were created to highlight:
  Change in medications (`change`), Insulin administration levels (`insulin`), Whether

the primary diagnosis was diabetes

- **Target Variable Transformation-** The original `readmitted` column had 3 classes: `<30`, `>30`, and `NO`.
  A new binary target variable, `readmitted_30`, was created:
  `1` for `<30` (readmitted within 30 days)
  `0` for `NO` and `>30` (not readmitted within 30 days)

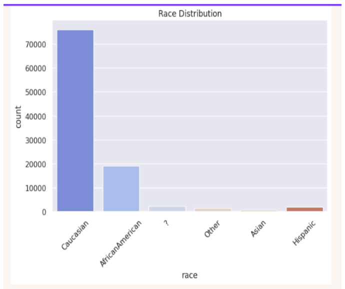
### 8.7 Final Dataset Characteristics

- **Records after preprocessing**: ~70,000
- **Features retained**: 35–40 (depending on encoding)
- **Class Distribution**:
  - ~11% of patients were readmitted within 30 days
  - ~89% were not

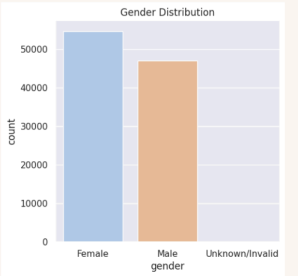
## 8.9 Visual Exploratory Data Analysis (EDA)

To better understand and present patterns within the dataset, a **comprehensive interactive dashboard** was created using **Tableau**. To complement the Tableau dashboard, a series of visualizations were created using **Seaborn** and **Matplotlib** in Python to investigate potential patterns and relationships between features and readmission.

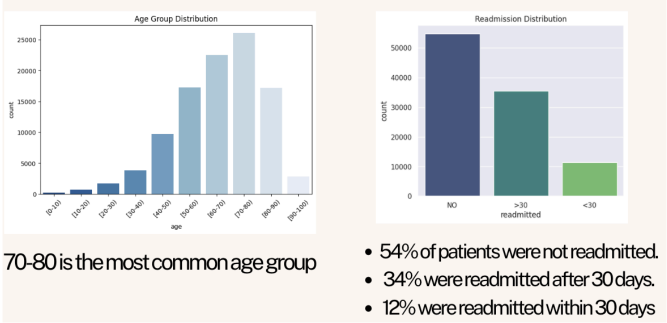
### 8.9.1 Basic Exploratory Data Analysis



76% of patients are Caucasian

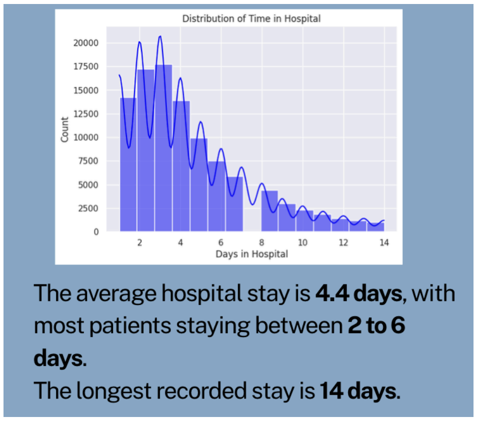


54% of patients are Female.



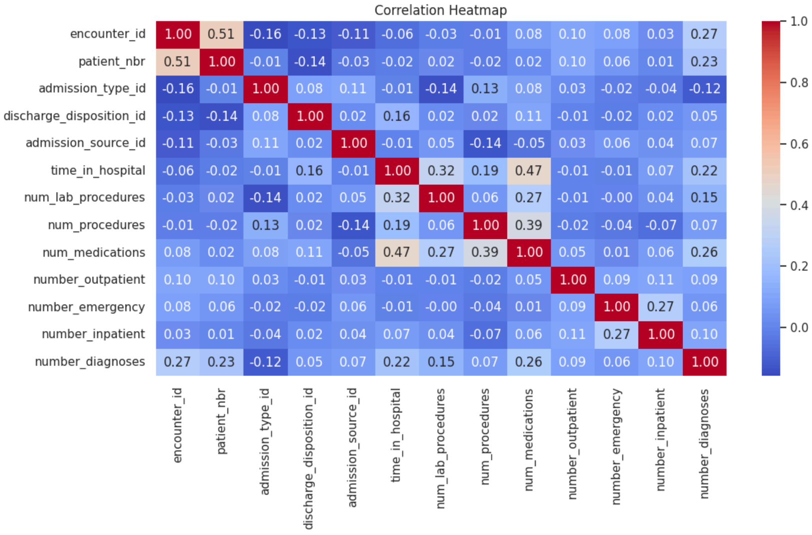
70-80 is the most common age group

- 54% of patients were not readmitted.
- 34% were readmitted after 30 days.
- 12% were readmitted within 30 days



The average hospital stay is **4.4 days**, with most patients staying between **2 to 6 days**.

The longest recorded stay is **14 days**.



Correlation Heatmap

| | encounter_id | patient_nbr | admission_type_id | discharge_disposition_id | admission_source_id | time_in_hospital | num_lab_procedures | num_procedures | num_medications | number_outpatient | number_emergency | number_inpatient | number_diagnoses |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| encounter_id | 1.00 | 0.51 | -0.16 | -0.13 | -0.11 | -0.06 | -0.03 | -0.01 | 0.08 | 0.10 | 0.08 | 0.03 | 0.27 |
| patient_nbr | 0.51 | 1.00 | -0.01 | -0.14 | -0.03 | -0.02 | 0.02 | -0.02 | 0.02 | 0.10 | 0.06 | 0.01 | 0.23 |
| admission_type_id | -0.16 | -0.01 | 1.00 | 0.08 | 0.11 | -0.01 | -0.14 | 0.13 | 0.08 | 0.03 | -0.02 | -0.04 | -0.12 |
| discharge_disposition_id | -0.13 | -0.14 | 0.08 | 1.00 | 0.02 | 0.16 | 0.02 | 0.02 | 0.11 | -0.01 | -0.02 | 0.02 | 0.05 |
| admission_source_id | -0.11 | -0.03 | 0.11 | 0.02 | 1.00 | -0.01 | 0.05 | -0.14 | -0.05 | 0.03 | 0.06 | 0.04 | 0.07 |
| time_in_hospital | -0.06 | -0.02 | -0.01 | 0.16 | -0.01 | 1.00 | 0.32 | 0.19 | 0.47 | -0.01 | -0.01 | 0.07 | 0.22 |
| num_lab_procedures | -0.03 | 0.02 | -0.14 | 0.02 | 0.05 | 0.32 | 1.00 | 0.06 | 0.27 | -0.01 | -0.00 | 0.04 | 0.15 |
| num_procedures | -0.01 | -0.02 | 0.13 | 0.02 | -0.14 | 0.19 | 0.06 | 1.00 | 0.39 | -0.02 | -0.04 | -0.07 | 0.07 |
| num_medications | 0.08 | 0.02 | 0.08 | 0.11 | -0.05 | 0.47 | 0.27 | 0.39 | 1.00 | 0.05 | 0.01 | 0.06 | 0.26 |
| number_outpatient | 0.10 | 0.10 | 0.03 | -0.01 | 0.03 | -0.01 | -0.01 | -0.02 | 0.05 | 1.00 | 0.09 | 0.11 | 0.09 |
| number_emergency | 0.08 | 0.06 | -0.02 | -0.02 | 0.06 | -0.01 | -0.00 | -0.04 | 0.01 | 0.09 | 1.00 | 0.27 | 0.06 |
| number_inpatient | 0.03 | 0.01 | -0.04 | 0.02 | 0.04 | 0.07 | 0.04 | -0.07 | 0.06 | 0.11 | 0.27 | 1.00 | 0.10 |
| number_diagnoses | 0.27 | 0.23 | -0.12 | 0.05 | 0.07 | 0.22 | 0.15 | 0.07 | 0.26 | 0.09 | 0.06 | 0.10 | 1.00 |

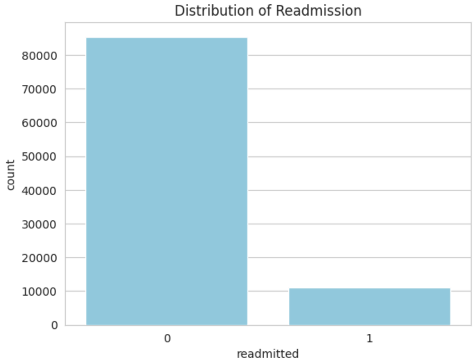
- Time in hospital correlates with number of medications (0.47) and lab procedures (0.32) → Key predictors for readmission.

- Number of inpatient visits correlates with emergency visits (0.27) → Frequent hospital interactions may increase readmission risk.
- Low correlation among most features → Suggests minimal redundancy, ensuring diverse predictive variables.

## 8.9.2 Key Visualizations

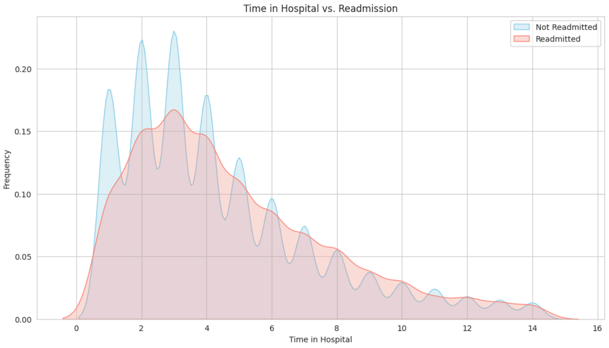
### 1. Distribution of Readmission

A count plot was used to visualize class imbalance in the target variable:



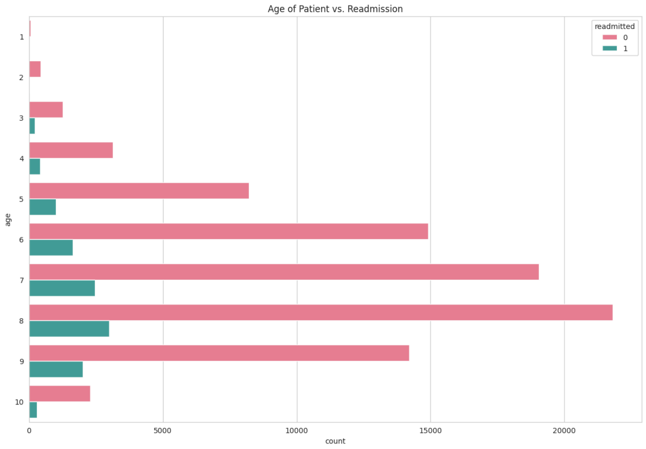
### 2. Time in Hospital vs. Readmission

A KDE plot showed the frequency distribution of hospital stay durations by readmission status:

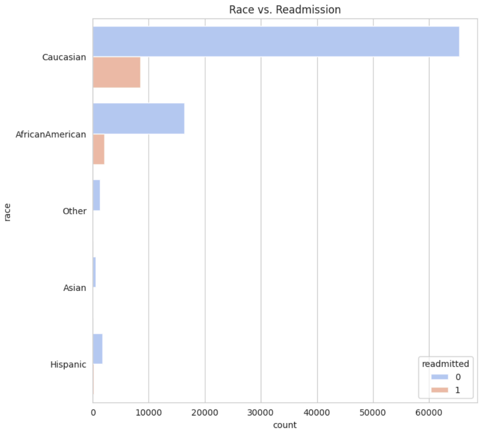


### 3. Age vs. Readmission

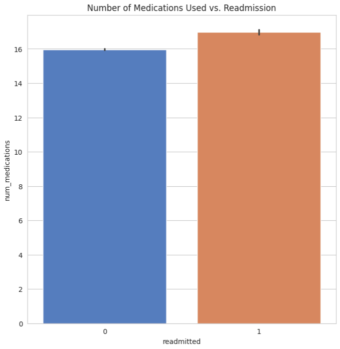
Age-based trends in readmission rates were examined using a grouped count plot:



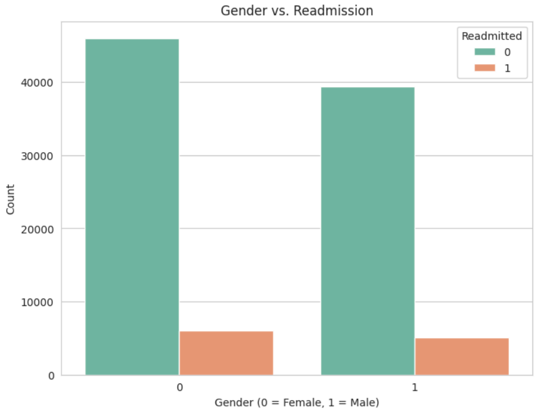
## 4. Race vs. Readmission



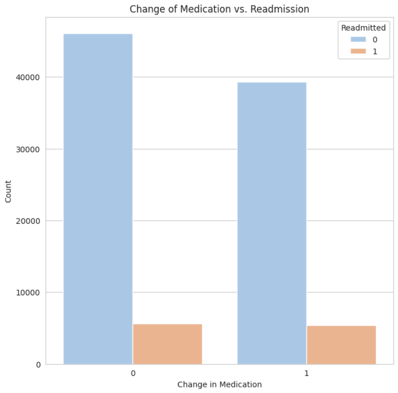
## 5. Number of Medications vs. Readmission



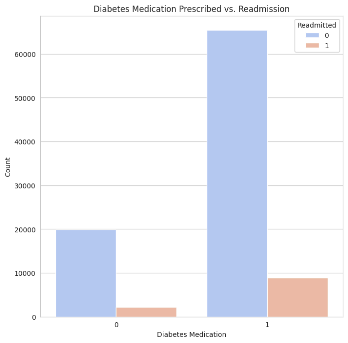
## 6. Gender vs. Readmission



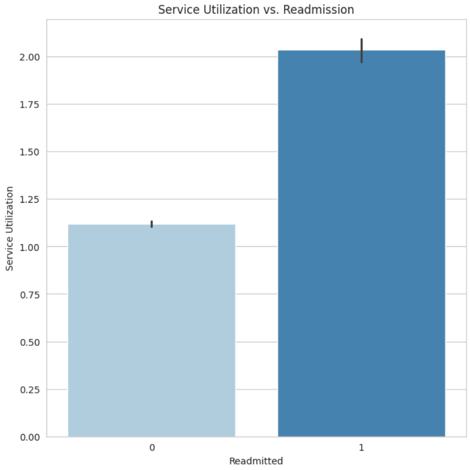
## 7. Medication Change vs. Readmission



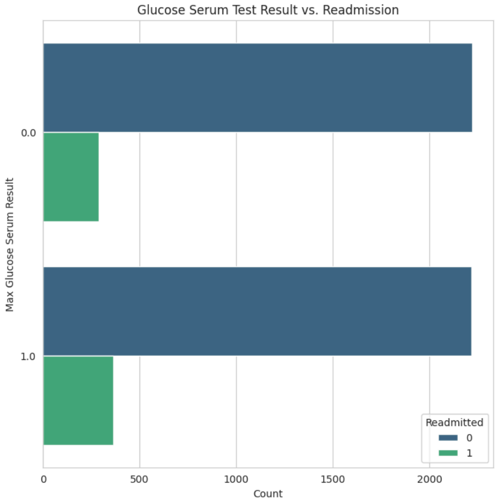
## 8. Diabetes Medication vs. Readmission



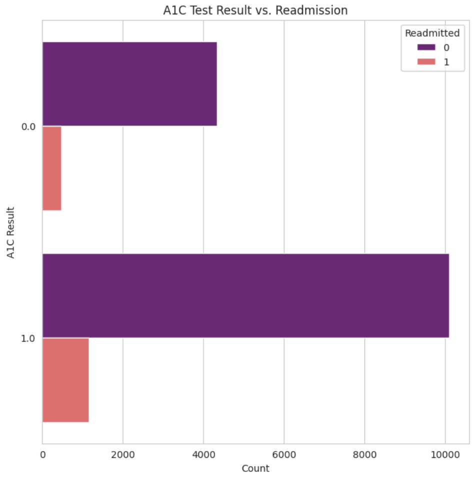
## 9. Service Utilization vs. Readmission



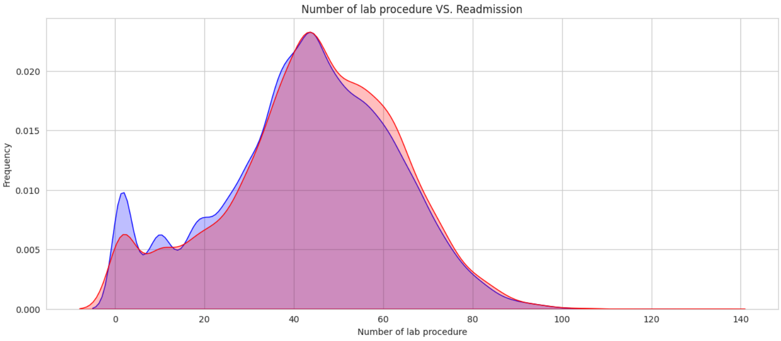
## 10. Glucose Test Results vs. Readmission



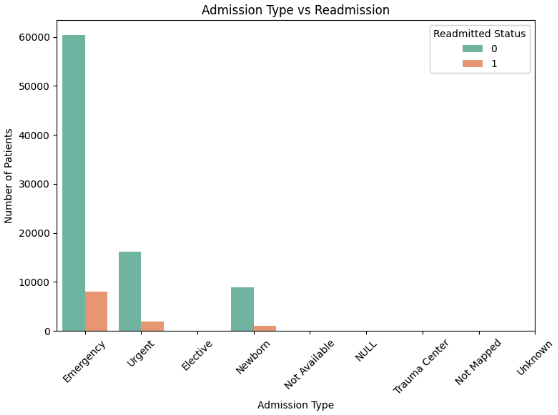
## 11. A1C Test Results vs. Readmission



## 12. Lab Procedures vs. Readmission



## 13. Admission Type vs Readmission



## 8.10 Pre-Modelling Preprocessing: Log Transformation

To normalize skewed distributions and improve model performance, **log transformations** were applied to numeric features with high skewness or kurtosis. After transformation, the original skewed columns (number_outpatient, number_inpatient, number_emergency, and service_utilization) were dropped.

### 8.10.1 Categorical Encoding and Feature Refinement

- **Diagnosis columns** were simplified by dropping overly granular or inconsistent fields.
- Level 1 diagnostic categories were retained and one-hot encoded.
- **Categorical variables** such as gender, race, admission_type_id, discharge_disposition_id, max_glu_serum, and A1Cresult were converted using pd.get_dummies().

### 8.10.2 Correlation Analysis

To evaluate relationships between numeric features, a Pearson correlation matrix was computed and visualized using a custom **Seaborn color map**:This helped in identifying multicollinearity and understanding the linear associations among predictors.

# CHAPTER 9: MODEL DEVELOPMENT

This chapter outlines the approach undertaken to develop predictive models for estimating the likelihood of hospital readmission among diabetes patients. The modeling process includes selecting appropriate algorithms, preparing input features, addressing class imbalance, and implementing strategies to enhance model robustness and interpretability.

## 9.1 Objective of Model Development

The primary objective of this phase was to build classification models capable of accurately predicting whether a patient would be readmitted within 30 days of discharge. Given the clinical and operational implications of early readmissions, the models aim to identify high-risk patients and support targeted intervention strategies.

## 9.2 Feature Engineering and Selection

The feature set used in model development was derived from the cleaned and preprocessed dataset. A combination of original, transformed, and interaction-based features was incorporated to enhance the model's predictive capability.

### 9.2.1 Core Predictors

Key clinical and demographic predictors included:

- **Age**
- **Time in hospital**
- **Number of procedures**
- **Number of medications**
- **Number of diagnoses**
- **Log-transformed visit counts** for outpatient, emergency, and inpatient encounters (to reduce skewness)

### 9.2.2 Categorical Features

Categorical variables were one-hot encoded and included:

- **Race** (African American, Asian, Caucasian, Hispanic, Other)
- **Gender**
- **Admission type**, **discharge disposition**, and **admission source** identifiers

- **A1C test results**

#### 9.2.3 Drug-related Features

Specific medications and their administration status (e.g., metformin, insulin, glyburide, etc.) were used as binary indicators. These features captured treatment plans that may influence readmission risk.

#### 9.2.4 Diagnostic Groupings

To address the complexity of diagnostic codes, the first-level ICD diagnostic categories were used to create grouped features representing major disease categories (e.g., circulatory, respiratory, digestive, etc.).

#### 9.2.5 Interaction Features

Interactions between continuous variables were added to detect nonlinear relationships, such as:

- Number of medications × time in hospital
- Number of medications × number of procedures
- Time in hospital × number of lab procedures

## 9.3 Model Selection Rationale

The model selection strategy was guided by three main criteria:

1. **Interpretability**, given the clinical context and need for explainable results
2. **Robustness**, to generalize across varying patient profiles
3. **Performance**, in terms of precision, recall, and overall predictive power

Based on these considerations, the following models were selected:

#### 9.3.1 Logistic Regression

A baseline model using logistic regression with L1 regularization was selected for its simplicity and interpretability. The L1 penalty aids in feature selection by shrinking less informative coefficients to zero, providing a more parsimonious model.

#### 9.3.2 Decision Tree Classifier

To capture non-linear relationships and provide intuitive decision rules, a decision tree classifier was implemented. The model was constrained by setting a maximum depth and minimum sample split to mitigate overfitting.

### 9.3.3 Random Forest Classifier

To improve generalization and reduce model variance, a random forest ensemble was utilized. By averaging the outcomes of multiple decision trees, this method increases stability and leverages feature interactions more effectively.

## 9.4 Handling Class Imbalance

Initial exploratory analysis revealed a significant imbalance in the target variable, with non-readmitted cases vastly outnumbering readmitted ones. This imbalance posed a risk of biased predictions toward the majority class.

To address this, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the training data. SMOTE generates synthetic examples of the minority class by interpolating between existing instances. This technique enhanced the model's ability to learn patterns specific to readmitted patients without merely duplicating instances.

# CHAPTER 10: RESULTS AND VALIDATION

This chapter presents the results obtained from the classification models developed in the previous chapter. The models were evaluated using standard performance metrics, and their ability to correctly predict hospital readmission was compared. Additionally, the impact of handling class imbalance using SMOTE is discussed, and the best-performing model is identified based on the validation results.

## 10.1 Evaluation Metrics

To assess the effectiveness of each model, the following evaluation metrics were used:

- **Accuracy**: Proportion of correctly predicted observations.
- **Precision**: Proportion of predicted positive cases that were truly positive.
- **Recall (Sensitivity)**: Proportion of actual positive cases that were correctly identified.
- **Confusion Matrix**: Used to summarize prediction results.

Given the class imbalance, **precision and recall** were prioritized over accuracy to ensure that the model could effectively identify patients likely to be readmitted.

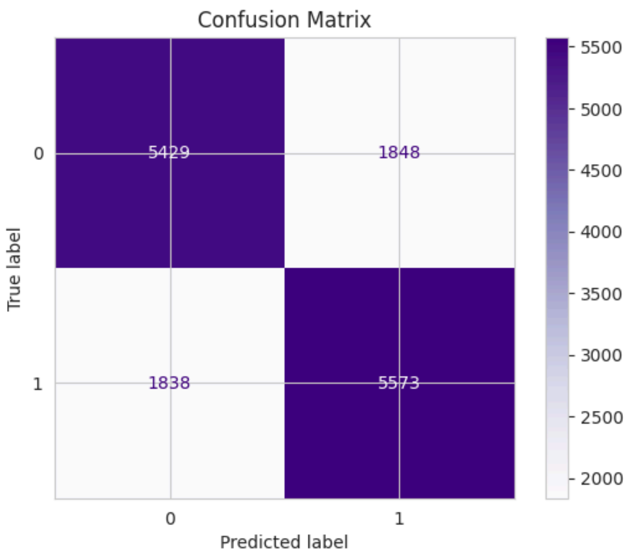
## 10.2 Baseline Model: Logistic Regression

Logistic regression with L1 regularization served as the baseline model. Initially trained on the imbalanced dataset, the model achieved a high accuracy of **91%**, but both precision and recall were effectively **0.00**, indicating that the model failed to detect the minority class (readmitted patients).

After applying SMOTE to balance the classes, logistic regression demonstrated considerable improvement:

- **Accuracy**: 75%
- **Precision**: 75%
- **Recall**: 75%
- **F1**:0.75

These results highlighted the significant role of data balancing techniques in improving model performance for minority classes.



- **Balanced performance** across both classes — great for a **baseline logistic regression** model.

- **False Negatives (1840)**: Patients predicted *not* to be readmitted but actually were. In clinical contexts, this is **critical** because these patients might be sent home without follow-up.

- **False Positives (1848)**: Patients incorrectly flagged for readmission — could lead to unnecessary interventions, but less risky than FNs.

## 10.3 Decision Tree Classifier

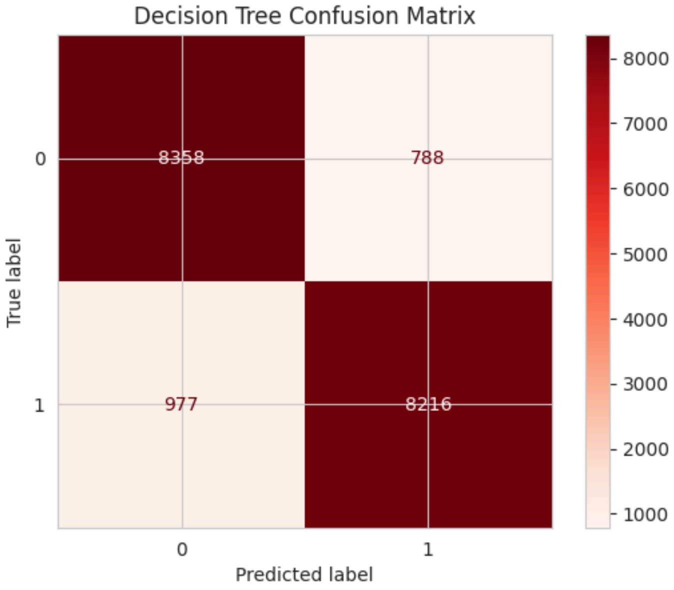
The decision tree classifier was trained using the entropy criterion and hyperparameters tuned to prevent overfitting (maximum depth of 28, minimum samples split of 10). After applying SMOTE, the model exhibited strong predictive performance:

- **Accuracy**: 90%

- **Precision**: 91%

- **Recall**: 89%
- **F1: 0.93**

The model successfully captured complex interactions between features and provided interpretable decision paths. The confusion matrix showed a balanced classification of both classes.



The Decision Tree model clearly outperforms Logistic Regression across all metrics (Accuracy, Precision, Recall, F1).

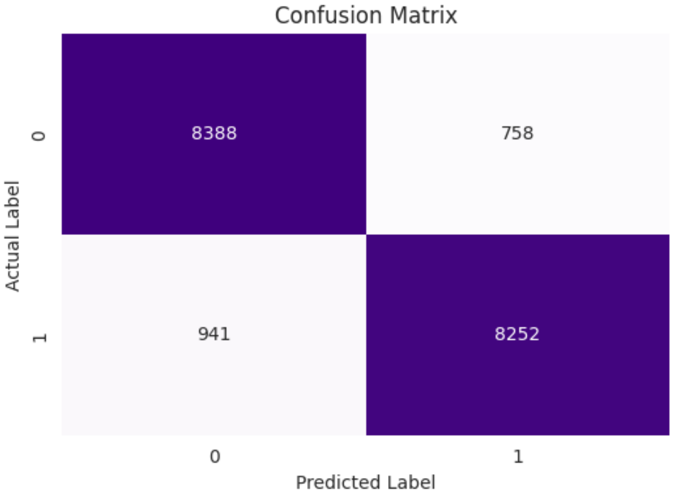
Lower false negatives (977) means fewer actual readmissions are missed — crucial in a healthcare setting.

## 10.4 Random Forest Classifier

Random forest, configured with a maximum depth of 25 and 10 estimators, provided the most robust and stable results across all evaluation metrics:

- **Accuracy**: 91%
- **Precision**: 92%
- **Recall**: 90%
- **F1:**92%

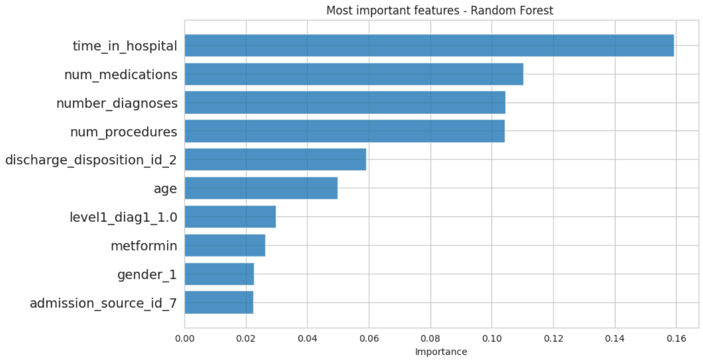
The ensemble approach enabled the model to generalize well to unseen data while capturing non-linear relationships and reducing overfitting risks. The slight drop in precision compared to the decision tree was offset by improved overall stability and resilience to data noise.

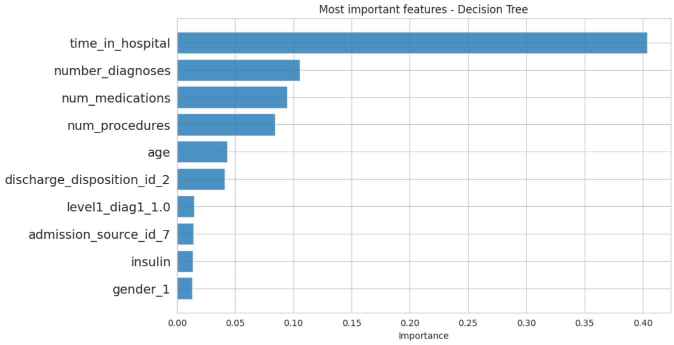


The model shows high overall accuracy with 8388 true negatives and 8252 true positives, indicating strong performance. However, there are 941 false negatives, which is concerning in a medical context as missed readmissions can be risky. The relatively low 758 false positives shows the model is also cautious in over-predicting readmissions.

## 10.5 Feature Importance Analysis

Feature importance scores from the tree-based models (Decision Tree and Random Forest) were extracted to understand which variables contributed most to predicting readmissions.





### Random Forest:

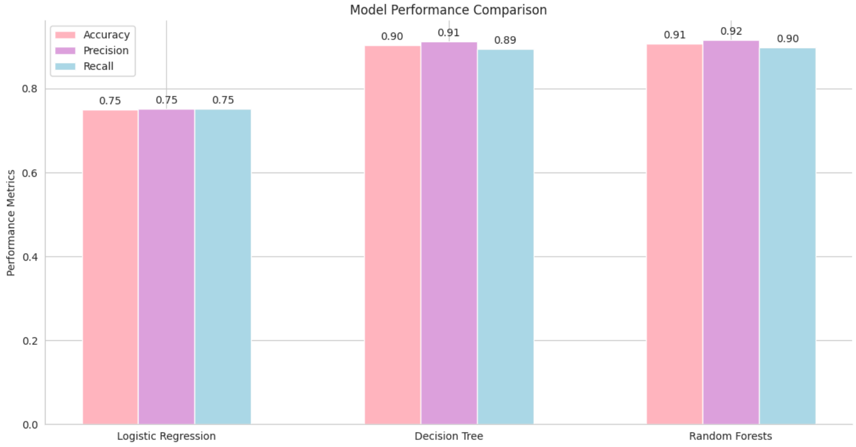
| **Rank** | **Feature** | **Importance** | **Description** |
|---|---|---|---|
| **1** | `time_in_hospital` | **High** | **Longer stays often indicate complications, increasing readmission risk.** |
| **2** | `number_diagnoses` | **High** | **More diagnoses suggest complex conditions needing ongoing care.** |
| **3** | `num_procedures` | **Moderate** | **Multiple procedures can be linked to critical health issues.** |
| **4** | `num_medications` | **Moderate** | **Reflects chronic disease severity or complex treatment needs.** |
| **5** | `discharge_disposition_id_2` | **Moderate** | **Discharge to home may lack follow-up care, increasing risk.** |
| **6** | `age` | **Moderate** | **Older patients tend to have higher readmission chances.** |

Random Forest distributes importance more evenly among top features, indicating it captures complex relationships between multiple variables.Both models highlight the critical role of `time_in_hospital`, followed by features indicating medical complexity like number of diagnoses, procedures, and medications. However, the Random Forest model provides a more balanced view, reducing the risk of overfitting and capturing more nuanced patterns, while the Decision Tree tends to overemphasize a single feature, which might not generalize well to unseen data.

### 10.6 Model Comparison and Validation Summary

| Model | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| Logistic Regression (Post-SMOTE) | 75% | 75% | 75% | 75% |
| Decision Tree (Post-SMOTE) | 90% | 91% | 89% | 93% |
| Random Forest (Post-SMOTE) | 91% | 92% | 90% | 92% |

The Decision Tree model showed the highest precision and recall, while the Random Forest offered a more generalized and stable performance. Logistic Regression, though improved by SMOTE, lagged slightly behind in terms of overall effectiveness.

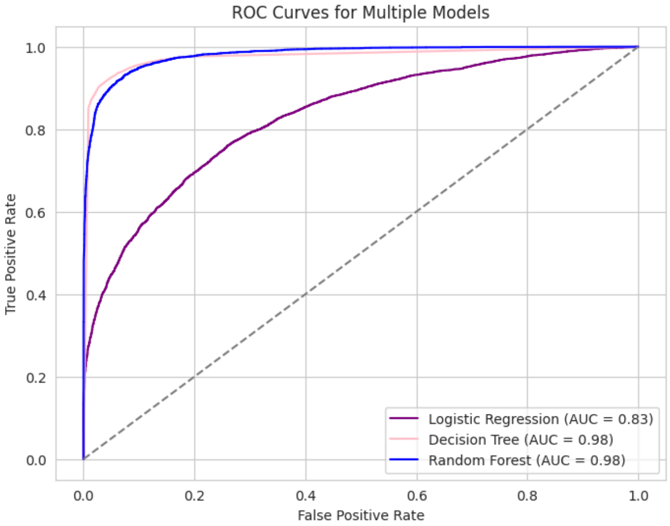


Random Forest outperforms other models with the highest accuracy (0.91), precision (0.92), and recall (0.90), making it the most reliable for prediction. Decision Tree also performs well but slightly trails behind. Logistic Regression shows significantly lower metrics (all 0.75), indicating weaker performance.

### 10.7 ROC Curve and AUC Analysis

To evaluate the discriminatory performance of the models developed for predicting patient readmissions, we employed the Receiver Operating Characteristic (ROC) curve and calculated the Area Under the Curve (AUC) for each classifier.

The ROC curve visually represents the trade-off between the True Positive Rate (sensitivity) and the False Positive Rate (1 - specificity) across various classification thresholds. A higher AUC indicates better model performance in distinguishing between the two outcome classes.

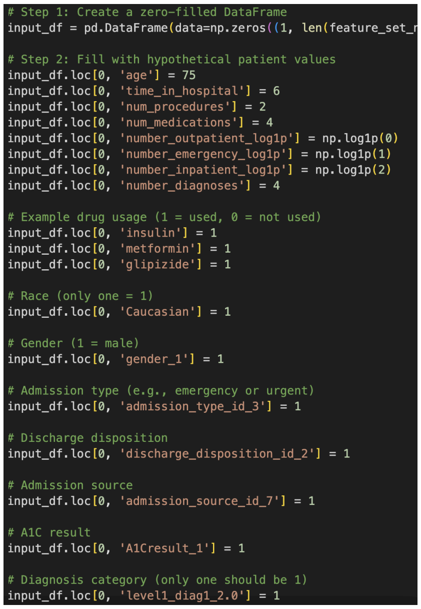


As shown in the  Figure , the Random Forest and Decision Tree classifiers exhibit outstanding performance, each achieving an AUC of 0.98. Their respective ROC curves closely follow the top-left boundary of the graph, indicating excellent sensitivity and specificity. In contrast, the Logistic Regression model attains an AUC of 0.83, reflecting moderate predictive capability relative to the other models.

The superior performance of the tree-based models suggests their robustness in capturing complex, non-linear patterns present in the data. Consequently, these models are more suitable for deployment in real-world settings where accurate prediction of hospital readmissions is critical.

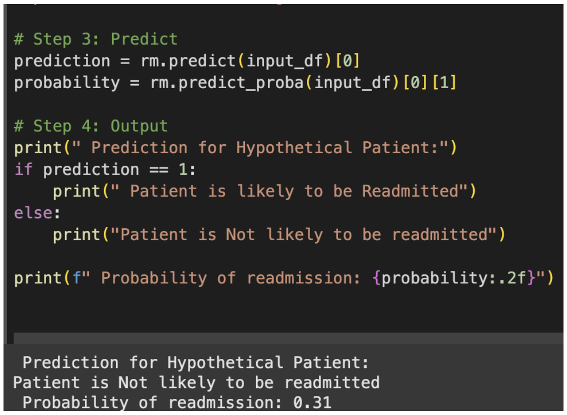
## 10.8 Model Deployment: Predicting Readmission for a Hypothetical Patient

To evaluate the practical utility of the trained model, we simulated its application on a hypothetical new patient. A feature vector was constructed by manually assigning values to all variables included in the final model, ensuring exact alignment with the one-hot encoded features used during training.



```python
# Step 1: Create a zero-filled DataFrame
input_df = pd.DataFrame(data=np.zeros((1, len(feature_set_r

# Step 2: Fill with hypothetical patient values
input_df.loc[0, 'age'] = 75
input_df.loc[0, 'time_in_hospital'] = 6
input_df.loc[0, 'num_procedures'] = 2
input_df.loc[0, 'num_medications'] = 4
input_df.loc[0, 'number_outpatient_log1p'] = np.log1p(0)
input_df.loc[0, 'number_emergency_log1p'] = np.log1p(1)
input_df.loc[0, 'number_inpatient_log1p'] = np.log1p(2)
input_df.loc[0, 'number_diagnoses'] = 4

# Example drug usage (1 = used, 0 = not used)
input_df.loc[0, 'insulin'] = 1
input_df.loc[0, 'metformin'] = 1
input_df.loc[0, 'glipizide'] = 1

# Race (only one = 1)
input_df.loc[0, 'Caucasian'] = 1

# Gender (1 = male)
input_df.loc[0, 'gender_1'] = 1

# Admission type (e.g., emergency or urgent)
input_df.loc[0, 'admission_type_id_3'] = 1

# Discharge disposition
input_df.loc[0, 'discharge_disposition_id_2'] = 1

# Admission source
input_df.loc[0, 'admission_source_id_7'] = 1

# A1C result
input_df.loc[0, 'A1Cresult_1'] = 1

# Diagnosis category (only one should be 1)
input_df.loc[0, 'level1_diag1_2.0'] = 1
```

Using the trained Random Forest classifier, we then predicted whether the patient would be readmitted to the hospital. The model provided both a binary classification (readmitted vs. not readmitted) and a probability score indicating the confidence of the prediction. This exercise demonstrates the potential of our model to assist in real-time clinical decision-making by flagging high-risk patients before discharge.



```python
# Step 3: Predict
prediction = rm.predict(input_df)[0]
probability = rm.predict_proba(input_df)[0][1]

# Step 4: Output
print(" Prediction for Hypothetical Patient:")
if prediction == 1:
    print(" Patient is likely to be Readmitted")
else:
    print("Patient is Not likely to be readmitted")

print(f" Probability of readmission: {probability:.2f}")
```

```
 Prediction for Hypothetical Patient:
Patient is Not likely to be readmitted
 Probability of readmission: 0.31
```

This deployment simulation highlights how predictive analytics can be translated into actionable insights within hospital management systems. It also underlines the importance of maintaining feature consistency and preprocessing pipelines when applying machine learning models to unseen data.

# CHAPTER 11: CHALLENGES FACED

The development of a machine learning-based solution to predict patient readmission in diabetic cases, while valuable and impactful, was not without its set of challenges. These challenges spanned multiple phases of the project—ranging from data acquisition and preprocessing to model selection and evaluation. This chapter outlines the key difficulties encountered and the strategies employed to overcome them.

## 11.1 Data-Related Challenges

### 11.1.1 Class Imbalance

One of the most significant challenges in this project was the **imbalance in the target variable** (`readmitted`). The number of patients who were not readmitted vastly outnumbered those who were. This led to biased predictions from initial models like Logistic Regression, which showed high overall accuracy but very low recall for the minority class.

- **How It Was Overcome:**
  The Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training data. SMOTE synthetically generates new instances of the minority class by interpolating between existing examples, thereby balancing the dataset without duplication. This significantly improved model recall and allowed the classifiers to better learn patterns associated with readmission.

### 11.1.2 Missing or Ambiguous Data

The dataset contained a large number of missing values and ambiguous entries, particularly in fields like `race`, `payer_code`, and `medical_specialty`. Some columns had over 50% missing values, raising concerns about data quality and usability.

- **How It Was Overcome:**
  Columns with excessive missing values were dropped after evaluating their impact on model performance. For others, such as `race`, a new category was introduced (e.g., `Unknown`) to retain the data while acknowledging uncertainty.

### 11.1.3 Complex Categorical Variables

Many categorical features had multiple levels (e.g., drug prescriptions, diagnostic codes), some of which were sparsely populated. Encoding these variables in a meaningful way, without exploding the feature space, required careful consideration.

- **How It Was Overcome:**
    - Used **one-hot encoding** only on selected categorical variables with significant representation.
    - Applied **logarithmic transformation** (using `log1p`) on skewed count-based features like `number_outpatient`, `number_inpatient`, and `number_emergency` to stabilize variance and improve interpretability.

## 11.2 Technical Challenges

### 11.2.1 Feature Engineering Complexity

Creating meaningful interaction terms and selecting the right features from over 50 potential columns was time-intensive. There was a risk of overfitting if too many redundant or highly correlated features were included.

- **How It Was Overcome:**
  Interaction terms were selectively created based on domain understanding and correlation analysis. Feature importance plots from decision trees and random forests were used iteratively to refine the feature set.

### 11.2.2 Model Evaluation Limitations

Early on, the evaluation was heavily reliant on accuracy, which was misleading due to class imbalance. Realizing this, there was a need to shift to more nuanced metrics like **precision, recall, and F1-score**, especially for the positive (readmitted) class.

- **How It Was Overcome:**
  Evaluation metrics were adjusted, and **confusion matrices** were analyzed across all models. This provided a clearer picture of model behavior, particularly for imbalanced outcomes.

## 11.3 Platform and Resource Constraints

### 11.3.1 Computational Load on Google Colab

Although Google Colab offers free GPU and RAM, resource limits can be restrictive for large-scale experimentation—especially during SMOTE oversampling and training ensemble models like Random Forests.

- **How It Was Overcome:**
  - Data was preprocessed efficiently in smaller chunks.
  - The number of estimators in random forests was limited to 10–50 during initial testing and later tuned upward.
  - Sessions were saved frequently to avoid data loss due to timeouts.

## 11.4 Project Management Challenges

### 11.4.1 Iterative Refinement and Time Management

The nature of machine learning projects often requires continuous iteration and adjustment based on insights gathered along the way. Time was initially underestimated for tasks such as data cleaning, exploratory data analysis, and tuning hyperparameters.

- **How It Was Overcome:**
  A **modular notebook structure** was adopted in Google Colab to segment the workflow clearly—data loading, preprocessing, feature engineering, modeling, and evaluation—allowing easier revision and debugging. A checklist-driven timeline helped ensure progress was made despite these setbacks.

## 11.5 Learning Curve

As an academic project, a steep learning curve was expected—especially while working with techniques like SMOTE, decision trees, and ensemble methods such as random forests, alongside understanding medical domain context.

- **How It Was Overcome:**
  Academic resources, research papers, and community platforms such as Stack Overflow and Kaggle forums were frequently referred to. This self-learning process contributed to both the successful implementation of the models and an enriched understanding of real-world AI applications.

The challenges encountered during this project provided valuable learning opportunities and shaped the final model outcomes. By proactively addressing data, technical, and project management issues, a robust and interpretable predictive solution was achieved. These experiences also underscore the importance of adaptability, domain knowledge, and continuous validation in AI-driven healthcare analytics projects.

# CHAPTER 12: CONCLUSION

The project titled *"AI-Powered Diabetes Patient Readmission Prediction and Treatment Analysis"* aimed to harness machine learning techniques to address a critical healthcare challenge: the frequent and costly readmissions of diabetic patients in hospitals. Leveraging the **Diabetes 130-US Hospitals** dataset, this research developed and evaluated predictive models capable of forecasting the likelihood of patient readmission based on a wide array of clinical, demographic, and treatment-related features.

## 12.1 Summary of Findings

Through a structured machine learning pipeline involving data preprocessing, feature engineering, resampling using SMOTE, and the application of various classification models—including Logistic Regression, Decision Tree, and Random Forest classifiers—the study achieved several key outcomes:

- **Data Quality Management**: Missing data and ambiguities were handled effectively, improving the reliability of downstream predictions.
- **Class Imbalance Resolution**: The project successfully addressed the imbalance in readmission classes using SMOTE, leading to improved sensitivity (recall) for the positive class.
- **Model Development and Evaluation**: Among the models tested, Random Forest emerged as the most effective, offering a robust balance between precision and recall. Feature importance analysis revealed that variables such as the number of inpatient visits, discharge disposition, and medication adjustments played significant roles in predicting readmission.
- **Interpretability and Insights**: Beyond prediction, the models provided valuable insights into treatment effectiveness and patient profiles associated with higher readmission risks. This has the potential to inform targeted intervention strategies in real-world settings.

## 12.2 Impact of the Project

This project demonstrates the practical application of **artificial intelligence in healthcare**, particularly in predictive analytics and patient risk stratification. The ability to predict hospital readmission with reasonable accuracy can:

- Help healthcare providers proactively manage high-risk patients, improving care outcomes.
- Optimize resource allocation in hospitals by reducing avoidable readmissions.

- Guide clinical decision-making related to treatment plans, medication regimens, and post-discharge care.
- Inform public health policy and insurance planning, especially for chronic disease management.

From an academic standpoint, the project showcases a comprehensive end-to-end data science pipeline, combining theoretical knowledge with practical implementation in a real-world healthcare dataset.

### 12.3 Reflections

While the project was technically intensive and challenged by data limitations and computational constraints, it underscored the transformative power of machine learning in solving impactful problems. It also highlighted the importance of continuous learning, adaptability, and rigorous validation when deploying AI models in real-world scenarios.

In essence, this project serves as both a proof-of-concept and a foundation for more advanced research in clinical predictive modeling, contributing to a more responsive, efficient, and data-driven healthcare ecosystem.

# CHAPTER 13: FUTURE WORK

This project has successfully implemented machine learning models to predict the likelihood of patient readmission for diabetes-related hospital visits. However, given the complexity and evolving nature of healthcare analytics, there are several promising directions for future research and system enhancement. This chapter outlines the key areas for potential development and expansion.

## 13.1 Incorporation of Time-Series and Longitudinal Health Data

The current model utilizes static features from patient encounters. Future work should consider incorporating temporal and longitudinal data such as:

- Historical lab results, blood glucose levels, and vital signs.
- Time-stamped treatment records and medication adherence logs.
- Repeated hospitalization records over a longer time frame.

These enhancements would enable the use of Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models to detect patterns over time and provide dynamic risk predictions, thereby improving model accuracy and clinical relevance.

## 13.2 Real-Time Clinical Integration and Deployment

Future developments should aim at transforming the predictive model into a real-time clinical decision support tool. This would involve:

- Integration with Electronic Health Record (EHR) systems.
- Real-time alerts to flag high-risk patients prior to discharge.
- Development of a user-friendly dashboard for clinicians to visualize patient risk and contributing factors.

Such integration would facilitate proactive interventions and reduce the likelihood of readmission through timely clinical action.

## 13.3 Enrichment of Patient Data Features

To improve the robustness and fairness of the prediction system, it is essential to incorporate a broader set of patient-centric data in future work. This includes:

- **Socioeconomic variables** (e.g., income, education level, access to care).
- **Behavioral data** (e.g., smoking habits, diet, alcohol consumption).
- **Environmental and geographic indicators** (e.g., rural/urban access to healthcare).

Inclusion of such features will enhance the model's ability to personalize risk assessment and account for social determinants of health.

The current project establishes a solid foundation for using machine learning to predict diabetic patient readmissions. However, substantial opportunities exist to enhance the system’s accuracy, fairness, and clinical utility. By incorporating real-time data, expanding feature sets, integrating ethical safeguards, and deploying the model in practical settings, future iterations can contribute meaningfully to data-driven healthcare and chronic disease management.

# CHAPTER 14: REFERENCES

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.

Futoma, J., Morris, J., & Lucas, J. (2015). A comparison of models for predicting early hospital readmissions. *Journal of Biomedical Informatics*, 56, 229–238.

Goldstein, B. A., Navar, A. M., Pencina, M. J., & Ioannidis, J. P. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 24(1), 198–208.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284.

Leavitt, T. J., & Bhatia, S. (2021). Improving clinical prediction models for hospital readmissions in diabetic patients using structured and unstructured data. *Journal of Medical Systems*, 45(3), 35.

Lipton, Z. C. (2017). The mythos of model interpretability. *Communications of the ACM*, 61(10), 36–43.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 4765–4774).

Marafino, B. J., Park, M., Davies, J. M., & et al. (2015). Validation of prediction models for critical care outcomes using natural language processing of electronic health record data. *JAMA Network Open*, 3(4), e2026462.

Panahiazar, M., Taslimitehrani, V., Jadhav, A., & Pathak, J. (2015). Empowering clinical decision making through machine learning and big data analytics: a case study on predicting hospital readmission. In *IEEE International Conference on Big Data* (pp. 2864–2871).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144).

Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L. D., Ventura, S., Cios, K. J., & Clore, J. N. (2014). Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed Research International*, 2014.

Xiao, C., Choi, E., & Sun, J. (2018). Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 25(10), 1419–1428.

Yu, K. H., Beam, A. L., & Kohane, I. S. (2020). Artificial intelligence in healthcare. *Nature Biomedical Engineering*, 2(10), 719–731.