



**Computing Masters Project**

**CIS4055-N-KB2-2023**

**PREDICTIVE ANALYTICS FOR DIABETIC PATIENT READMISSION USING  
MACHINE LEARNING AND POWERBI**

**SCHOOL OF COMPUTING, ENGINEERING & DIGITAL TECHNOLOGIES**



**D3684573**

**SUBMISSION DATE 3 SEPTEMBER 2024**

**SUPERVISOR  
ALSMADI HIBA**

## **ACKNOWLEDGEMENT**

I would like to express my deepest gratitude to my supervisors, Hiba, Yordanka, and Mansha, for their exceptional guidance, patience, and expert advice throughout the course of this project. Their invaluable insights and unwavering support have been crucial in shaping this research and significantly contributed to my academic development in the field of data science. My heartfelt thanks also goes to my family, whose steadfast support was instrumental throughout my studies. Their encouragement has been a source of strength during challenging times. This project would not have been possible without the collective support and encouragement from these individuals, as well as many others who, although not mentioned by name, have played a vital role in my journey. I am profoundly appreciative of all the guidance and opportunities provided to me.

## **ABSTRACT**

This research addresses the critical issue of high hospital readmission rates among diabetic patients by integrating Machine Learning (ML) algorithms with Power BI for advanced data visualization. The study aims to develop actionable strategies to reduce readmission rates, thereby improving patient outcomes and optimizing healthcare resources. Through a comparative analysis of various classification models, with and without Principal Component Analysis (PCA), the effectiveness of these models in predicting both short-term (within 30 days) and long-term (after 30 days) readmissions was evaluated. The Support Vector Machine (SVM) model emerged as the most effective, achieving a 74% accuracy and an F1-score of 0.74, while the Random Forest model also demonstrated strong performance with 71% accuracy and an F1-score of 0.71. The application of PCA yielded mixed results, enhancing the performance of some models while reducing the effectiveness of others.

Power BI played a crucial role in visualizing these findings, creating interactive dashboards that facilitated a comprehensive analysis of readmission metrics. These dashboards incorporated key factors such as race, age, admission type, and medication use, and featured tools like a dynamic decomposition tree, a heatmap for tracking trends across demographic groups, and a risk stratification tool for forecasting patient outcomes. The insights generated from these visualizations support healthcare professionals in making data-driven decisions, which can lead to a quantifiable reduction in readmission rates, improved patient care, and more efficient allocation of healthcare resources. This project demonstrates the potential of combining advanced ML techniques with robust data visualization to address complex healthcare challenges.

## **TABLE OF CONTENTS**

### **Table of Contents**

1. INTRODUCTION .....	10
1.2 Rationale / Justification of the Study .....	11
1.3 Aim .....	12
1.4 Objectives: .....	12
1.5 Business Intelligence: .....	13
1.6 Research Questions.....	13
1.7 Thesis Structure.....	14
2. LITERATURE REVIEW .....	16
2.1 Defining Hospital Readmissions .....	16
2.2 Challenges in Predicting Diabetic Patient Readmissions.....	17
2.3 Factors and Predictors.....	18
2.4 Limitations and Considerations .....	18
2.5 Practical Applications and Future Research.....	19
2.6 Machine Learning In Hospital Readmission Prediction .....	19
2.7 The Potential of Machine Learning .....	23
2.8 Benefits of Machine Learning for Readmission Prediction .....	23
2.9 Challenges and Considerations in Machine Learning for Healthcare .....	23
2.10 The Role of Business Intelligence .....	24
2.11 Successful Implementation of Business Intelligence .....	24
2.12 Business Intelligence and Organizational Factors .....	25
2.13 Key Capabilities for Business Intelligence Success.....	25
2.14 Integration of Machine Learning and Business Intelligence .....	26
2.15 Case Studies and Practical Applications .....	26
2.16. Future Directions and Research Opportunities .....	29
2.17 Identified Gaps.....	29
2.18 Data Quality and Diversity: .....	30
2.19 Model Performance and Comparison: .....	30
2.20 Operational Data Integration: .....	30
2.21 Real-time Implementation: .....	30
2.22 Intervention Effectiveness: .....	30
2.23 Ethical and Privacy Concerns: .....	31
2.24 Interdisciplinary Collaboration: .....	31
2.25 Cost-effectiveness and Value-for-Money Analysis: .....	31
3. METHODOLOGY.....	33
3.1 Suggested Model.....	33
3.2 Problem Definition and Understanding .....	34

3.3 Data Collection.....	35
3.4 Exploratory Data Analysis (EDA) .....	40
3.4.1 Univariate Analysis .....	40
3.4.2 Number of Lab Procedures: .....	44
3.4.3 Correlation matrix .....	45
3.4.4 Distribution of num_lab_procedures based on Readmission Status .....	46
3.4.5 Distribution of Discharge Dispositions by Readmission Status .....	47
3.5 Key Performance Indicators.....	48
3.6 Data Cleaning.....	67
3.7 Data Pre-processing .....	68
3.7.1 Data Encoding and Transformation .....	69
3.7.2 Correlation Matrix and Feature Selection.....	70
3.7.3 Feature Engineering .....	72
3.7.4 Feature Transformation and normalisation .....	72
3.7.5 Standard Scaler .....	73
3.7.6 Summary of Scaled Data: .....	73
3.7.7 Dimensionality Reduction and PCA Impact.....	74
3.7.8 Distribution of Classes Before and After Applying SMOTE.....	76
3.7.9 Class Distribution After SMOTE:.....	78
3.8 Data Splitting .....	78
3.9 Machine Learning Models .....	78
3.9.1 Logistic Regression (LR):.....	78
3.9.2 KNN: .....	79
3.9.3 Gradient Boosting Machines (GBMs):.....	80
3.9.4 Random Forest (RF):.....	81
3.9.5 Support Vector Machine (SVM): .....	82
3.9.6 Evaluations Metrics .....	84
3.10 Predictive Analytics for Diabetic Patient Readmission Using Business Intelligence Tools .....	87
3.11 Data Preprocessing in Power BI .....	88
3.12 Data Modelling .....	89
3.13 Dax and Measures .....	91
3.14 Data Visualisation.....	92
4. IMPLETEMENTATION, RESUTS AND DISCUSSION .....	93
4.1 Implementation.....	93
4.2 Model Performance Results and Discussion .....	93
4.3 Model Performance With PCA.....	94
4.3.1 Confusion matrix.....	94

4.3.2 Feature Importance .....	95
4.3.3 Risk Stratification.....	96
4.3.4 LIME .....	97
4.4 Model Performance Without PCA .....	98
4.4.1 Confusion Matrix.....	98
4.4.2 Feature Importance .....	99
4.4.3 Risk Stratification.....	100
4.4.4 LIME Explanations .....	101
4.4.5 ROC AUC.....	102
4.4.6 Performance Comparison .....	103
4.5 Power BI.....	105
4.6 Summary of Findings .....	110
5. Discussions: Machine Learning and Power BI Findings .....	111
5.2 Factors Contributing to High Readmission Rates: .....	112
5.3 Predictive Analytics for High-Risk Patients:.....	112
5.4 Impact of Targeted Interventions:.....	113
5.5 Effect of Machine Learning on Healthcare Costs: .....	113
5.6 Gaps in Diabetes Management and Post-Discharge Care:.....	113
5.7 Improvement in Identifying High-Risk Patients: .....	113
5.8 Effective Strategies for Resource Allocation: .....	113
5.9 Impact of Predictive Analytics and BI Tools on Health Outcomes: .....	113
5.6 Limitations.....	114
5.7 Ethical Consideration .....	114
5.7.1 Anonymization and Data Handling: .....	114
5.7.2 Informed Consent: .....	114
5.7.3 Compliance with Data Protection Regulations: .....	115
5.7.4 Prevention of Bias and Fairness: .....	115
5.7.5 Transparency and Accountability: .....	115
6. Conclusion .....	116
6.1 Recommendations.....	116
7. REFERENCES .....	118



## **TABLE OF FIGURES**

Figure 1: Suggested Model Diagram.....	8
Figure 2: Checking First Few Rows.....	40
Figure 3: Age Distribution .....	42
Figure 4: Gender Distribution .....	42
Figure 5: Race Distribution .....	43
Figure 6: Readmitted Distribution .....	43
Figure 7: Time in Hospital by Readmission Status .....	44
Figure 8: Number of Lab Procedures .....	45
Figure 9: Correlation Matrix.....	46
Figure 10: Distribution of Number of Lab Procedures Based on Readmission Status .....	47
Figure 11: Discharge Disposition by Readmission Status .....	48
Figure 12: Total Patients card visual.....	48
Figure 13: Female percentage time in hospital card visual .....	49
Figure 14: Male percentage time in hospital card visual.....	49
Figure 15: Total hospital Readmission card visual.....	50
Figure 16: Readmission Probability card visual.....	50
Figure 17: Time in hospital for male card visual .....	51
Figure 18: Time in hospital for female card visual .....	52
Figure 19: Average time in hospital per gender card visual.....	53
Figure 20: Total time in hospital card visual.....	53
Figure 21: Discharge Disposition .....	54
Figure 22:Diabetic Patients Based on Medication.....	55
Figure 23: Total Readmission by Age .....	55
Figure 24: Total Readmission by Gender .....	56
Figure 25: Patients with Change in medication .....	57
Figure 26: Total Readmission by Race.....	57
Figure 27: Number of Insulin by Age .....	58
Figure 28: Total Patients by Peer group .....	58
Figure 29: Parameter (Readmission Rate).....	59
Figure 30: Parameter (Average time in Hospital) .....	59
Figure 31: Parameter (Average Number of Medication Per Age group) .....	59
Figure 32: Sum of Total in Hospital by Race and Total Patients by Race .....	60
Figure 33: Medications .....	61
Figure 34: Number of Diagnoses, Procedures and Lab .....	62
Figure 35 : SVM Risk Stratification.....	62
Figure 36: Tuned GBM Classification Report KPI .....	63
Figure 37: Tuned KNN Classification Report KPI .....	63
Figure 38: Tuned Random Classification Report KPI .....	63
Figure 39: Tuned Logistic Regression Classification Report KPI .....	64
Figure 40: Tuned SVM Classification Report KPI.....	64
Figure 41: Model Comparison KPI .....	65
Figure 42: SVM LIME Explanation .....	65
Figure 43: Log Regression Feature importance .....	66
Figure 44: Random Forest Feature importance .....	66
Figure 45: GBM Feature importance .....	66
Figure 46: SVM Feature importance .....	67
Figure 47: KNN Feature importance.....	67
Figure 48: Handling Missing Values .....	68



Figure 49: Data Saved for PowerBI Analytics.....	68
Figure 50: Numerical Features Correlation Matrix.....	71
Figure 51: The cumulative explained variance code .....	75
Figure 52: Cumulative Explained Variance for Different PCA settings.....	76
Figure 53: The PCA Application.....	76
Figure 54: Class distribution of Diabetic Readmission .....	77
Figure 55: Class distribution of Diabetic Readmission using SMOTE .....	77
Figure 56: Data Splitting .....	78
Figure 57: Logistic Regression Sigmoid Function .....	79
Figure 58: KNN Model.....	80
Figure 59: Flow diagram of GBM .....	81
Figure 60: Flow diagram of Random Forest .....	82
Figure 61: Support Vector Machine (SVM).....	83
Figure 62: Accuracy (ACC) Formula.....	85
Figure 63: Precision (PREC) Formula .....	85
Figure 64: Recall (REC) Formula .....	86
Figure 65: Confusion Matrix .....	86
Figure 66: Applied Steps .....	88
Figure 67: Data Pre-processing/Removing Redundant columns.....	89
Figure 68: Data Preprocessing/creating New Column.....	89
Figure 69: Modelling.....	90
Figure 70: Modelling with ML results imported .....	91
Figure 71: Dashboard Organisation .....	92
Figure 72: Confusion Matrix for Random Forest .....	95
Figure 73: Random Forest Feature Importance .....	96
Figure 74: Risk Stratification .....	97
Figure 75: LIME Explanations .....	98
Figure 76: Confusion Matrix for SVM .....	99
Figure 77: SVM Feature Importance .....	100
Figure 78: SVM Risk Stratification.....	101
Figure 79: LIME Explanations for SVM .....	102
Figure 80: ROC AUC.....	103
Figure 81: Comparison of Classification Report With PCA.....	104
Figure 82: Comparison of Classification Report Without PCA.....	105
Figure 83: Power BI Overview.....	106
Figure 84: Parameter Dashboard .....	107
Figure 85: Medication Dashboard .....	108
Figure 86: ML Results .....	109
Figure 87: ML Feature Importance .....	109

# **1. INTRODUCTION**

## **1.1 Background**

Hospital readmission for diabetic patients is a significant challenge in healthcare, affecting both patient outcomes and financial burdens. Diabetic patients face unique challenges due to the chronic and complex nature of their condition, which often requires continuous monitoring, extensive medication management, and frequent adjustments to their treatment plans. These challenges increase the likelihood of hospital readmissions, which not only underscore potential gaps in diabetes management and post-discharge care but also contribute significantly to rising healthcare costs. As highlighted by Calver et al. (2006) and Johansen et al. (1994), a substantial portion of hospital expenses is attributed to a small fraction of patients, particularly those with chronic conditions like diabetes, who often experience multiple hospital admissions for the same issue. The American Diabetes Association (2007) estimates that the cost of diabetes in the U.S. reached \$174 billion, with inpatient care accounting for the majority of these expenses.

Moreover, Kassin et al. (2012) and Stefan et al. (2013) emphasize that the 30-day readmission rate following an initial hospitalization has become a critical performance metric for hospitals, used by the Centers for Medicare and Medicaid Services (CMS) to evaluate patient care quality. High readmission rates are increasingly scrutinized as indicators of substandard patient care. Therefore, addressing this issue is crucial for improving patient outcomes, enhancing the quality of care, and reducing healthcare costs.

To tackle this challenge, the integration of Machine Learning (ML) and Business Intelligence (BI) tools offers a novel and powerful solution. As Negash (2004) and Rohloff (2011) suggest, BI involves the use of information and specialized analytical tools to support informed decision-making across various organizational settings. Mettler and Vimarlund (2009) highlight that a fundamental feature of BI is its ability to integrate data from multiple internal and external sources, creating a comprehensive information platform that significantly enhances decision-making processes. Tremblay et al. (2012) further note that BI is widely recognized for its substantial benefits to healthcare organizations, including optimizing patient care, improving clinical outcomes, streamlining operations, and supporting strategic decisions.

This project, titled "Predictive Analytics for Diabetic Patient Readmission Using Machine Learning and Power BI," leverages ML and BI tools to predict hospital readmissions among diabetic patients. The integration of Power BI dashboards enhances this approach by providing healthcare professionals with interactive

visualizations that display key metrics such as 'Average Time in Hospital,' 'Average Number of Medications per Age Group,' and 'Readmission Rate' segmented by race. These visualizations, supported by DAX measures, enable the analysis of patterns in patient insulin usage and hospitalization time across different demographic groups, aiding in the identification of trends and disparities that inform patient care and resource allocation.

The medication dashboard illustrates medication usage across different races, with an equal distribution observed. It also includes a risk stratification component derived from the top-performing ML model, classifying patients into risk categories based on predicted readmission risks. This analysis reveals disparities in treatment efficacy, particularly in Class 1 (<30 or readmitted with 30 days) patients, who show a higher rate of readmission.

Furthermore, the machine learning dashboard highlights the superior performance of the SVM model, which, even when evaluated by macro averages, outperforms other models in predicting readmission risks. The feature importance and LIME explanations focus on key factors such as diagnoses and medications, providing clear insights into the variables most influential in-patient outcomes. These insights empower healthcare providers to make data-driven decisions, optimize treatment plans, and allocate resources more effectively, ultimately improving patient care and healthcare delivery. By integrating these advanced analytical tools, the project aims to identify high-risk patients before readmission occurs, allowing for targeted interventions that lead to better patient outcomes and more efficient resource utilization.

## **1.2 Rationale / Justification of the Study**

The project addresses several key objectives to tackle the issue comprehensively:

- **Reducing high hospital readmission rates:** Firstly, it aims to reduce high hospital readmission rates by predicting which diabetic patients are at high risk of readmission, thereby enabling early interventions, personalized care plans, and proactive follow-up measures. This approach ensures that patients receive the necessary support and care to manage their condition effectively, preventing complications that lead to readmissions.
- **Lowering healthcare costs:** By reducing the frequency of readmissions, the project can significantly cut healthcare expenses, making the system more efficient by minimizing the need for repeated hospital stays and emergency interventions. Predicting and preventing readmissions can lead to better allocation of healthcare resources, ensuring that high-risk patients receive appropriate attention and support.

- **Improving diabetes management and post-discharge care:** The project aims to identify gaps in diabetes management and ensure continuity of care post-discharge, leading to enhanced patient outcomes. By using ML models to pinpoint specific factors contributing to readmissions, healthcare providers can address these gaps with improved management protocols, patient education, and follow-up care strategies.
- **Facilitating proactive interventions:** Facilitating proactive interventions is also a major focus. Utilizing predictive insights allows healthcare providers to support high-risk patients before complications arise. This proactive approach contrasts with reactive care, which often comes too late to prevent readmissions. By implementing enhanced patient education, regular follow-up appointments, and remote monitoring, healthcare providers can reduce the likelihood of readmissions by addressing potential issues early.

Overall, the integration of ML and BI tools in this project aims to empower healthcare professionals with actionable insights and predictive capabilities. This will enhance patient care, reduce hospital readmissions, and optimize healthcare resource utilization, ultimately leading to a more efficient and effective healthcare system for managing diabetes.

### **1.3 Aim**

To leverage machine learning and business intelligence to predict the likelihood of hospital readmissions among diabetic patients, enabling targeted interventions and improved resource utilization.

### **1.4 Objectives:**

#### **Machine Learning:**

- Develop ML models to accurately predict the likelihood of hospital readmissions for diabetic patients based on various clinical and demographic factors.
- Utilize ML algorithms to identify diabetic patients who are at high risk of readmission, allowing healthcare providers to prioritize interventions and allocate resources effectively.
- Generate personalized care plans for high-risk patients by leveraging ML insights to tailor treatment strategies, follow-up schedules, and post-discharge care protocols.
- Optimize the allocation of healthcare resources by using ML predictions to guide staffing levels, bed availability, and the provision of additional support services for high-risk patients.

## **1.5 Business Intelligence:**

- Create interactive dashboards and visualizations using BI tools to present readmission prediction results, trends, and insights in a clear and accessible manner for healthcare providers.
- Use BI analytics to identify patterns and trends in readmission rates, patient demographics, and clinical variables to uncover potential risk factors and opportunities for intervention.
- Implement BI dashboards to monitor the performance of predictive models, track readmission rates over time, and evaluate the effectiveness of interventions and care plans implemented based on ML predictions.
- Provide decision support tools through BI dashboards to help healthcare providers make informed decisions about resource allocation, patient management, and intervention strategies based on real-time data and predictive insights.
- Use BI analytics to identify areas for quality improvement in diabetes management and post-discharge care, enabling healthcare organizations to implement targeted interventions and protocols to enhance patient outcomes and reduce readmissions.

## **1.6 Research Questions**

- What factors contribute most significantly to high hospital readmission rates among diabetic patients?
- How can predictive analytics be used to identify diabetic patients at high risk of readmission?
- What impact does targeted intervention have on readmission rates of diabetic patients?
- How does the use of machine learning in predicting readmission risks affect healthcare costs?
- Which gaps in diabetes management and post-discharge care are most associated with high readmission rates?
- How can machine learning models improve the identification of high-risk diabetic patients for proactive intervention?
- What are the most effective strategies for resource allocation to reduce readmission rates among diabetic patients?
- How do predictive analytics and BI tools impact the quality of life and health outcomes for diabetic patients?

## **1.7 Thesis Structure**

This research work is organized into six chapters: Introduction, Literature Review, Methodology, Implementation, Experimental Results and Discussions, and Conclusion and Future Work.

### **Chapter One: Introduction:**

This chapter introduces the prevalence and impact of hospital readmissions among diabetic patients and explores the challenges in accurately predicting these readmissions. It outlines the research questions, objectives, significance of the study, and provides an overview of the research structure.

### **Chapter Two:**

Literature Review Chapter two reviews existing research on traditional methods for predicting hospital readmissions, machine learning techniques, and business intelligence tools used in healthcare. It also examines previous studies on diabetes management and post-discharge care, highlighting the necessity for predictive analytics and BI to reduce readmission rates.

### **Chapter Three:**

Methodology This chapter details the systematic and sequential procedures used in this study, including data collection, preprocessing, feature engineering, model development, and evaluation. It also explains the selection of specific ML algorithms and BI tools, and their integration into the study.

### **Chapter Four:**

Implementation Chapter four describes the implementation process, including the development and deployment of ML models and BI dashboards. It discusses how predictive models were integrated with BI tools to create interactive dashboards for healthcare providers, facilitating informed decision-making.

### **Chapter Five:**

Experimental Results and Discussions This chapter presents the experimental results, comparing the performance of different ML models using various evaluation metrics. It discusses the findings in relation to the research questions and objectives, emphasizing the effectiveness of predictive models and BI tools in identifying high-risk patients and reducing readmissions.

### **Chapter Six:**

Conclusion and Future Work The final chapter summarizes the entire study, its findings, and limitations. It discusses the implications of the research for healthcare practice and provides recommendations for future research, focusing on further

improving predictive models and BI tools to enhance patient outcomes and optimize healthcare resources.

## **2. LITERATURE REVIEW**

Hospital readmissions pose a significant challenge in healthcare, creating financial burdens and potentially hindering patient recovery. This issue is particularly pronounced among patients with diabetes mellitus (DM). According to Soh et al. (2020), adult diabetic patients account for 10% to 25% of all unplanned hospital readmissions within a 30-day timeframe, underscoring the urgent need for effective strategies to mitigate readmission rates in this patient population.

### **2.1 Defining Hospital Readmissions**

Hospital readmissions are typically characterized as unexpected returns to the hospital within a set timeframe, usually within 30 days after discharge. Chin et al. (2016) and Gerhardt et al. (2013) describe readmissions as a common phenomenon in healthcare, highlighting its prevalence among diabetic patients, particularly those initially admitted for diabetes-related conditions. While previous studies such as those by Chin et al. (2016) and Gerhardt et al. (2013) provide foundational definitions of hospital readmissions, they often overlook the complexity introduced by specific patient populations, like those with diabetes mellitus (DM). These studies tend to generalize readmission patterns without accounting for the unique challenges faced by diabetic patients, such as the need for ongoing medication management and the high incidence of comorbidities. By focusing specifically on diabetic patients, this study addresses these nuances, providing a more targeted analysis that is crucial for developing effective intervention strategies. Research conducted by Rubin et al. (2016) and Ostling et al. (2017) highlights that patients with diabetes mellitus (DM) experience higher readmission rates compared to individuals with other primary diagnoses.

Advanced machine learning models are increasingly used for predicting hospital readmissions due to their high accuracy, yet their lack of interpretability limits practical application. Gao et al. (2023) tackled this issue by creating an interpretable machine learning model designed to forecast hospital readmissions within 30 and 90 days. They employed a two-step Extracted Regression Tree method to achieve this; they first trained a black box model and then extracted interpretable decision trees, achieving similar predictive performance to neural networks while maintaining interpretability.

The extracted regression trees identified key readmission risk factors, such as factors like patient age, Charlson score, source of admission, number of previous visits, and length of hospital stay, which were consistent with clinical knowledge. This method bridges the gap between explainable models with moderate accuracy and non-



explainable models with high accuracy, enhancing trust and practical adoption in clinical settings.

The study validated the extracted trees against data from a big hospital in Southeast Asia, showing that explainable models can match the predictive power of black box models and provide clinically relevant insights. Critical factors identified include patient age, Charlson and van Walraven scores, admission source, previous visit frequency, and duration of hospital stay, along with secondary diagnoses and discharge locations, which hold significance in certain scenarios.

In summary, the study demonstrated that combining advanced machine learning with interpretability enhances the utility of predictive models in healthcare practice, addressing the historical trade-off between interpretability and accuracy. Future research should test this approach across diverse hospital settings and incorporate more clinical data to further validate and refine the model.

## **2.2 Challenges in Predicting Diabetic Patient Readmissions**

Despite the high prevalence of readmissions among diabetic patients, understanding the specific factors predicting these readmissions remains a considerable challenge. Research conducted by Rubin et al. (2016), Eby et al. (2015), and Collins et al. (2017) underscores considerable gaps in understanding, pointing out the intricate and multifaceted nature of factors influencing readmission risks. These studies emphasize that despite the identification of various patient characteristics linked to 30-day readmissions, creating reliable and precise prediction models remains a difficult task. Rubin et al. (2016) performed a retrospective analysis of 44,203 discharges involving diabetic patients, leading to the creation of the Diabetes Early Readmission Risk Indicator (DERRI). This model pinpointed ten critical predictors of 30-day readmission and effectively categorized patients into quintiles based on their readmission risk. Patients in the highest risk quintile had nearly a 40% chance of being readmitted, underscoring the model's potential effectiveness. The DERRI demonstrated satisfactory accuracy and reliability in both training and validation samples, highlighting its promise as a tool for predicting individual patient readmission risks.

Research by Rubin et al. (2016) and Collins et al. (2017) highlights significant gaps in understanding the predictors of readmission among diabetic patients, yet their models often rely on conventional statistical methods that may fail to capture the complex, non-linear relationships between variables. These traditional approaches can be limited by their assumption of linearity and may not adequately account for interactions between multiple factors. In contrast, this study leverages advanced machine learning

(ML) techniques, which are better suited to handle complex datasets and can reveal hidden patterns that are not apparent with traditional methods.

### **2.3 Factors and Predictors**

The 30-day readmission rate for diabetic patients generally falls between 10% and 21%. In Rubin et al.'s study, this rate was 20.4%, which is on the higher end of the range observed in urban populations. This suggests that urban settings might have higher readmission risks due to factors like healthcare infrastructure and population density. Common reasons for readmission included diabetes complications, heart failure, complications from procedures, chest discomfort, difficulty breathing, acute kidney injury, and infections of the urinary tract.

Several new predictors of 30-day readmission were identified, including recent hospital discharge within the 90 days preceding admission, pre-admission insulin use, and specific laboratory values like serum hematocrit, sodium, and creatinine levels. Notably, patients residing within 5 miles from the hospital had higher readmission rates compared to those living farther away. This contrasts with other studies, suggesting that local healthcare infrastructure differences could influence readmission patterns. For instance, in Boston, patients living farther from the study hospital might have access to other healthcare facilities, potentially affecting where they are readmitted.

The work of Rubin et al. (2016) provides valuable insights into the factors influencing 30-day readmissions, yet the study's reliance on a single dataset limits the generalizability of its findings. Moreover, the use of basic statistical models may not fully exploit the richness of the data. By integrating multiple datasets and applying machine learning models, this study builds on Rubin et al.'s findings by offering more robust and generalizable predictions. Additionally, while Rubin et al. focus on clinical factors, this study expands the analysis to include operational data, thus providing a more comprehensive understanding of the predictors of readmission.

### **2.4 Limitations and Considerations**

The DERRI model's C-statistic of 0.70 compares favourably with other readmission models for diabetic patients and broader populations, which typically have C-statistics ranging from 0.56 to 0.72. The DERRI's advantage is its reliance on easily accessible predictors at the time of admission, which supports the early identification of patients at high risk and enables prompt intervention. Nonetheless, the study has some limitations. Since it was conducted at a single urban academic medical centre, the results may not be applicable to other contexts. Additionally, the retrospective design of the data collection process restricts the inclusion of certain factors, such as hemoglobin A1c levels, diabetes type and duration, and social determinants of health,

which were not captured. The limitations of the DERRI model, such as its single-site focus and retrospective design, underscore the need for more versatile and prospective studies. While the DERRI model has proven useful, its application in diverse healthcare settings remains untested. This study addresses these limitations by employing a broader dataset and incorporating real-time data analysis, which enhances the model's adaptability and relevance across different healthcare environments.

## **2.5 Practical Applications and Future Research**

The practical application of the DERRI model is envisioned as a readmission risk prediction tool, similar to the American College of Cardiology/American Heart Association CV Risk Calculator, could be integrated into electronic medical record systems. Such a tool would allow for the real-time identification of patients at high risk, enabling healthcare providers to apply targeted interventions to lower readmission rates. Future studies should aim to validate the DERRI model across various settings and assess the impact of targeted interventions for high-risk patients. This may include evaluating multi-faceted discharge bundles that offer patient-centred education, coordination of care around the discharge period, and support following discharge.

## **2.6 Machine Learning In Hospital Readmission Prediction**

A notable study on hospital readmission utilizing machine learning was conducted by Michailidis et al. (2022), which analysed data from Sismanogleio Public Hospital in Komotini, Greece, spanning 2018 to 2019. The anonymized dataset, sourced from the hospital's information system, includes a range of variables. The study applies four machine learning models: support vector machines (SVM) with linear and radial basis function (RBF) kernels, as well as weighted and balanced random forests. Michailidis et al. address data imbalance by reporting a readmission rate of one in every 6.42 cases (1741 readmissions out of 11,172 total cases), with readmission defined as re-admission within 30 days of discharge.

The data is divided into training (90%) and testing (10%) subsets using stratified random sampling to preserve the readmission ratio. Model performance is assessed using sensitivity and AUC metrics. This research is distinctive for incorporating three types of real-world data which are; medical-clinical, administrative/demographic, and operational data, highlighting that the inclusion of operational data offers a valuable contribution beyond the typical focus on electronic medical records (EMR) or medical and demographic data.

The key findings reveal that the primary factors for predicting readmissions are the diagnosis codes recorded at both admission and discharge. Notably, the clinic's

occupancy rate, an operational metric, was identified as the third most crucial predictor, even ranking higher than patient age and length of stay. Additionally, while the number of doctors and nurses in the clinic does influence readmissions, its impact is less significant compared to the other variables. These findings suggest that operational factors are crucial in forecasting readmissions and should be considered by hospital management.

Comparative results of the four models show that SVM-linear and SVM-RBF models attain sensitivities of 0.59 and 0.60, indicating that they correctly identify about 59-60% of readmission cases. However, both models suffer from a high false-positive rate, with a precision of 0.31, indicating good sensitivity but a tendency for false alarms regarding readmissions. Overall, Michailidis et al.'s study underscores the importance of incorporating diverse data types, particularly operational data, in predicting hospital readmissions and highlights areas for further research and potential improvements in hospital management practices.

Adhiya et al. (2024) analysed readmission rates in New York, focusing on patients treated for skin conditions from July 2014 and June 2015. Their study used data from the western part of New York, Rochester, and included 22,388 patient records. The study found that readmission rates varied significantly among different service settings: ambulatory surgery had a rate of 0.93%, inpatient hospital services had a rate of 0.83%, and outpatient hospital services had a rate of 8.93%. Analysis included factors such as gender, age, type of claim, line of business, healthcare group subcategory, and monthly fluctuations. For example, men had a higher readmission rate (7%) compared to women (4%), and patients aged 21-40 had the highest readmission rate at 13%.

To predict 30-day readmissions, seven machine learning models were evaluated using a 5-fold cross-validation method, with 80% of data allocated for training and 20% for validation. The study employed the SMOTE (Synthetic Minority Over-sampling Technique) technique to address data imbalance, enhancing predictive accuracy. Metrics such as accuracy, F1 score, AUC, precision, and recall were used for model assessment. Random forests (RF) and extreme gradient boosting (XG) were the most effective, achieving accuracy and F1 scores of 0.85, and AUCs of 0.90 and 0.89, respectively. The study underscored the importance of predictors like the month of admission and the line of business in forecasting readmissions.

Lu and Uddin (2022) created a stacking-based machine learning model to predict 30-day hospital readmissions among diabetic patients. Their method tackled the issue of class imbalance using Random Under-Sampling and employed SelectFromModel for

feature selection. The stacked model they developed outperformed individual machine learning techniques such as CatBoost and XGBoost, achieving an accuracy of 68.63% and an AUC of 0.6736 on the test data. In contrast, CatBoost and XGBoost recorded accuracies of 61.53% and 61.49%, respectively. The study also emphasized the importance of model interpretability by using Local Interpretable Model-agnostic Explanations (LIME) to clarify individual predictions. Significant predictors of readmission identified included the number of inpatients, primary diagnosis, and discharge information. Lu and Uddin's work demonstrates the value of stacking models and interpretability in improving predictive accuracy and offering valuable insights for healthcare professionals.

Mohamed Alloghani et al. (2021) conducted a comprehensive study using machine learning techniques to identify predictors of hospital readmission among diabetic patients. Their research focused on a dataset comprising 100,000 cases, of which 78,363 were diabetic patients, revealing that over 47% experienced readmission. The study employed classifiers including Linear Discriminant Analysis (LDA), Random Forest, k-Nearest Neighbor (kNN), Naïve Bayes, J48, and Support Vector Machine (SVM) to analyse patterns influencing readmission. They found that factors such as being female, Caucasian, outpatient status, and receiving less intensive medical procedures and medications were associated with higher readmission rates. Notably, patients discharged without significant clinical improvements or proper insulin administration despite positive HbA1c results were more prone to readmission.

In terms of predictive performance, Naïve Bayes emerged with the greatest Area Under the Curve (AUC) at 0.640, followed closely by Random Forest with an AUC of 0.602. These results underscored the efficiency of these machine learning algorithms in discerning readmission risks among diabetic patients. The study highlighted the critical role of thorough medical assessments, accurate diagnoses, and appropriate medication regimens in mitigating readmission rates, particularly among vulnerable demographics such as women and Caucasians.

Alajmani and Elazhary (2019) in their study, addressed the critical issue of predicting hospital readmissions for diabetic patients using machine learning techniques. Their research aimed to enhance resource allocation and enhance the quality of care by identifying high-risk patients who might be readmitted. This study is notable for its comparative approach, assessing the effectiveness of five widely used machine learning algorithms: logistic regression (LR), multi-layer perceptron (MLP), Naïve Bayes (NB) classifier, decision tree (DT), and support vector machine (SVM).

The dataset used in the study comprised 3090 instances with 18 attributes for diabetic patients aged 30-50. Each algorithm was meticulously trained and tested. For logistic regression, the model was constructed using grid search to optimize accuracy and hyper-parameters. The support vector machine was trained with an RBF kernel, with grid search determining that the highest accuracy, 0.9246, was achieved with  $C=10$  and  $\gamma=0.3$ . The decision tree was developed using the 'gini' function to evaluate splits, with the best performance noted at a maximum depth of 15. The Naïve Bayesian classifier was implemented using Gaussian Naive Bayes, assuming a natural distribution of attributes. The multi-layer perceptron consisted of an 18-input network with 5 neurons in the buried layer, trained using stochastic gradient descent.

The study's results were evaluated using accuracy, recall, precision, and F1 score, under a 10-fold cross-validation scheme. The SVM appeared as the best-performing model, achieving the superior accuracy of 0.9522. The decision tree followed with an accuracy of 0.9251, and the multi-layer perceptron achieved an accuracy of 0.8358. The Nave Bayesian classification and logistic regression had accuracies of 0.6907 and 0.6865, respectively.

Alajmani and Elazhary concluded that the support vector machine significantly outperformed the other methods, highlighting its potential as the most effective technique for predicting diabetic patients hospital readmission. Conversely, the Naïve Bayesian classifier and logistic regression were identified as the least effective, underscoring the variability in performance among different machine learning algorithms for this specific application. These studies contribute to the growing body of literature on using ML to predict hospital readmissions, emphasizing the importance of incorporating diverse data types and addressing data imbalance for improved model performance.

The studies highlight the utility of machine learning in identifying key predictors and improving readmission forecasts, which is crucial for healthcare management and planning. Studies like Michailidis et al. (2022) demonstrate the potential of machine learning in predicting hospital readmissions but often suffer from a narrow focus on specific types of data, such as EMR or demographic information. Michailidis et al.'s inclusion of operational data marks an improvement, yet their models, such as SVM and Random Forest, still face challenges related to overfitting and interpretability.

This study builds upon these efforts by not only incorporating a broader range of data types but also comparing the performance of multiple models, including those with enhanced interpretability, thus offering a more nuanced approach to readmission prediction.

## 2.7 The Potential of Machine Learning

Machine learning offers an encouraging method to addressing the challenges of readmission prediction in diabetic patients. Jayatilake and Ganegoda (2021) emphasize that ML excels at analyzing vast and complex datasets, including medical records, reports, and imaging data. This capability allows ML algorithms to uncover subtle patterns and relationships within patient data that might be missed by traditional methods.

By analysing historical data on diabetic patients, including demographics, medical history, laboratory results, and details of the initial hospitalization, ML models can learn to identify characteristics and patterns associated with an increased risk of readmission. This information can then be utilised to develop predictive models that flag high-risk patients, enabling healthcare providers to intervene proactively.

Jayatilake and Ganegoda (2021) highlight the strength of ML in handling complex datasets but do not address the practical challenges of integrating these models into existing healthcare workflows. Their discussion of ML's potential lacks consideration of real-world implementation issues, such as model interpretability and user acceptance. This study goes a step further by applying interpretable machine learning models and validating them within a clinical setting, ensuring that the results are not only accurate but also actionable and easily understood by healthcare professionals.

## 2.8 Benefits of Machine Learning for Readmission Prediction

The potential benefits of using ML for readmission prediction in diabetic patients are numerous. Identifying high-risk patients early enables the implementation of focused interventions, such as:

- **Personalized discharge planning:** Developing tailored discharge plans that address specific risk factors identified by the ML model.
- **Enhanced patient education:** Providing targeted education to patients at high risk of readmission on medication adherence, self-care management, and recognizing signs of complications.
- **Improved care management:** Enhancing communication and teamwork among healthcare professionals involved in a patient's treatment plan.

By employing these interventions, healthcare providers have the potential to enhance patient outcomes and decrease the overall incidence of readmissions.

## 2.9 Challenges and Considerations in Machine Learning for Healthcare

While ML offers significant promise, implementing it in healthcare settings presents certain challenges. Habebh and Gohel (2021) highlight the importance of ensuring practical implementation and interpretation of ML models within real-world workflows.

Integrating these models smoothly with existing healthcare information systems and ensuring user adoption by healthcare professionals are crucial for successful implementation.

Additionally, the effectiveness and applicability of machine learning models are significantly influenced by the quality and volume of training data. Any biases in the data can result in biased algorithms, potentially hindering their effectiveness. Addressing these challenges requires robust data collection practices, careful model selection and training techniques, and ongoing monitoring to ensure model performance and mitigate bias. While Habehh and Gohel (2021) acknowledge the challenges of implementing ML in healthcare, their discussion remains largely theoretical. They do not provide concrete solutions for overcoming issues such as data bias and model integration. This study addresses these gaps by employing techniques like SMOTE to mitigate data imbalance and using Local Interpretable Model-agnostic Explanations (LIME) to enhance model transparency, ensuring that the ML models developed are both fair and practically applicable in a clinical context.

## **2.10 The Role of Business Intelligence**

ML is an effective tool for uncovering insights from healthcare data. However, to translate these insights into actionable knowledge for improved decision-making, business intelligence (BI) plays a critical role.

Wixom and Watson (2001) define BI systems as tools that facilitate the analysis of data from various sources. These systems enable users to explore trends, identify patterns, and make decisions based on data. In the context of readmission prediction in diabetic patients, BI tools can play a vital role in leveraging the power of ML models for patient care.

## **2.11 Successful Implementation of Business Intelligence**

Several factors contribute to successful BI implementation, as outlined by Wixom and Watson (2001) and further emphasized by other researchers (Seah et al., 2010; Ramamurthy et al., 2008; Isik et al., 2011). These factors include:

- **Strong management support:** Active leadership commitment from hospital administration is essential for providing resources and ensuring project success.
- **Dedicated project champion:** A knowledgeable and enthusiastic individual can spearhead the BI implementation and address challenges.
- **Sufficient resources:** Adequate financial and human resources are necessary to support the implementation process.



- **Effective user participation:** Involving end-users in the design and implementation phases ensures that the BI system meets their needs and is user-friendly.
- **Appropriate technical team skills:** A skilled technical team is crucial for developing, implementing, and maintaining the BI system.
- **High-quality source system data:** The accuracy and reliability of the BI system is based on the quality of the data being analysed.

Seah et al. (2010) emphasize the critical role of senior management support and leadership in the successful implementation of Business Intelligence (BI) systems. While their study provides valuable insights into the importance of organizational backing, it tends to generalize the factors necessary for BI success, overlooking the specific challenges faced in healthcare environments, such as integrating BI with existing Electronic Health Records (EHR) systems and ensuring compliance with healthcare regulations. This study builds on Seah et al.'s findings by contextualizing the need for leadership support within the healthcare sector, particularly in managing hospital readmissions among diabetic patients. Strong leadership is essential not only for resource allocation but also for fostering a culture that embraces data-driven decision-making, which is critical for the successful deployment of both BI and ML tools in predicting readmissions.

## **2.12 Business Intelligence and Organizational Factors**

Seah et al. (2010) stress the importance of strong support and leadership from senior management to achieve successful results. Ramamurthy, Sen, and Sinha (2008) use the diffusion of innovation (DOI) theory to describe Business Intelligence (BI) as a major IT infrastructure advancement. They contend that effective implementation depends on organizational elements such as firm management dedication, organizational size, and its ability to absorb new technologies. They also point out that the characteristics of the BI system, including its perceived benefits and simplicity, play a crucial role in successful adoption.

## **2.13 Key Capabilities for Business Intelligence Success**

Isik et al. (2011) suggest that the effectiveness of Business Intelligence (BI) implementations relies on several critical factors, including access to high-quality data, proper user permissions, and smooth integration with other systems. They also emphasize the need to thoroughly define and comprehend the decision-making environment, including the kinds of decisions to be supported and the operational context in which the BI system will function. Isik et al. (2011) identify several critical factors that contribute to the successful implementation of BI systems, including

access to high-quality data, proper user permissions, and smooth integration with other systems. Their research highlights the technical and organizational prerequisites for BI success but falls short of addressing the specific challenges posed by healthcare data, such as the sensitivity and variability of patient information. In this study, the emphasis is placed on ensuring that the BI systems used in conjunction with ML models can handle the unique demands of healthcare data, including maintaining patient privacy and dealing with the complexities of real-time data integration. By extending Isik et al.'s framework, this research not only considers the technical aspects of BI but also the ethical and legal challenges involved in its application within the healthcare sector.

#### **2.14 Integration of Machine Learning and Business Intelligence**

Integrating ML with BI can significantly enhance the ability of healthcare organizations to predict and manage hospital readmissions. By leveraging the predictive power of ML models and the analytical capabilities of BI systems, healthcare providers can gain comprehensive insights into patient data, enabling more informed decision-making.

The integration of ML with BI offers a powerful toolset for healthcare organizations, enabling them to predict and manage hospital readmissions more effectively. While previous studies, such as those by Ramamurthy, Sen, and Sinha (2008), have discussed the importance of organizational readiness and the characteristics of BI systems, they often do not explore how these systems can be optimized through the incorporation of ML techniques.

This study addresses this gap by demonstrating how ML can enhance BI systems' predictive capabilities, particularly in the context of readmission prediction for diabetic patients. By leveraging the strengths of both technologies, this research provides a more comprehensive approach to managing hospital readmissions, enabling healthcare providers to make more informed decisions based on a deeper understanding of patient data.

#### **2.15 Case Studies and Practical Applications**

Several studies have confirmed the effectiveness of combining ML and BI in healthcare settings. For example, Howell et al. (2009) and Png et al. (2018) discuss how conventional inpatient medical records can be leveraged to identify patients at risk of readmission and to create predictive models. These studies highlight the practical applications of ML and BI in improving patient outcomes and reducing readmissions.

Howell et al. (2009) conducted a study using routine inpatient data from Queensland public hospitals to develop a predictive algorithm for identifying patients at risk of

readmission following an emergency admission for specific conditions. Their dataset included demographic characteristics, co-morbidities, and prior utilization measures of 17,699 patients. They employed logistic regression to build the algorithm, which identified 16 predictors associated with readmission risk. The algorithm demonstrated a sensitivity of 44.7% at a 50% risk score threshold, with a corresponding false positive rate of 37.5%. The AUC curve was 0.65, indicating modest discriminatory power. Comparison with UK and US studies revealed similarities in predictive performance metrics. The algorithm's sensitivity was slightly lower than that reported in the UK (54.3%) and US (57.9%) studies, while the false positive rate was comparable. Likelihood ratios (LR+) were generally fair, and LR- were poor across risk score thresholds, suggesting limitations in sensitivity for detecting readmissions. The study highlighted demographic factors, co-morbidities, and prior hospital utilization as significant predictors, emphasizing the challenge of accurately forecasting readmission risk solely based on inpatient data.

From a business intelligence (BI) perspective, the discussion on cost implications in the study by Howell et al. (2009), is crucial as it explores the economic feasibility and value-for-money aspects of implementing predictive algorithms for readmission risk management in healthcare. BI in healthcare involves analyzing cost-effectiveness and return on investment (ROI) of interventions. The study referenced UK and US studies that built business cases to demonstrate potential cost savings through targeted case-management for patients at high risk of readmission (>70% predicted risk). This approach aligns with BI principles of using data-driven insights to justify and optimize resource allocation.

BI strategies in healthcare often differentiate between interventions that are strictly cost-saving (rare) versus those that provide value-for-money by improving health outcomes relative to costs incurred. The study suggests that while case-management might not be cost-saving in the traditional sense, it could be considered value-for-money given the high costs associated with readmissions and the potential to improve health outcomes and quality of life. BI frameworks help prioritize interventions based on cost-effectiveness and potential impact on health outcomes. In the context of readmission risk prediction, BI can inform healthcare decision-makers about allocating resources to interventions that provide the greatest cost-effectiveness, thus enhancing healthcare efficiency delivery and patient outcomes.

Insights from cost analysis in BI studies inform policy-makers and healthcare administrators about the financial implications of implementing predictive models. This includes understanding the upfront costs of intervention versus potential savings in

reduced readmissions, guiding decisions on investment in healthcare IT infrastructure and patient care management programs. In summary, the cost analysis discussed in Howell et al.'s study underscores the role of BI in healthcare by providing evidence-based insights into the economic viability and value-for-money of predictive modelling for readmission risk. This approach supports informed decision-making in healthcare policy, resource allocation, and patient care management.

Zhao et al (2023), in their study explore the predictors of unplanned 30-day hospital readmissions following surgery within an Asian surgical patient population. Conducted at a university-affiliated tertiary hospital in Singapore, this two-year retrospective cohort study involved 2744 patients aged 45 and above undergoing intermediate or high-risk non-cardiac surgeries. The study aimed to identify significant predictors of readmission to potentially mitigate healthcare costs and complications associated with such events.

The study found that 9.1% of patients experienced unplanned 30-day readmissions. Through multivariable logistic regression analysis, several significant predictors were identified: higher American Society of Anaesthesiologists (ASA) Classification grades (adjusted OR 1.51), obesity (adjusted OR 1.66), asthma (OR 1.70), renal disease (OR 2.03), malignancy (OR 1.68), chronic obstructive pulmonary disease (OR 2.46), cerebrovascular disease (OR 1.73), and anaemia (OR 1.45). These findings underscore the complex interplay of medical, healthcare system, and sociocultural factors in determining readmission risk. From a business intelligence perspective, the study's insights are critical for developing predictive models and risk management strategies tailored to the Asian surgical population. By identifying high-risk patients early, healthcare providers can allocate resources more efficiently, improve patient outcomes, and reduce the financial burden of unplanned readmissions. Implementing targeted interventions based on these predictors could lead to cost savings by minimizing the frequency of readmissions, thereby optimizing the use of hospital beds, medical staff, and other healthcare resources.

However, the authors also highlight the limitations of their retrospective study, noting that causality cannot be established. They advocate for future prospective studies to validate their findings and assess the effectiveness of interventions designed to lower readmission rates. This comprehensive approach could enhance the clinical relevance and applicability of predictive models, ultimately improving healthcare quality and efficiency in the context of a uniquely Asian surgical patient population.

Howell et al. (2009) and Png et al. (2018) provide practical examples of how conventional inpatient medical records can be utilized to identify patients at risk of

readmission. However, these studies typically rely on logistic regression models, which may not capture the complexity of patient data or the dynamic nature of healthcare environments. While Howell et al.'s study offers valuable insights into demographic and clinical predictors of readmission, the predictive power of their model is limited, as indicated by the modest AUC of 0.65. This study builds on these findings by employing more advanced ML algorithms, which are better suited to handle the intricacies of healthcare data and offer higher predictive accuracy. Moreover, by integrating ML with BI, this research enhances the practical application of predictive models, making them more actionable for healthcare providers.

## **2.16. Future Directions and Research Opportunities**

The integration of ML and BI in healthcare is still an evolving field with significant potential for future research and development. Key areas for further investigation include:

- **Enhancing ML algorithms:** Continuous improvement of ML algorithms to increase their accuracy, speed, and generalizability.
- **Addressing data quality issues:** Developing methods to ensure high-quality data collection and management practices.
- **Ethical and privacy considerations:** Addressing ethical concerns and ensuring patient privacy in the use of ML and BI technologies.
- **Interdisciplinary collaboration:** Encouraging collaboration between healthcare professionals, data scientists, and IT specialists to maximize the benefits of ML and BI.

Hospital readmissions, particularly among diabetic patients, remain a significant challenge in healthcare. Machine learning and business intelligence offer promising solutions to predict and manage readmissions effectively. By leveraging these advanced technologies, healthcare providers can improve patient outcomes, reduce readmission rates, and optimize resource allocation. However, successful implementation requires addressing various challenges, including data quality, ethical considerations, and practical integration into existing workflows. Continued research and collaboration are essential to fully realize the potential of ML and BI in transforming healthcare decision-making.

## **2.17 Identified Gaps**

Based on the previous works above, the gaps in the existing research on hospital readmissions, especially among diabetic patients, include several key areas that can be addressed in future studies and applications. These gaps are directly related to the integration of machine learning (ML) and business intelligence (BI) to improve

predictive models and real-time data analysis for healthcare providers. Here are the primary gaps identified:

### **2.18 Data Quality and Diversity:**

- While several studies utilize diverse datasets, there is still a need for more comprehensive and high-quality data, including social determinants of health, lifestyle factors, and detailed clinical information such as hemoglobin A1c levels and diabetes type and duration.
- Current datasets often lack granularity and completeness, which can limit the accuracy and generalizability of predictive models.

### **2.19 Model Performance and Comparison:**

- Although various ML models have been tested (such as SVM, random forests, XGBoost), there is a need for systematic comparisons and validations of these models across different settings and populations.
- The performance metrics (such as accuracy, AUC, precision) vary significantly among studies, indicating a need for standardization in evaluating ML models.

### **2.20 Operational Data Integration:**

- The literature highlights the importance of including operational data (such as clinic occupation rate, number of doctors and nurses) in predictive models. However, this integration is not yet widespread.
- More research is needed to explore how operational factors interact with clinical and demographic data to influence readmission risks.

### **2.21 Real-time Implementation:**

- There is a significant gap in the real-time application of predictive models and BI systems in clinical settings. Most studies focus on retrospective data analysis, with limited practical application in real-time patient management.
- Future research should focus on developing and testing real-time predictive tools integrated into electronic medical records (EMRs) to provide timely alerts and interventions.

### **2.22 Intervention Effectiveness:**

- While predictive models can identify high-risk patients, there is limited research on the effectiveness of subsequent interventions. Studies should evaluate the impact of targeted interventions (such as personalized discharge planning, patient education) on reducing readmission rates.
- There is also a need for randomized controlled trials to assess the effectiveness of different intervention strategies informed by predictive models.

### **2.23 Ethical and Privacy Concerns:**

- Implementing machine learning and business intelligence in healthcare brings up ethical and privacy concerns, especially regarding the security of patient data and the need for informed consent.
- Future research should address these concerns by developing frameworks for ethical data use and ensuring compliance with privacy regulations.

### **2.24 Interdisciplinary Collaboration:**

- Successful implementation of ML and BI systems requires collaboration between healthcare professionals, data scientists, and IT specialists. However, there is limited discussion in the literature on best practices for fostering such interdisciplinary collaboration.
- More case studies and practical guidelines are needed to illustrate how interdisciplinary teams can work together effectively.

### **2.25 Cost-effectiveness and Value-for-Money Analysis:**

- The literature often lacks comprehensive cost-benefit analyses of implementing ML and BI systems in healthcare settings.
- Future research should include detailed economic evaluations to justify investments in these technologies and demonstrate their value for money in improving patient outcomes and reducing readmissions. From a BI perspective, Howell et al.'s (2009) discussion on the cost implications of predictive algorithms highlights the economic feasibility and value-for-money aspects of managing readmission risks in healthcare. However, their study does not fully explore the potential of BI to optimize resource allocation based on predictive insights. This research expands on Howell et al.'s findings by incorporating BI strategies that differentiate between cost-saving interventions and those that provide value-for-money. By utilizing BI frameworks to prioritize interventions based on cost-effectiveness, this study aims to enhance healthcare delivery efficiency and patient outcomes. The integration of ML models with BI tools allows for more precise predictions, enabling healthcare administrators to make informed decisions about resource allocation that balance cost with the quality of care.

Zhao et al. (2023) highlight significant predictors of unplanned 30-day hospital readmissions following surgery in an Asian patient population, contributing valuable insights into the factors that influence readmissions. However, the retrospective nature of their study limits its ability to establish causality, and the generalizability of their findings may be constrained by the specific demographic and cultural context. This

study builds upon Zhao et al.'s work by applying prospective data collection methods and testing the predictive models across diverse patient populations. Additionally, by integrating ML with BI, this research aims to develop more generalized and robust models that can be applied across different healthcare settings, thus enhancing the practical utility of readmission predictions.



### **3. METHODOLOGY**

This chapter delivers an in-depth exploration of the classification techniques employed to achieve the objectives and tackle the research questions introduced in Chapter 1 of this project. The methodology implemented in this project spans across multiple phases of the data science lifecycle, ensuring a robust and systematic approach to the analysis.

#### **3.1 Suggested Model**

Figure 1 presents eight-step overview of the proposed model for diabetic patients' readmission classification, encapsulating distinct phases that outline the comprehensive tasks conducted throughout this study. Consequently, this chapter is systematically divided into eight sections, each offering an in-depth analysis of every phase. It begins with a detailed understanding of the problem, progresses through data collection, and concludes with the Power BI analytics and visualisation of the model results as Key Performance Indicators (KPIs). The design of this model is meticulously crafted to ensure precise classification of diabetic patients' readmission. Also, the model was deployed on a Windows 10 operating system, utilizing advanced hardware that includes a 12th Gen Intel(R) Core (TM) i7-12700 CPU running at 2.10 GHz, complemented by 64.0 GB of installed RAM. This 64-bit operating system is optimized for x64-based processors, offering robust performance and pen support for enhanced functionality.

Additionally, for the development of the algorithms, Jupyter Notebook version 6.5.4, with its web-based interactive computing capabilities, was selected. This IDE provided a highly flexible and intuitive environment for both coding and project management, substantially boosting the efficiency and effectiveness of the model development process. The code was crafted in Python 3.x, ensuring compatibility with all versions within the 3.x series, thereby offering a robust and scalable foundation for the project.

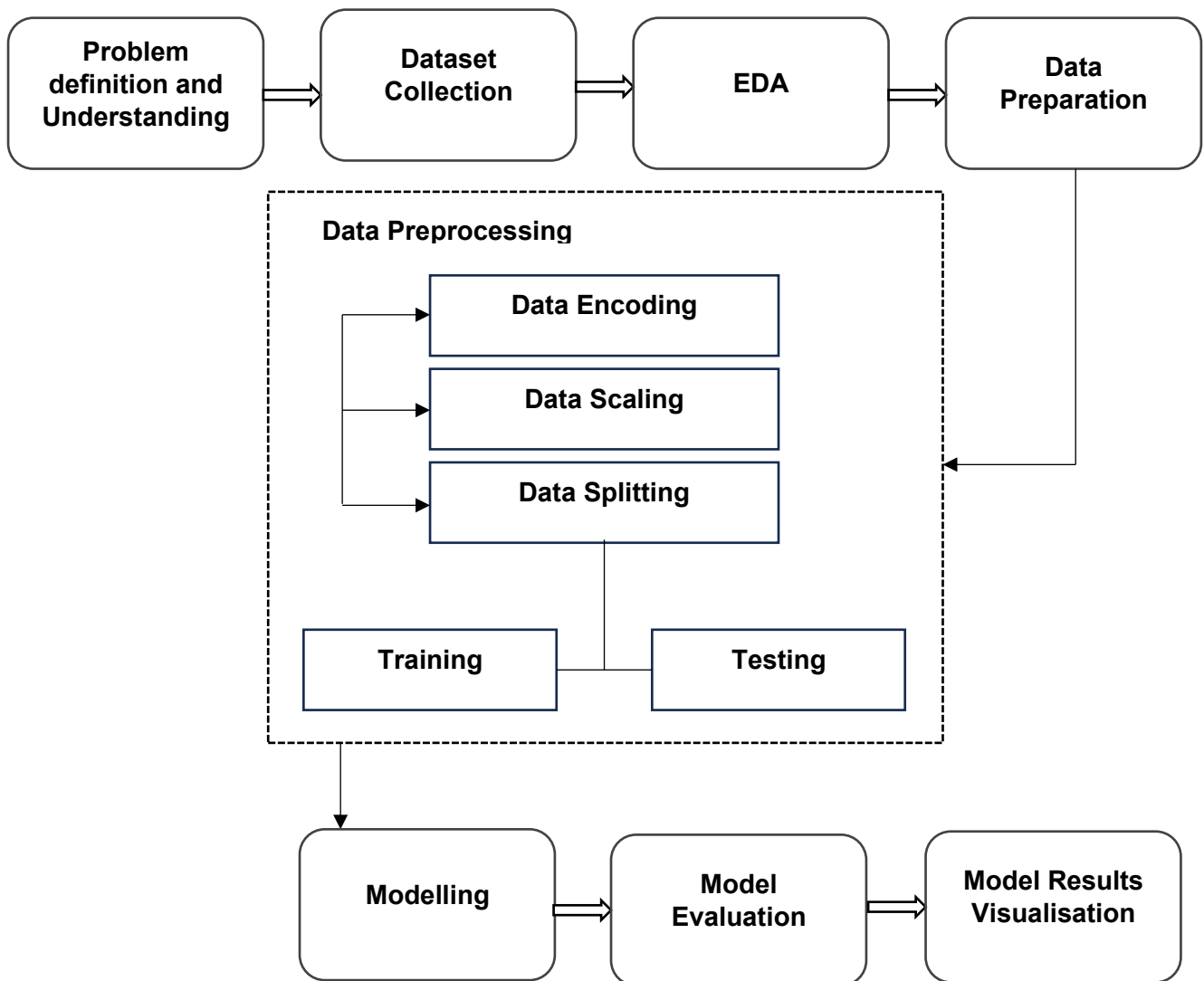


Figure 1: Suggested Model Diagram

### 3.2 Problem Definition and Understanding

In the context of predictive analytics for diabetic patient readmission using machine learning and Power BI (BI), the initial phase is pivotal to the project's overall success. This foundational step is critical as it sets the stage for all subsequent activities. During this phase, the project's goals and objectives are clearly defined, establishing a precise problem statement to be addressed through the application of ML algorithms and Power BI analytics. Do the current stewards of the organization's data perceive a Business Intelligence (BI) implementation as a threat? Those who are most familiar with the organization's data might feel that their value is tied to being the exclusive holders of this crucial information. To successfully implement BI, it is crucial to gain the cooperation of these key data custodians. They must be convinced to share their insights and clarify their data analyses. It's important to illustrate how a well-integrated BI system will enable them to transition from managing data to playing a pivotal role in guiding the organization. By focusing on strategic decision-making, they can leverage their expertise to influence the organization's direction, rather than merely

overseeing data management. (Data Analysis with Microsoft Power BI, Larson, December 6, 2019)

The clarity and depth of understanding achieved in this preliminary stage directly influence the effectiveness of the subsequent phases. For this project, the objective is to accurately predict diabetic patient readmission by leveraging a diverse set of features or independent variables. This process involves applying and evaluating various ML algorithms to identify the most effective approach for classifying readmissions, ensuring optimal results in the analysis.

### **3.3 Data Collection**

The dataset utilized in this research, focusing on the hospital readmission of diabetic patients, was obtained from Kaggle. It includes 101,767 patient records and 50 attributes, collected from 130 hospitals across the U.S. over a decade (1999-2008). It is pertinent to note, that only 1,500 rows and 50 attributes were used to efficiently process the data due to data complexity and processing time. Diabetic patients are at a higher risk of readmission compared to non-diabetic patients, making the reduction of readmission rates a critical factor in significantly lowering medical costs. The dataset's target variable is the readmission status, categorized as "No" for those who were not, "<30" for patients who were readmitted within 30 days and ">30" for those who were readmitted after 30 days. Following data importation, a comprehensive exploratory data analysis (EDA) was undertaken using a suite of Python visualization tools, including Seaborn, Matplotlib, and custom Python scripts. These tools were instrumental in analysing and visually representing the relationships between various features within the dataset. During the EDA phase, univariate, bivariate, and multivariate analyses were conducted to deeply explore the data's characteristics, with these insights further enhanced through detailed visualizations. The next critical phase involved preparing the data for modelling. This preparation included several key steps: encoding categorical variables, removing outliers, scaling the data, applying Principal Component Analysis (PCA), and addressing data imbalance through SMOTE, a crucial step to ensure the predictive accuracy of the models. The dataset was then divided into distinct training and testing subsets.

In the preprocessing phase, 865 attributes were initially employed after encoding and standardizing the data. Principal Component Analysis (PCA) was then applied, reducing the number of attributes to 4 with 85% covariance. However, the results with PCA were less effective compared to the original dataset without PCA. The table below provides detailed descriptions of each data feature.

Table 1: Features in the Diabetic Hospital Readmission Dataset

Serial	Column Name	Description	Data Type
1.	encounter_id	Encounter ID	int64
2.	patient_nbr	Patient number	int64
3.	race	Patient's race	object
4.	gender	Patient's gender	object
5.	age	Patient's age group	object
6.	weight	Patient's weight	object
7.	admission_type_id	Admission type ID	int64
8.	discharge_disposition_id	Discharge disposition ID	int64
9.	admission_source_id	Admission source ID	int64
10.	time_in_hospital	Length of stay in the hospital	int64
11.	payer_code	Patient's payer code	object
12.	medical_specialty	Admitting medical specialty	object
13.	num_lab_procedures	Number of laboratory procedures performed	int64
14.	num_procedures	Number of procedures performed	int64
15.	num_medications	Number of medications prescribed	int64
16.	number_outpatient	Number of outpatient visits in the past year	int64
17.	number_emergency	Number of emergency	int64

		department visits in the past year	
18.	number_inpatient	Number of inpatient admissions in the past year	int64
19.	diag_1	Primary diagnosis code	object
20.	diag_2	Secondary diagnosis code	object
21.	diag_3	Tertiary diagnosis code	object
22.	number_diagnoses	Number of diagnoses codes assigned	int64
23.	max_glu_serum	Maximum blood glucose serum level	object
24.	A1Cresult	Hemoglobin A1c test result	object
25.	metformin	Whether the patient is on metformin	object
26.	repaglinide	Whether the patient is on repaglinide	object
27.	nateglinide	Whether the patient is on nateglinide	object
28.	chlorpropamide	Whether the patient is on chlorpropamide	object
29.	glimepiride	Whether the patient is on glimepiride	object

30.	acetoexamide	Whether the patient is on acetoexamide	object
31.	glipizide	Whether the patient is on glipizide	object
32.	glyburide	Whether the patient is on glyburide	object
33.	tolbutamide	Whether the patient is on tolbutamide	object
34.	pioglitazone	Whether the patient is on pioglitazone	object
35.	rosiglitazone	Whether the patient is on rosiglitazone	object
36.	acarbose	Whether the patient is on acarbose	object
37.	miglitol	Whether the patient is on miglitol	object
38.	troglitazone	Whether the patient is on troglitazone	object
39.	tolazamide	Whether the patient is on tolazamide	object
40.	examide	Whether the patient is on examide	object

41.	citoglipton	Whether the patient is on citoglipton	object
42.	insulin	Whether the patient is on insulin	object
43.	glyburide-metformin	Whether the patient is on glyburide-metformin	object
44.	glipizide-metformin	Whether the patient is on glipizide-metformin	object
45.	glimepiride-pioglitazone	Whether the patient is on glimepiride-pioglitazone	object
46.	metformin-rosiglitazone	Whether the patient is on metformin-rosiglitazone	object
47.	metformin-pioglitazone	Whether the patient is on metformin-pioglitazone	object
48.	change	Change in medication	object
49.	diabetesMed	Whether the patient is on any diabetic medications	object
50.	readmitted	Whether the patient was readmitted within	object

		a month or more than a month	
--	--	---------------------------------	--

### 3.4 Exploratory Data Analysis (EDA)

Data Exploration, also known as EDA, is an necessary step in the data analysis procedure that begins after the dataset is acquired. During this stage, a comprehensive examination of the data is conducted to uncover its structural composition, statistical summaries, and intrinsic characteristics. This involves analysing the data's shape, types, and patterns, exploring relationships between variables, and identifying potential issues like missing values and outliers. The insights derived from this analysis are instrumental in shaping decisions and strategies for the succeeding phases of the proposed model, guaranteeing a more informed and effective method. The following EDA were carried out:

	encounter_id	patient_nbr	race	gender	age	weight	\
0	2278392	8222157	Caucasian	Female	[0-10)	?	
1	149190	55629189	Caucasian	Female	[10-20)	?	
2	64410	86047875	AfricanAmerican	Female	[20-30)	?	
3	500364	82442376	Caucasian	Male	[30-40)	?	
4	16680	42519267	Caucasian	Male	[40-50)	?	
	admission_type_id	discharge_disposition_id	admission_source_id	\			
0	6		25	1			
1	1		1	7			
2	1		1	7			
3	1		1	7			
4	1		1	7			
	time_in_hospital	...	citoglipton	insulin	glyburide-metformin	\	
0	1	...	No	No	No		
1	3	...	No	Up	No		
2	2	...	No	No	No		
3	2	...	No	Up	No		
4	1	...	No	Steady	No		
	glipizide-metformin	glimepiride-pioglitazone	metformin-rosiglitazone	\			
0	No	No	No	No			
1	No	No	No	No			
2	No	No	No	No			
3	No	No	No	No			
4	No	No	No	No			
	metformin-pioglitazone	change	diabetesMed	readmitted			
0	No	No	No	NO			
1	No	Ch	Yes	>30			
2	No	No	Yes	NO			
3	No	Ch	Yes	NO			
4	No	Ch	Yes	NO			
[5 rows x 50 columns]							
	encounter_id	patient_nbr	admission_type_id	\			
count	1.499000e+03	1.499000e+03	1499.000000				
mean	5.781669e+06	2.666213e+07	3.635757				
std	2.728809e+06	3.651056e+07	2.263533				
min	1.252200e+04	1.152000e+03	1.000000				
25%	3.577698e+06	1.450206e+06	1.000000				
50%	5.722026e+06	5.042762e+06	2.000000				

Figure 2: Checking First Few Rows

#### 3.4.1 Univariate Analysis

According to Camizuli and Carranza (2018), Exploratory Data Analysis (EDA) is a crucial technique that employs descriptive statistics and visual tools to gain an in-depth understanding of data. EDA is instrumental in uncovering trends, identifying outliers, and validating assumptions, serving as a foundational step before applying more complex statistical methods. This approach provides a comprehensive initial assessment, laying the groundwork for subsequent, more detailed analyses.



Visualizing data with tools like histograms, bar charts, and box plots is a standard approach for understanding univariate distributions. An in-depth analysis was carried out on the distribution patterns of each numerical and categorical feature within the dataset. The observations drawn from these visualizations provide valuable insights into the characteristics and distribution of the data are outlined below:

- Figure 3 shows that the dataset consists of 1,499 instances categorized into 10 unique age groups. Also, the most frequent age group is [70-80), with 356 instances, indicating that this age group is the most prevalent in the dataset.
- Figure 4 indicates that the gender distribution is relatively balanced, with 794 females (53%) and 705 males (47%). This near-equal representation suggests that gender is unlikely to introduce significant bias in the dataset.
- Figure 5 reveals that the majority of the individuals in the dataset are Caucasian (71.5%), followed by African American (25%). Other racial categories, including "Other," "Hispanic," and "Asian," are underrepresented, with counts of 24, 18, and 11, respectively. This imbalance may affect the generalizability of the analysis across different racial groups.
- Figure 6 implies that the readmission status is divided into three categories: "NO," ">30," and "<30".
- 729 instances (48.6%) are not readmitted, while 614 instances (41%) are readmitted after more than 30 days.
- A smaller portion, 156 instances (10.4%), is readmitted within 30 days.
- This distribution shows a significant portion of patients experiencing readmission, indicating that it is a common occurrence.

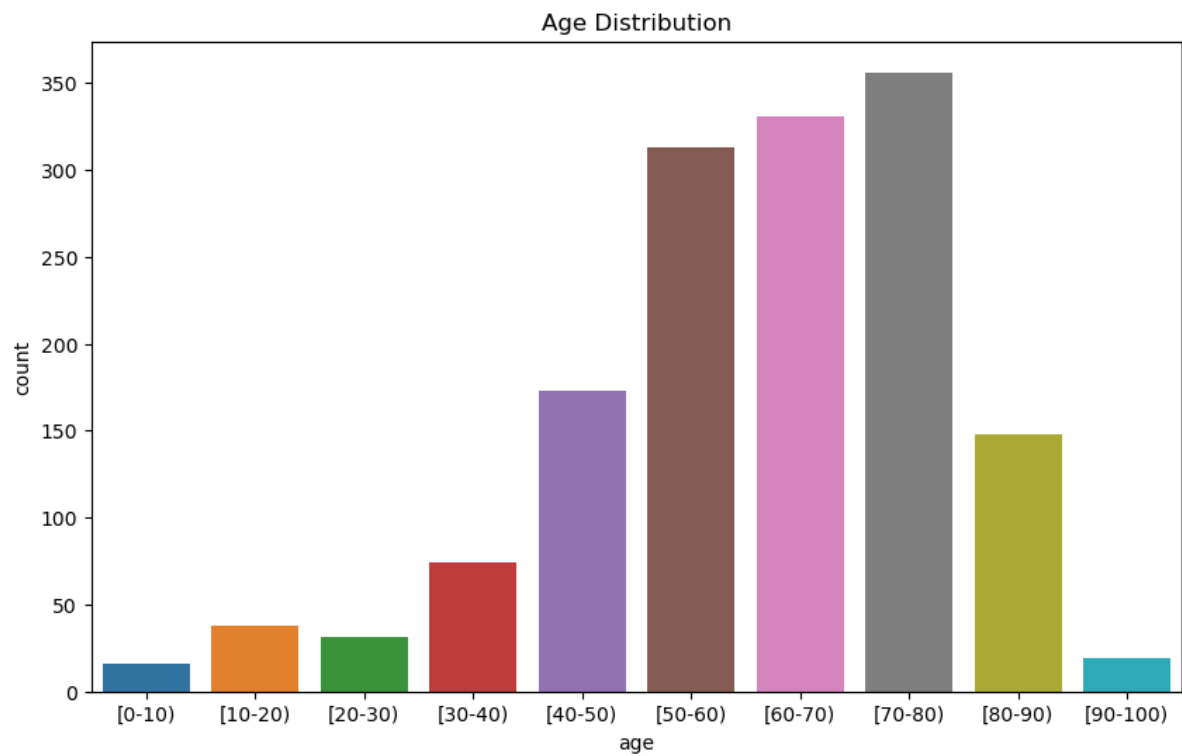


Figure 3: Age Distribution

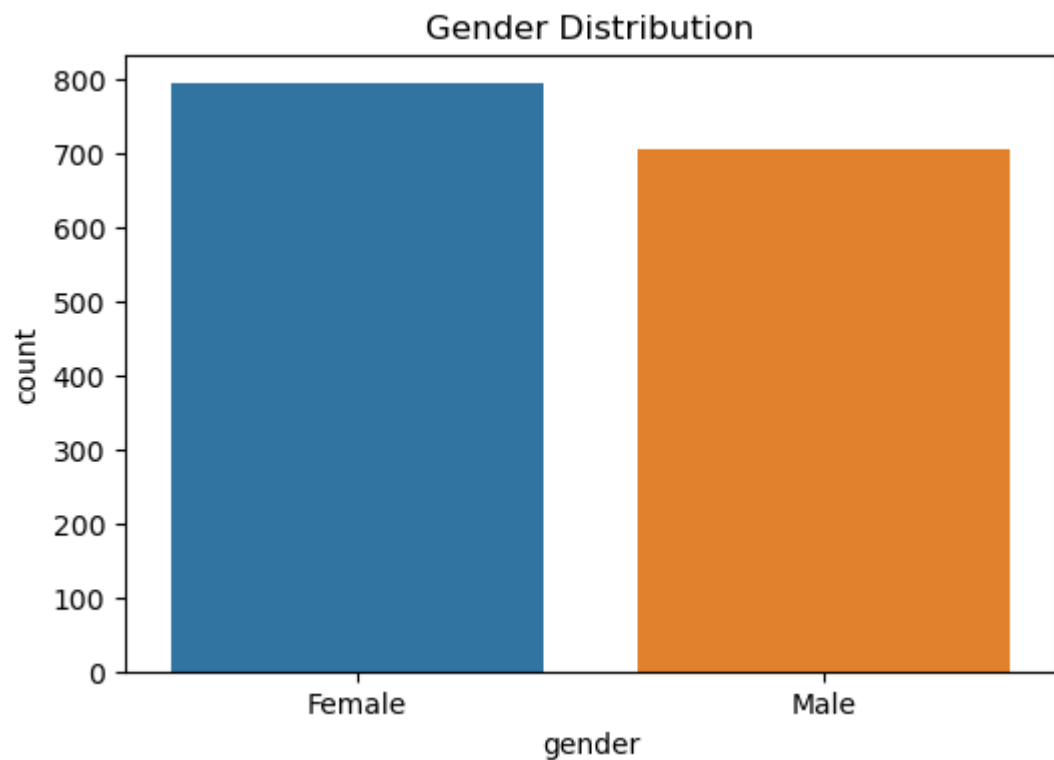


Figure 4: Gender Distribution

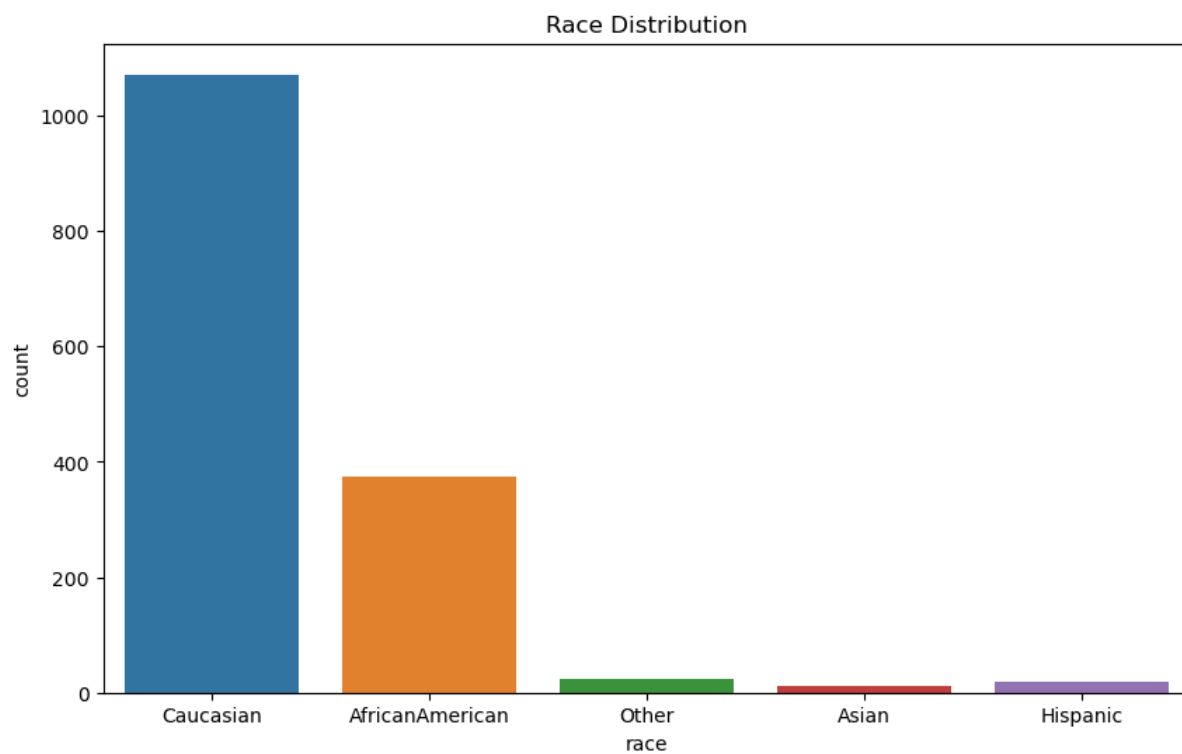


Figure 5: Race Distribution

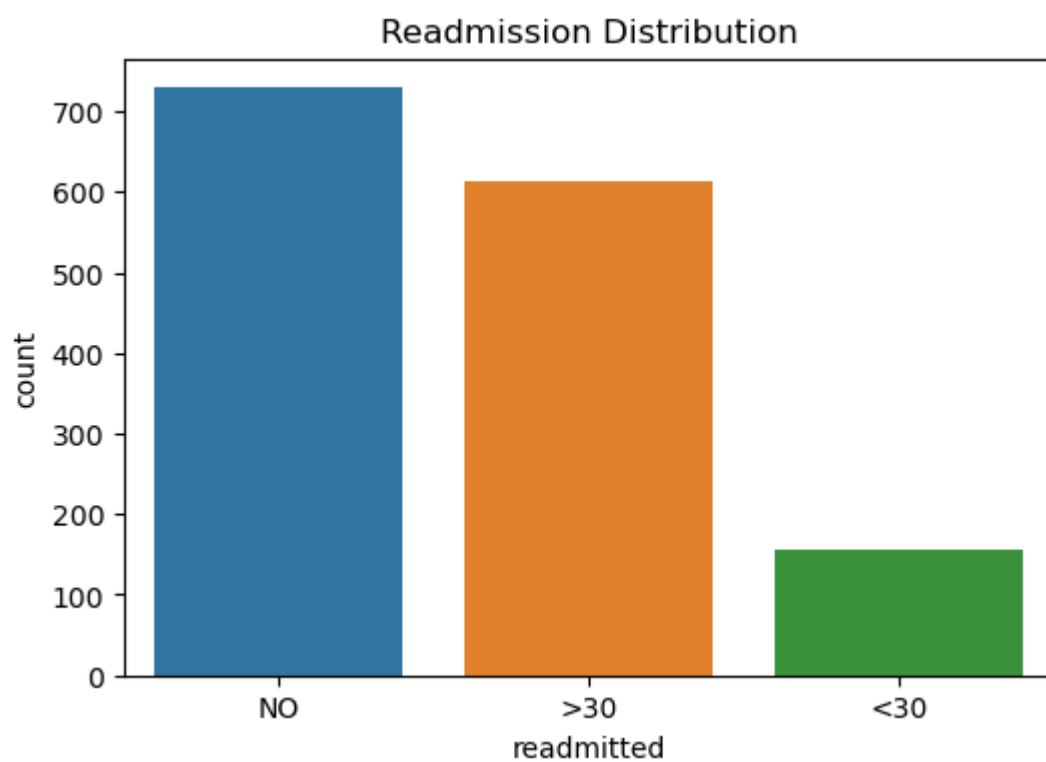


Figure 6: Readmitted Distribution

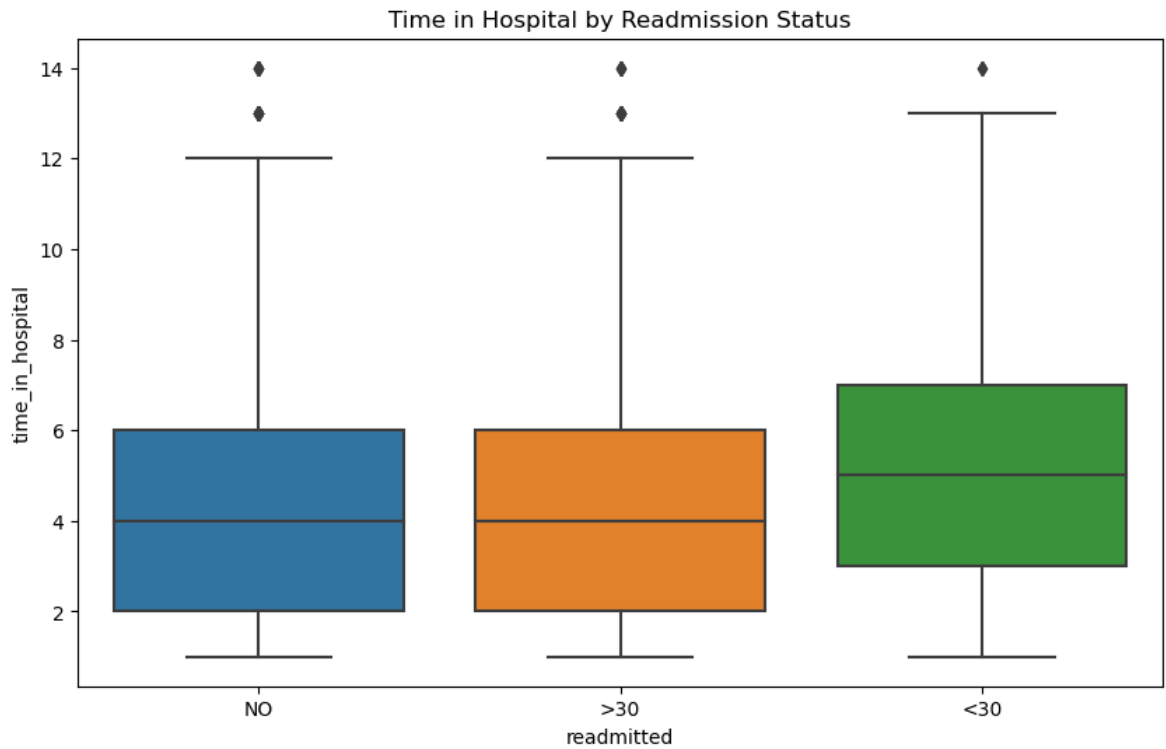


Figure 7: Time in Hospital by Readmission Status

### 3.4.2 Number of Lab Procedures:

- The number of lab procedures performed ranges from 1 to 105, with an average of approximately 50 procedures per patient.
- The standard deviation is 16.64, indicating moderate variability around the mean.
- The median (50%) number of lab procedures is 49, with the interquartile range (IQR) spanning from 39 to 61.
- The wide range and variability suggest that the number of lab procedures could be a significant factor in patient outcomes and may vary considerably among different patients.

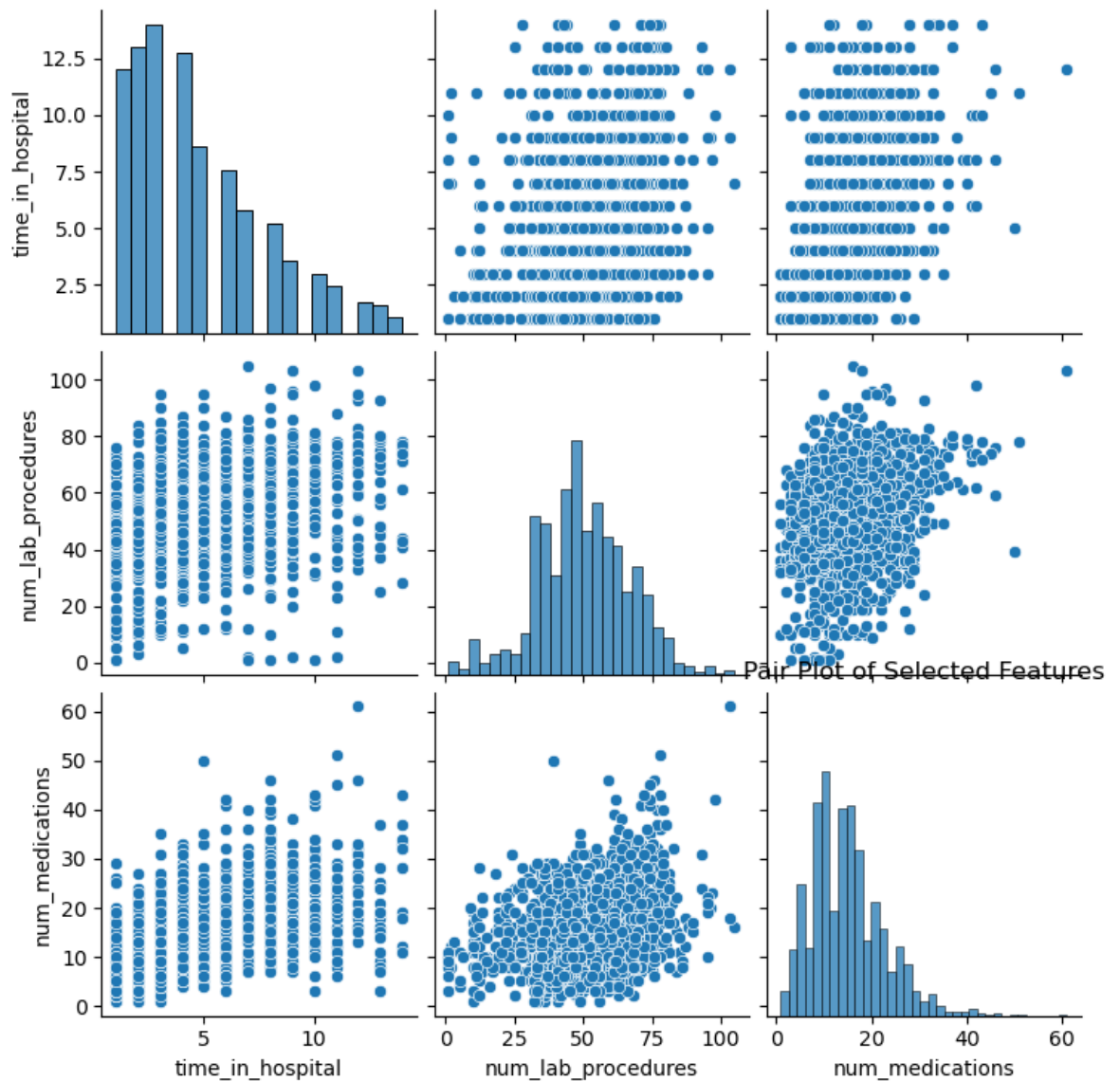


Figure 8: Number of Lab Procedures

#### 4.4.3 Correlation matrix

The initial correlation matrix analysis reveals key relationships among numerical features in the dataset. A strong positive correlation is observed between `num_medications` and `time_in_hospital` (0.52), indicating that longer hospital stays are associated with increased medication use. Similarly, there is a notable positive correlation between `num_procedures` and `num_medications` (0.46), suggesting that more procedures correlate with a higher number of medications. The correlation between `number_diagnoses` and `num_medications` (0.36) further supports the trend that more diagnoses lead to more medication.

Conversely, significant negative correlations are present, such as between `admission_type_id` and `discharge_disposition_id` (-0.36), implying an inverse relationship between admission types and discharge outcomes. The analysis also

highlights weak correlations for identifier features like encounter\_id and patient\_nbr, and minimal links between number\_outpatient and number\_emergency (0.11).

A notable issue is the strong correlation of 0.81 between admission\_type\_id and discharge\_disposition\_id, which may indicate redundancy and could impact model performance. Overall, the correlation matrix provides a crucial understanding of feature interactions, guiding effective feature selection and aiding in the development of more accurate models for analyzing hospital readmissions.

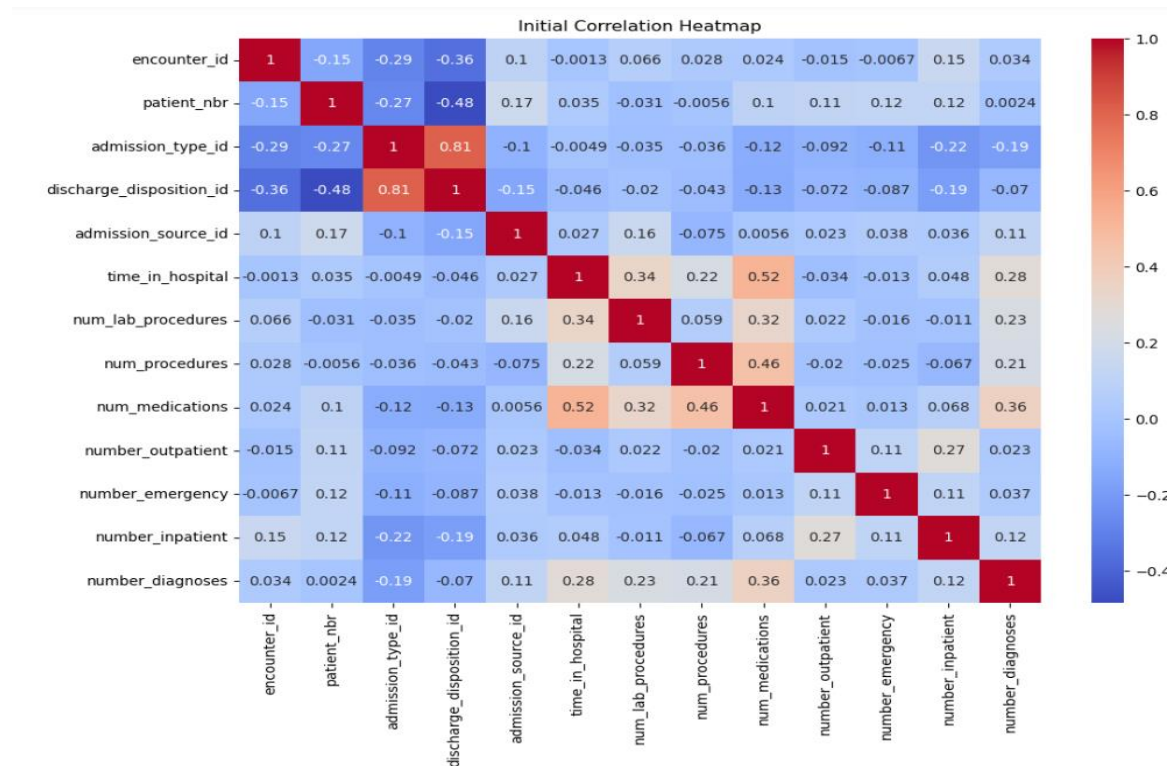


Figure 9: Correlation Matrix

**3.4.4 Distribution of num\_lab\_procedures based on Readmission Status:** The histogram provides a visual representation of the distribution of num\_lab\_procedures across different readmission statuses. The data is segmented by readmission status into three categories: Not Readmitted (Class 0), Readmitted within 30 days (Class 1), and Readmitted after 30 days (Class 2). The histogram uses a stacked bar approach with kernel density estimation (KDE) to illustrate the density of lab procedures performed.

From the histogram, several trends are apparent:

- **Readmitted Patients:** Patients who were readmitted within 30 days (Class 1) tend to have a higher density of lab procedures compared to those who were not readmitted (Class 0). This suggests that patients who are readmitted within a short period might undergo more frequent lab tests, possibly due to more severe or complex health conditions.

- **Not Readmitted Patients:** The distribution for patients who were not readmitted (Class 0) shows a broader spread of num\_lab\_procedures with a peak at lower values. This indicates that a significant portion of non-readmitted patients underwent fewer lab procedures, reflecting potentially less severe or less complex cases.
- **Readmitted After 30 Days:** Patients readmitted after 30 days (Class 2) also show a noticeable distribution of lab procedures, but the peak is not as pronounced as that for Class 1. This suggests that while there is a higher number of procedures for this group compared to non-readmitted patients, it is less concentrated compared to the readmission within 30 days group.

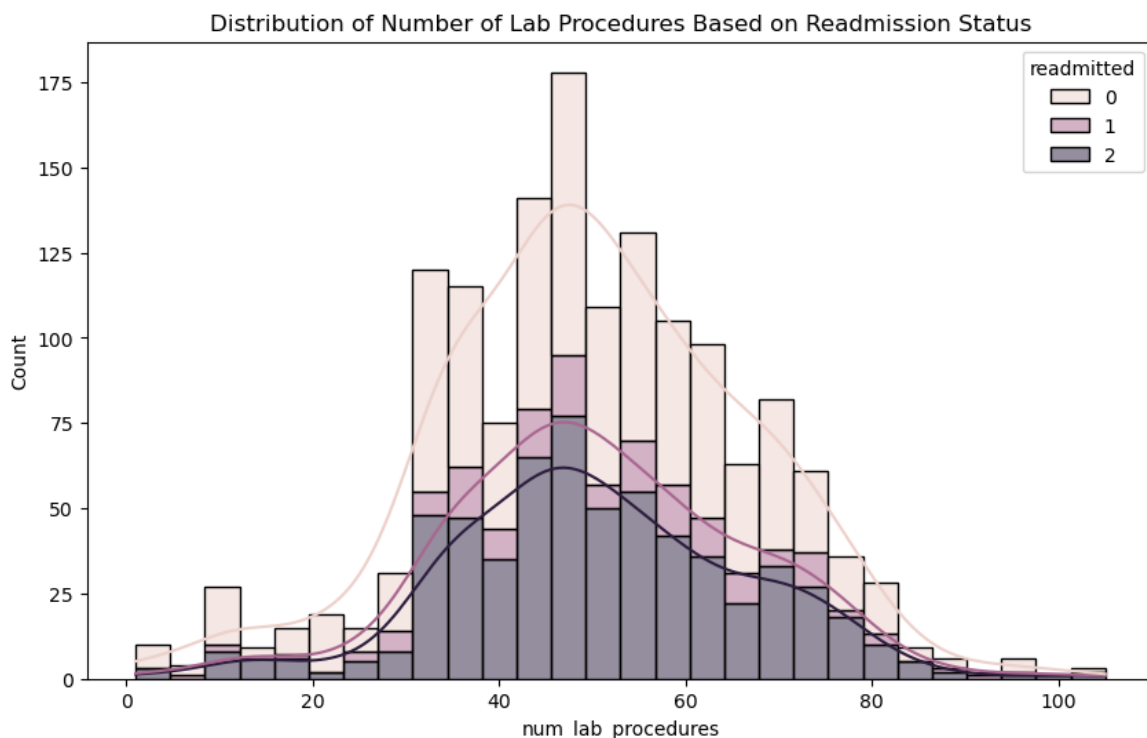


Figure 10: Distribution of Number of Lab Procedures Based on Readmission Status

**3.4.5 Distribution of Discharge Dispositions by Readmission Status:** The distribution of discharge dispositions across different readmission statuses highlights various levels of association with readmission risk. Discharge Disposition IDs with higher counts, such as 1, 25, and 6, are particularly relevant for understanding readmission patterns, while those with minimal representation may require further scrutiny or may have limited impact.

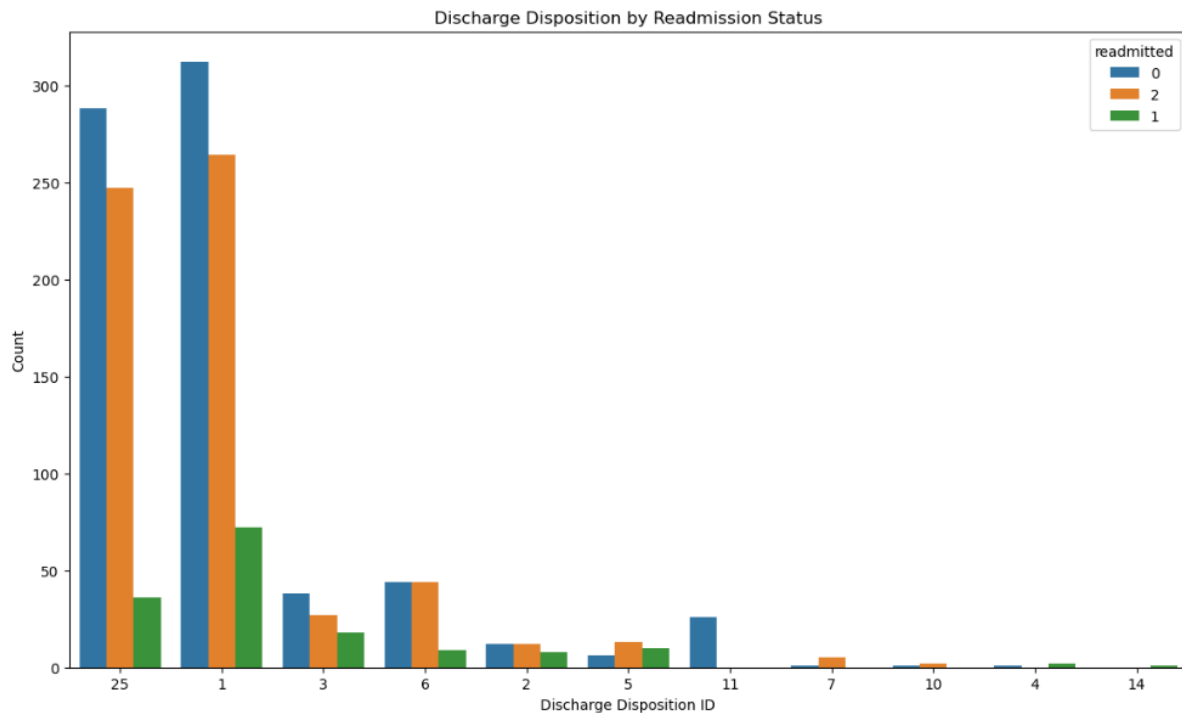
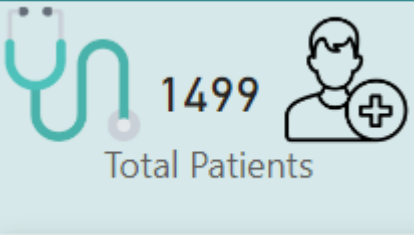


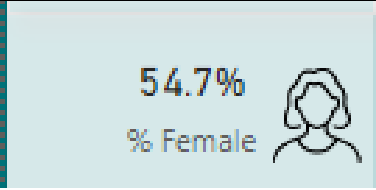
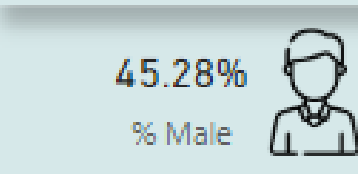
Figure 11: Discharge Disposition by Readmission Status

### 3.5 Key Performance Indicators

In Power BI, Key Performance Indicators (KPIs) are metrics used to evaluate the success of an organization or a specific activity in achieving objectives. KPIs provide a visual representation of performance relative to set targets, helping users track progress, identify trends, and make data-driven decisions. In the context of this project, KPIs were essential for monitoring key aspects of diabetic hospital readmissions, allowing for the analysis of trends and the effectiveness of strategies implemented to reduce readmission rates. The Key Performance Indicators (KPIs) utilized in this study, along with the DAX code implemented were to generate various results based on these KPIs, will be detailed.

<b>KPI 1: Total Patients</b>	<b>BI Question: What is the total number of the patients.</b>
 <p><b>Figure 12: Total Patients card visual</b></p>	<p>DAX Formula: Total Patients =  COUNTROWS('Cleaned_Diabetes_Data_Without_PCA')</p> <p>Visual: Card</p>
<b>KPI 2: Percentage of time in hospital</b>	<b>BI Question: What is the percentage of time in hospital for female patients?</b>





 <p><b>Figure 13: Female percentage time in hospital card visual</b></p>	<p>DAX Formula: Percentage Time in Hospital (Female) =</p> <pre> DIVIDE(     [Time in Hospital (Female)],     [Total Time in Hospital],     0 ) Visual: Card </pre>
<p><b>KPI 3: Percentage of time in hospital for male Card Visual</b></p>	<p><b>BI Question: What is the percentage of time in hospital for male patients?</b></p>
 <p><b>Figure 14: Male percentage time in hospital card visual</b></p>	<p>DAX Formula: Percentage Time in Hospital (Male) =</p> <pre> DIVIDE(     [Time in Hospital (Male)],     [Total Time in Hospital],     0 ) Visual: Card </pre>
<p><b>KPI 4: Total Hospital Readmission Card Visual</b></p>	<p><b>BI Question: What is the What is the total hospital Readmissions?</b></p>

<div data-bbox="199 235 279 280">770</div> <div data-bbox="95 309 386 342">Total Readmissions</div> <div data-bbox="76 409 458 481"> <b>Figure 15: Total hospital Readmission card visual</b> </div>	<p>DAX Formula: Total Readmissions =</p> <pre> IF(   ISBLANK(     CALCULATE(       COUNTROWS('Cleaned_Diabetes_Data_Without_PCA'),       'Cleaned_Diabetes_Data_Without_PCA'[readmitted] = "Yes"     )   ),   0,   CALCULATE(     COUNTROWS('Cleaned_Diabetes_Data_Without_PCA'),     'Cleaned_Diabetes_Data_Without_PCA'[readmitted] = "Yes"   ) ) </pre> <p>Visual: Card</p>
<b>KPI 5: Hospital readmission</b>	<b>BI Question: What is the possibility of hospital readmission?</b>
<div data-bbox="234 1406 367 1444">51.37%</div> <div data-bbox="140 1469 464 1503">Readmission Probability</div> <div data-bbox="76 1563 443 1635"> <b>Figure 16: Readmission Probability card visual</b> </div>	<p>DAX Formula: Readmission Rate = DIVIDE([Total Readmissions], [Total Patients])</p> <p>Visual: Card</p>
<b>KPI 6:Time in hospital for male</b>	<b>BI Question: What is the total time in hospital for male?</b>

<div data-bbox="78 163 443 360"><div data-bbox="220 219 311 264">3182</div><div data-bbox="110 284 421 320">Time in Hospital (Male)</div></div> <div data-bbox="78 378 488 450"><p>Figure 17: Time in hospital for male card visual</p></div>	<p>DAX Formula: Time in Hospital (Male) =</p> <pre>IF(   ISBLANK(     CALCULATE(       SUM('Cleaned_Diabetes_Data_Without_PCA'[time_in_hospital]),       'Cleaned_Diabetes_Data_Without_PCA'[gender] = "Male"     )   ),   0,   CALCULATE(     SUM('Cleaned_Diabetes_Data_Without_PCA'[time_in_hospital]),     'Cleaned_Diabetes_Data_Without_PCA'[gender] = "Male"   ) )</pre> <p>Visual: Card</p>
<p>KPI 6:Time in hospital for female card visual</p>	<p>BI Question: What is the total time in hospital for female?</p>

<div data-bbox="78 163 486 342"><div data-bbox="229 219 320 259">3845</div><div data-bbox="106 277 448 313">Time in Hospital (Female)</div></div> <div data-bbox="78 360 486 432"><p>Figure 18: Time in hospital for female card visual</p></div>	<p>DAX Formula: Time in Hospital (Female) =</p> <pre>IF(     ISBLANK(         CALCULATE(             SUM('Cleaned_Diabetes_Data_Without_PCA'[time_in_hospital]),             'Cleaned_Diabetes_Data_Without_PCA'[gender] = "Female"         )     ),     0,     CALCULATE(         SUM('Cleaned_Diabetes_Data_Without_PCA'[time_in_hospital]),         'Cleaned_Diabetes_Data_Without_PCA'[gender] = "Female"     ) )</pre> <p>Visual: Card</p>
<p>KPI 6:Time in hospital for male card visual</p>	<p>BI Question: What is the time in hospital per Gender?</p>

 <p><b>Figure 19: Average time in hospital per gender card visual</b></p>	<p>DAX Formula: Average Time in Hospital per Gender =  <code>AVERAGE('Cleaned_Diabetes_Data_Without_PCA'[time_in_hospital])</code></p> <p>Visual: Card</p>
<p><b>KPI 7: Total Time in Hospital</b></p>	<p><b>BI Question: What is the total time in hospital?</b></p>
 <p><b>Figure 20: Total time in hospital card visual</b></p>	<p>DAX Formula: Total Time in Hospital =  <code>SUM('Cleaned_Diabetes_Data_Without_PCA'[time_in_hospital])</code></p> <p>Visual: Card</p>

The decomposition tree KP below is crucial for both our business intelligence and machine learning objectives by providing a detailed hierarchical breakdown of hospital readmissions for diabetic patients, starting from total readmissions and analyzing them by race, age group, admission type, discharge disposition, and gender. This tool supports our ML objectives by helping validate and refine predictive models, identifying high-risk patient groups, and tailoring personalized care plans based on the specific factors that most influence readmissions. In the context of BI, the decomposition tree enhances our ability to create interactive dashboards that visually present key trends and insights, identify patterns in readmission data, and monitor the performance of predictive models, all of which are essential for informed decision-making and resource allocation in healthcare. Additionally, it revealed that the Caucasian race have the highest number in all categories while the Asian race have the lowest count of all categories.

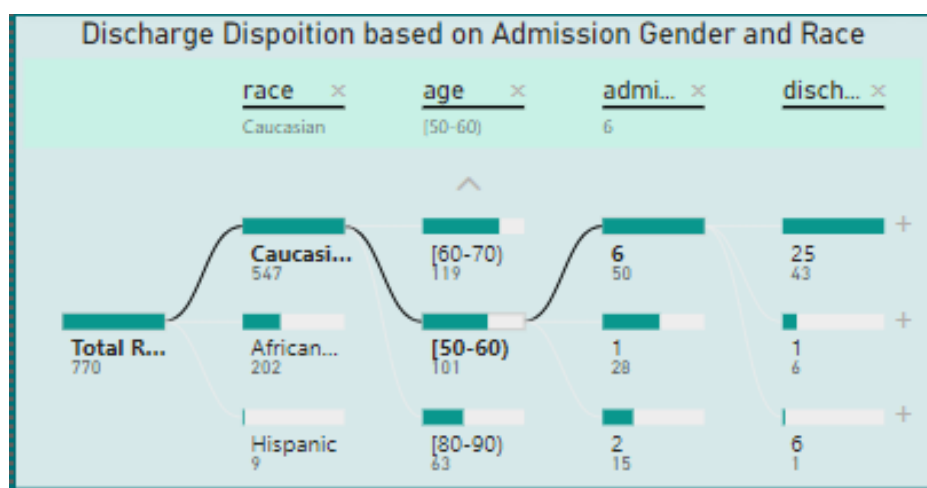


Figure 21: Discharge Disposition

The Diabetic Patients Based on Medication KPI below, indicates that 80.45% of patients are on diabetes medication, is highly relevant to both our business intelligence (BI) and machine learning (ML) objectives. From a BI perspective, this KPI helps in identifying the prevalence of medication use among diabetic patients, which is critical for understanding how medication adherence impacts hospital readmission rates. This insight allows healthcare providers to explore correlations between medication use and patient outcomes, enabling targeted interventions to improve patient adherence and reduce readmissions.

In terms of our ML objectives, this KPI is essential for building accurate predictive models that take into account medication use as a key factor in determining the likelihood of readmission. By incorporating medication data into our ML algorithms, we can better identify high-risk patients and generate personalized care plans that consider their medication needs. Additionally, this KPI supports the optimization of healthcare resources by guiding decisions on medication management and post-discharge care, ensuring that patients receive the appropriate follow-up based on their medication profiles. Thus, this KPI directly connects to our goal of leveraging ML insights and BI analytics to enhance patient outcomes and reduce hospital readmissions.

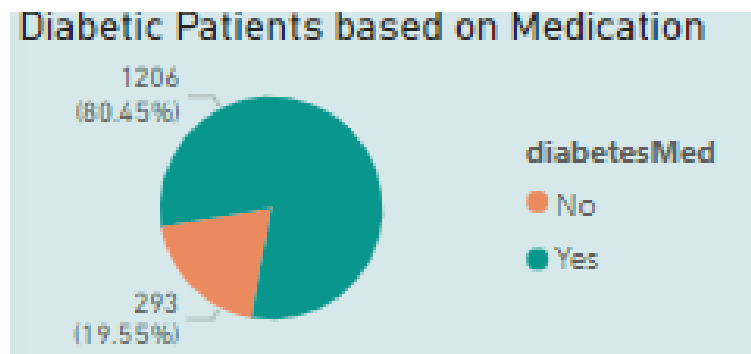


Figure 22: Diabetic Patients Based on Medication

The "Total Readmission by Age" KPI, which identifies the 70 to 80 years age group as having the highest hospital readmissions, is essential for achieving our business intelligence and machine learning objectives. This KPI helps healthcare providers analyze patterns in readmissions among older patients, enabling targeted interventions and better resource allocation. In machine learning, it enhances the accuracy of predictive models, allowing for personalized care plans that reduce readmissions in this high-risk age group.

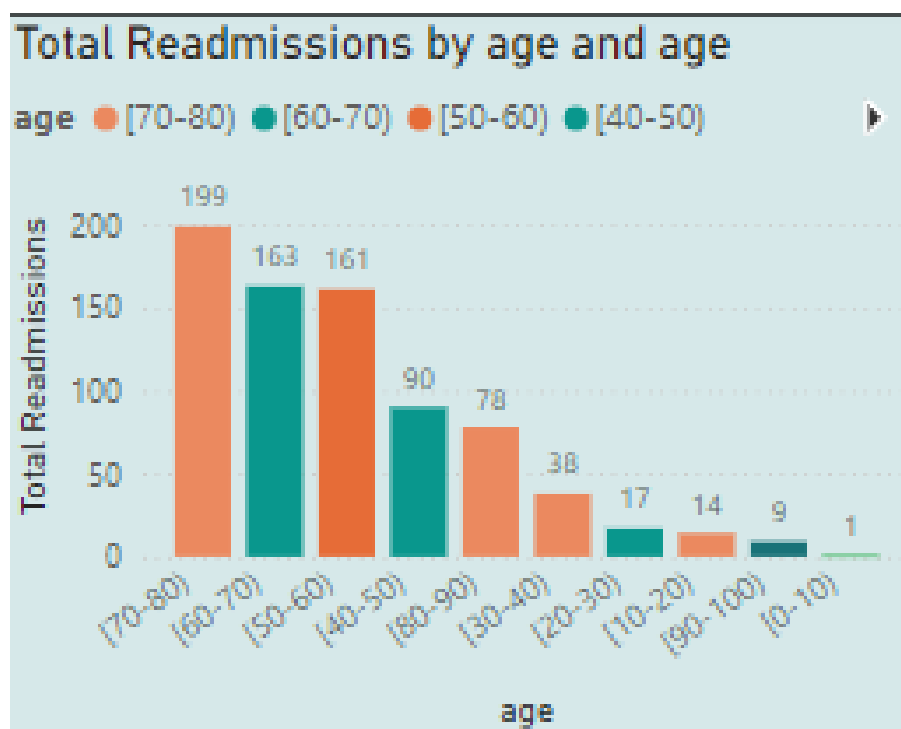


Figure 23: Total Readmission by Age

The "Total Readmission by Age" KPI, which identifies the 70 to 80 years age group as having the highest hospital readmissions, is essential for achieving our business intelligence and machine learning objectives. This KPI helps healthcare providers analyze patterns in readmissions among older patients, enabling targeted interventions and better resource allocation. In machine learning, it enhances the accuracy of predictive models, allowing for personalized care plans that reduce readmissions in this high-risk age group.

readmissions in this high-risk age group. Overall, this KPI supports more effective healthcare delivery and improved patient outcomes.

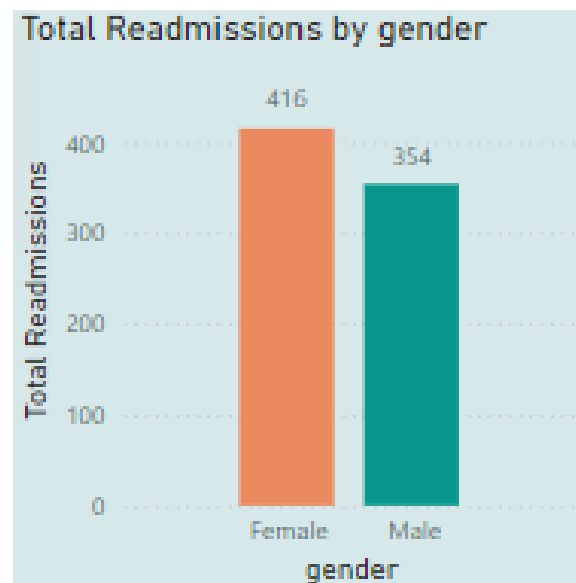


Figure 24: Total Readmission by Gender

The Total Patients with Changes in Medication KPI, which shows an equal split of 10 patients with and 10 without changes in medication, plays a significant role in both business intelligence and machine learning objectives. In terms of business intelligence, this KPI helps identify trends related to medication adjustments and their impact on hospital readmissions, enabling healthcare providers to monitor the effectiveness of treatment changes. It supports decision-making by highlighting whether changes in medication correlate with reduced readmission rates or if further intervention is needed.

For machine learning objectives, this KPI provides critical input data that can enhance the predictive accuracy of models. By analyzing the relationship between medication changes and patient outcomes, the models can better predict which patients are at higher risk of readmission and may benefit from personalized care plans.



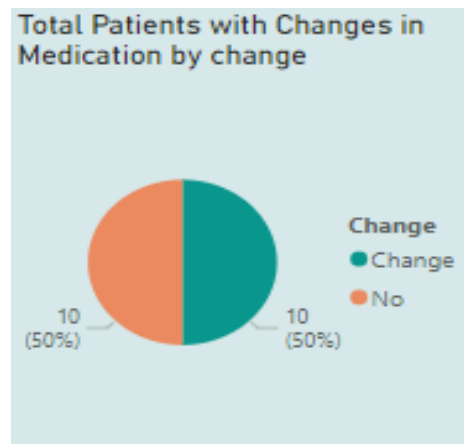


Figure 25: Patients with Change in medication

The Total Readmission by Race KPI shows that Caucasian patients have the highest readmission rates. This is crucial for both business intelligence and machine learning objectives. For business intelligence, it highlights potential racial disparities, guiding targeted interventions and resource allocation. For machine learning, it helps in refining predictive models by incorporating racial data, leading to more accurate and equitable predictions.

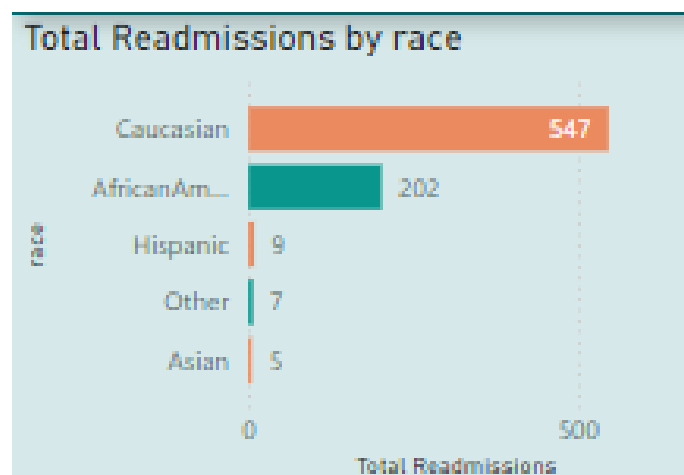


Figure 26: Total Readmission by Race

The KPI on insulin usage reveals increased prescriptions for patients in the 30–40 and 70–80 age groups. This information is key for business intelligence as it highlights medication trends, aiding in targeted care and resource planning. For machine learning, it offers crucial data for refining models to predict readmission risks and optimize treatment strategies based on medication patterns.

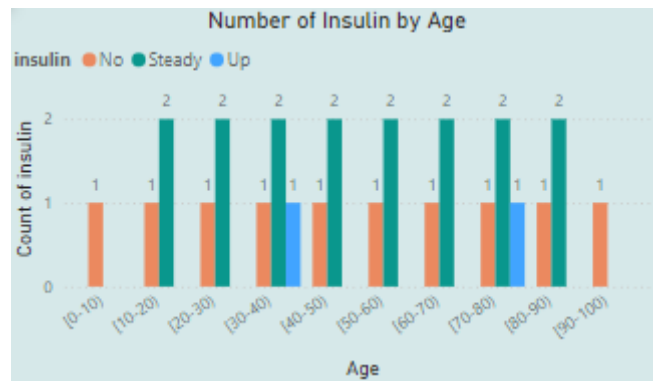


Figure 27: Number of Insulin by Age

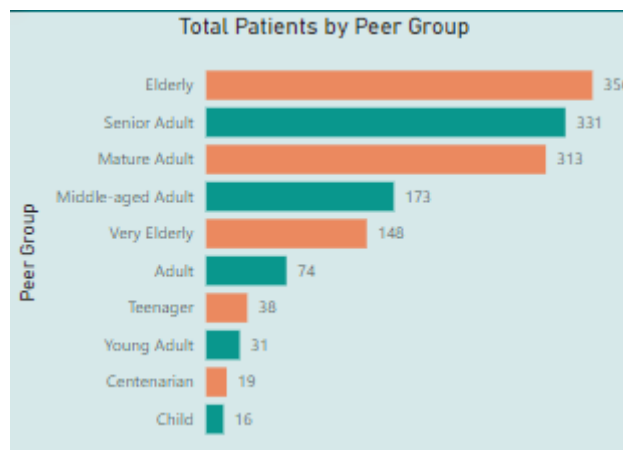


Figure 28: Total Patients by Peer group

The parameter KPI below provides a detailed and interactive view of patient readmission rates and medication averages across different age groups and races as shown in figure 28 - 30. This KPI is crucial for business intelligence as it enables stakeholders to analyse readmission trends by age group and race, identifying high-risk populations such as the 70-80 and 80-90 age groups.

For machine learning objectives, this KPI helps refine predictive models by offering insights into age and race-related risk factors. The ability to view average medication per age group and average time in the hospital further enhances the model's accuracy by incorporating these variables into the analysis. The use of color-coded elements in the KPI makes it easier for stakeholders to interpret the data, providing a clear visual representation of critical metrics. This facilitates better decision-making and strategic planning for reducing readmissions and optimizing patient care.

Parameter										
Readmission Rate										
race	[0-10)	[10-20)	[20-30)	[30-40)	[40-50)	[50-60)	[60-70)	[70-80)	[80-90)	[90-100)
Other	0.00%	0.00%	0.00%		0.00%	0.00%	50.00%	40.00%	100.00%	
Hispanic			50.00%	100.00%	50.00%	20.00%	100.00%	50.00%	66.67%	
Caucasian	6.67%	41.67%	57.14%	47.83%	54.81%	49.51%	48.57%	56.63%	51.22%	47.06%
Asian						80.00%	0.00%	50.00%		
AfricanAmerican		30.77%	57.14%	55.56%	48.48%	59.78%	53.33%	54.55%	55.00%	50.00%

Figure 29: Parameter (Readmission Rate)

Parameter										
Average Time in Hospital										
race	[0-10)	[10-20)	[20-30)	[30-40)	[40-50)	[50-60)	[60-70)	[70-80)	[80-90)	[90-100)
Other	3	4	5		4	5	5	8	7	
Hispanic			3	2	1	5	6	7	8	
Caucasian	2	3	4	4	4	4	5	5	5	5
Asian						5	4	5		
AfricanAmerican		3	3	5	4	4	5	5	4	5

Figure 30: Parameter (Average time in Hospital)

Parameter										
Average Number of Medications per Age Group										
race	[0-10)	[10-20)	[20-30)	[30-40)	[40-50)	[50-60)	[60-70)	[70-80)	[80-90)	[90-100)
Other	15	15	15	15	15	15	15	15	15	15
Hispanic	15	15	15	15	15	15	15	15	15	15
Caucasian	15	15	15	15	15	15	15	15	15	15
Asian	15	15	15	15	15	15	15	15	15	15
AfricanAmerican	15	15	15	15	15	15	15	15	15	15

Figure 31: Parameter (Average Number of Medication Per Age group)

The "Sum of Time in Hospital by Race" and "Total Patients by Race" KPIs reveal that the Caucasian race has the highest number of patients and the longest average hospital stays. This information is crucial for understanding demographic patterns, guiding resource allocation, and refining machine learning models. It helps identify

which racial groups may need additional support and improves predictions for hospital resource needs and patient outcomes based on race.

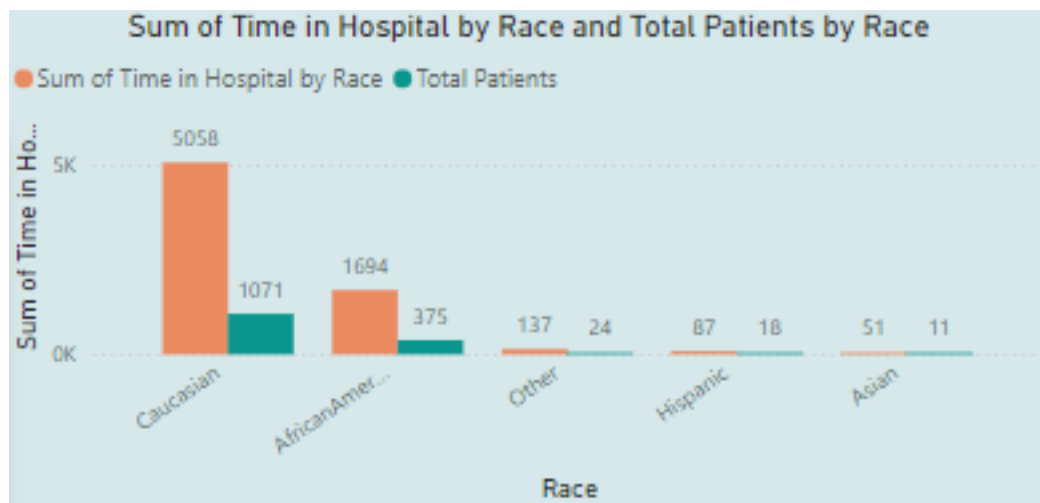


Figure 32: Sum of Total in Hospital by Race and Total Patients by Race

The Medications KPI reveals that medication distribution is consistent across all racial groups, highlighting an equitable approach to treatment. This KPI is relevant to both business intelligence and machine learning objectives as it provides insights into medication equality, which can influence the accuracy of predictive models. In machine learning, understanding uniform medication distribution helps ensure that the models do not inadvertently incorporate bias related to medication access or treatment. For business intelligence, this KPI supports transparency and helps verify that treatment practices are equitable across different demographic groups. Ensuring that medication distribution does not skew results allows stakeholders to make more informed decisions and improve patient care strategies without concern for racial disparities in medication access.

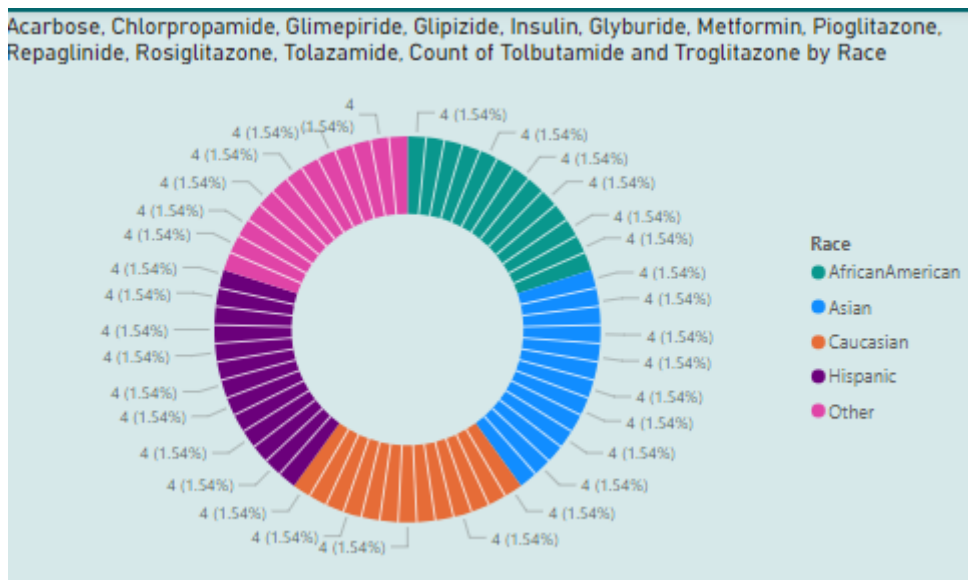


Figure 33: Medications

The "Number of Diagnoses, Procedures, and Lab Procedures" KPI is critical for both business intelligence and machine learning objectives as it provides a comprehensive view of healthcare utilization across different racial groups. This KPI shows that, while the Caucasian race has the highest count of diagnoses, procedures, and lab tests, this is proportional to their higher patient population compared to other races. This balanced distribution across races ensures that the data is representative and that the machine learning models used for predicting readmissions or other outcomes are not biased by differences in healthcare service utilization.

From a business intelligence perspective, this KPI is valuable for identifying and validating that healthcare resources and services are evenly distributed and used across racial groups, which can influence strategic planning and resource allocation. For machine learning, understanding that all races have an equal number of diagnoses, procedures, and lab tests supports the development of fair and accurate predictive models. It ensures that the models are based on consistent and unbiased data, leading to more reliable predictions and interventions. This KPI, therefore, aligns with objectives to ensure equitable healthcare delivery and accurate predictive analytics.

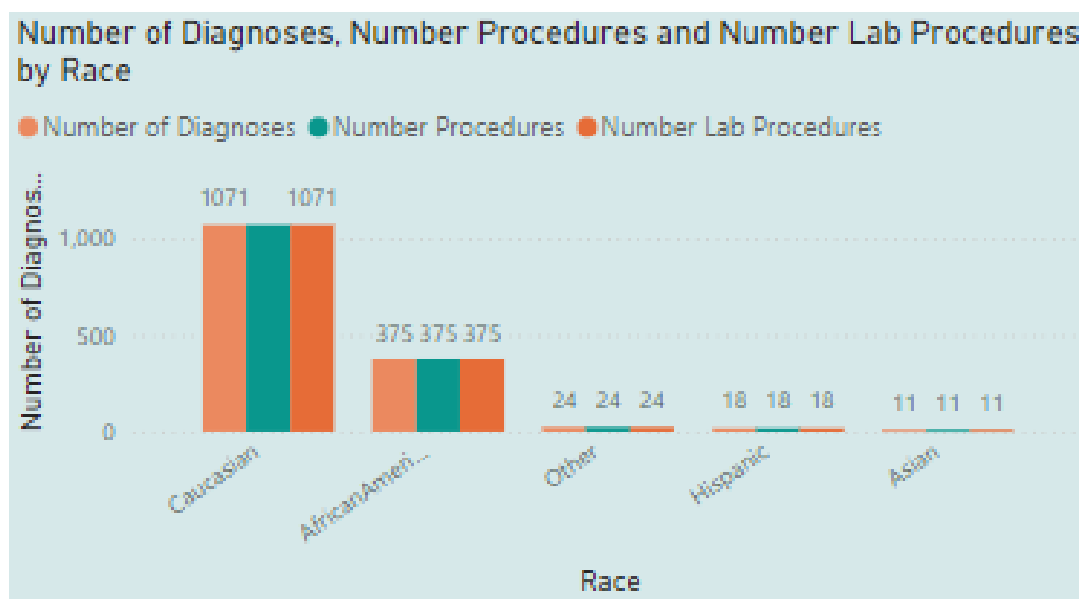


Figure 34: Number of Diagnoses, Procedures and Lab

As part of the ML results, the SVM Risk Stratification KPI is essential for evaluating the effectiveness of the SVM model in identifying high-risk patients. It shows that Class 2 has 147 predicted high-risk patients out of 150 actual cases, highlighting the model's accuracy. This is further discussed in the results chapter. It supports machine learning objectives by validating the model's performance in risk identification, allowing for targeted interventions and efficient resource allocation. For business intelligence, it provides valuable insights into patient risk distribution, helping stakeholders make informed decisions and improve healthcare delivery by focusing on high-risk groups.

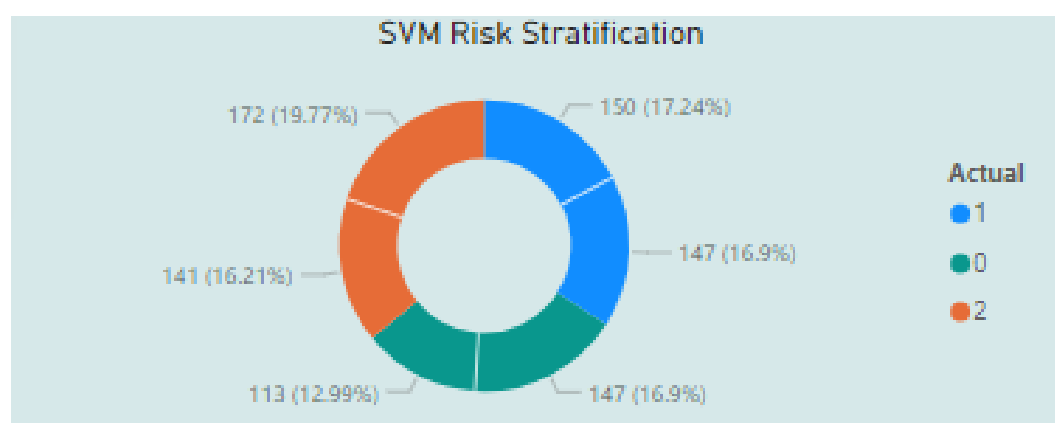


Figure 35 : SVM Risk Stratification

The detailed discussion of the algorithms' classification report, feature importance and LIME KPIs are provided in Chapter 4. However, screenshots of these KPIs can be found in Figures 36 to 41, offering a visual overview of the key performance indicators relevant to the study's objectives.

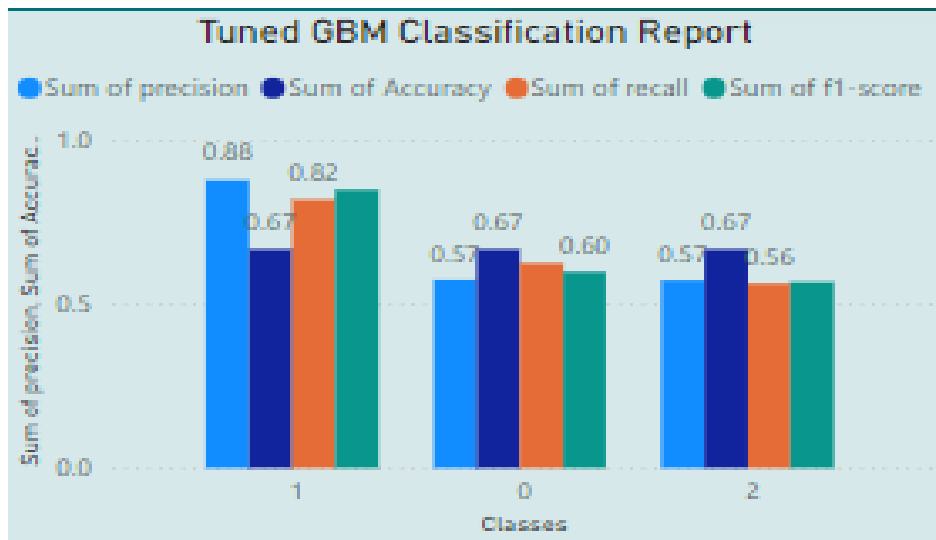


Figure 36: Tuned GBM Classification Report KPI

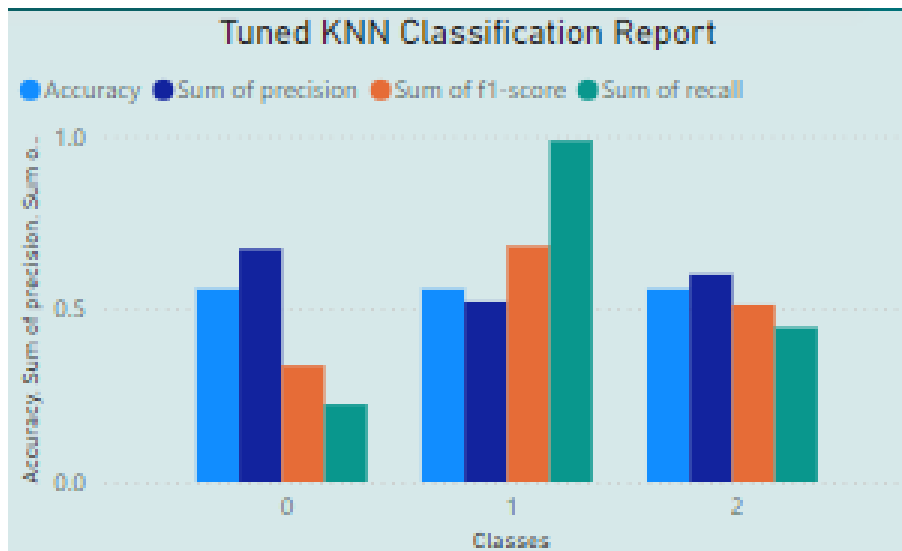


Figure 37: Tuned KNN Classification Report KPI

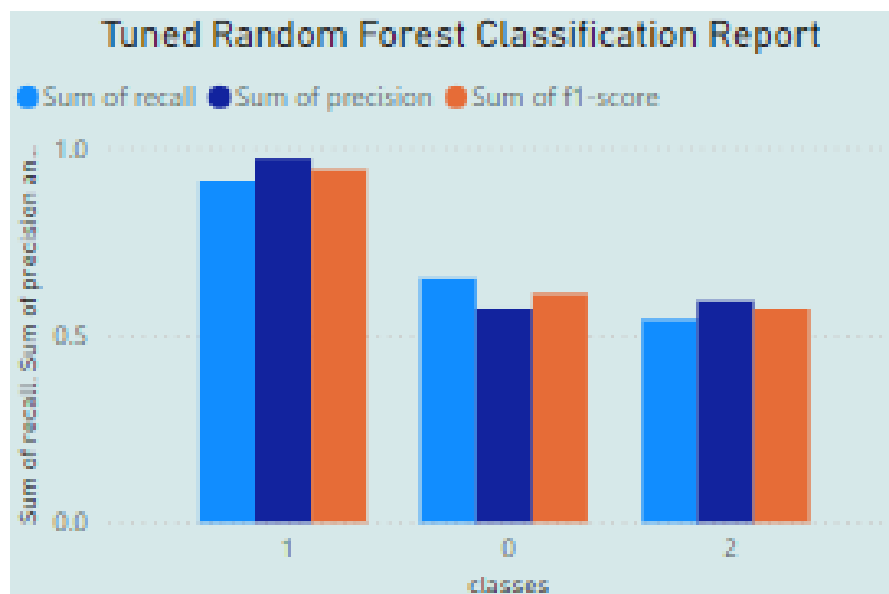


Figure 38: Tuned Random Classification Report KPI

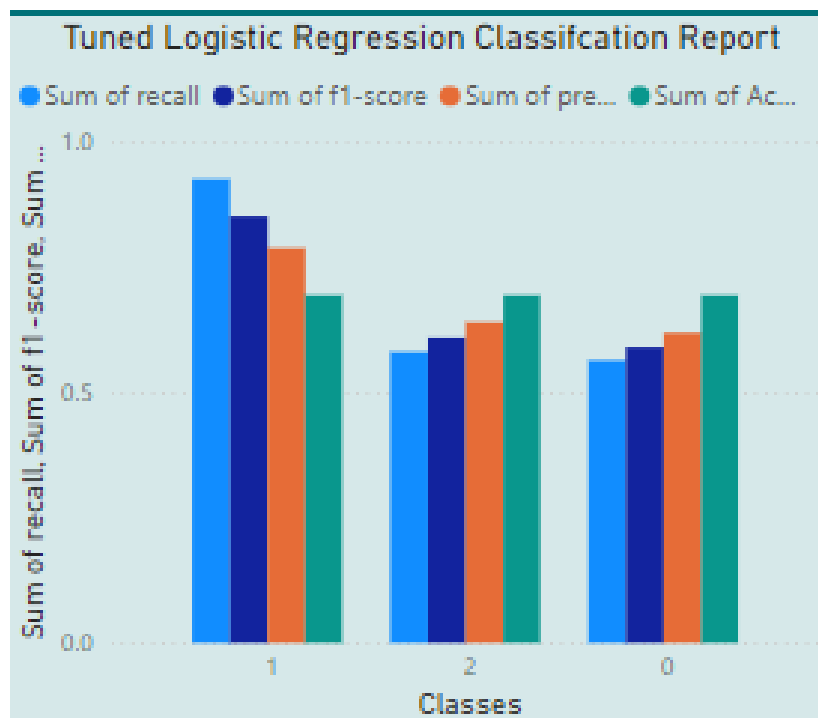


Figure 39: Tuned Logistic Regression Classification Report KPI

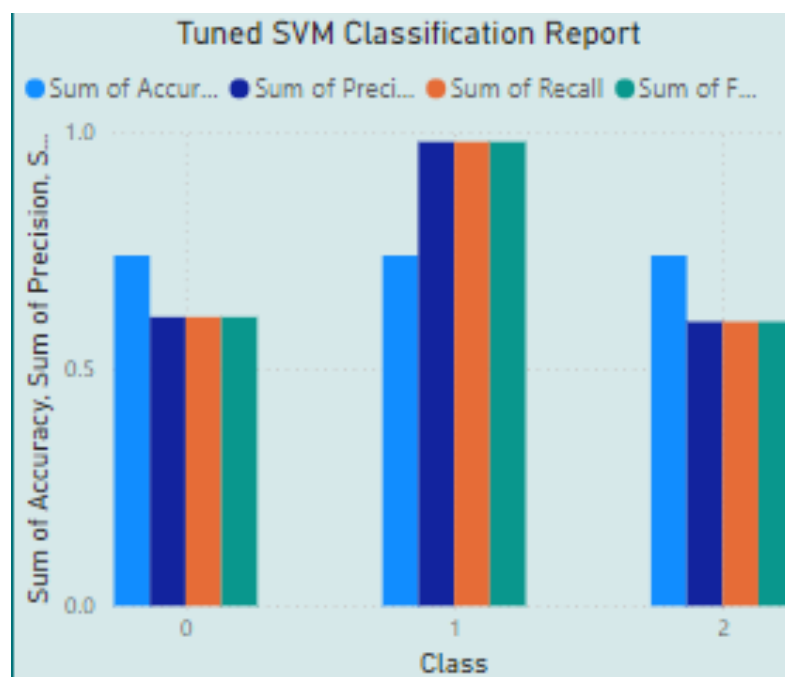


Figure 40: Tuned SVM Classification Report KPI



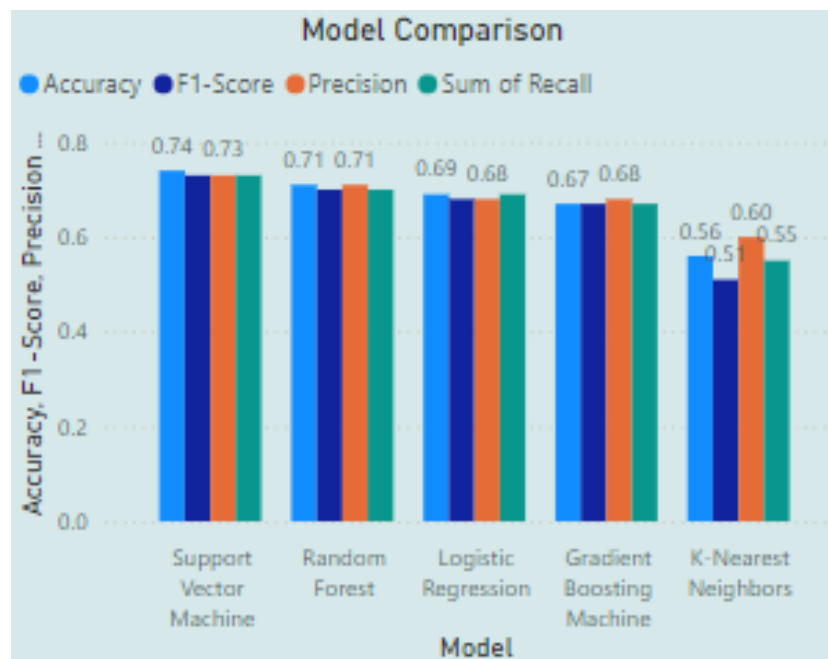


Figure 41: Model Comparison KPI

The LIME and Feature Importance KPIs are thoroughly discussed in Chapter 4, aligning with the study's objectives. For reference, screenshots of these KPIs are also provided from figure 42 - 47.

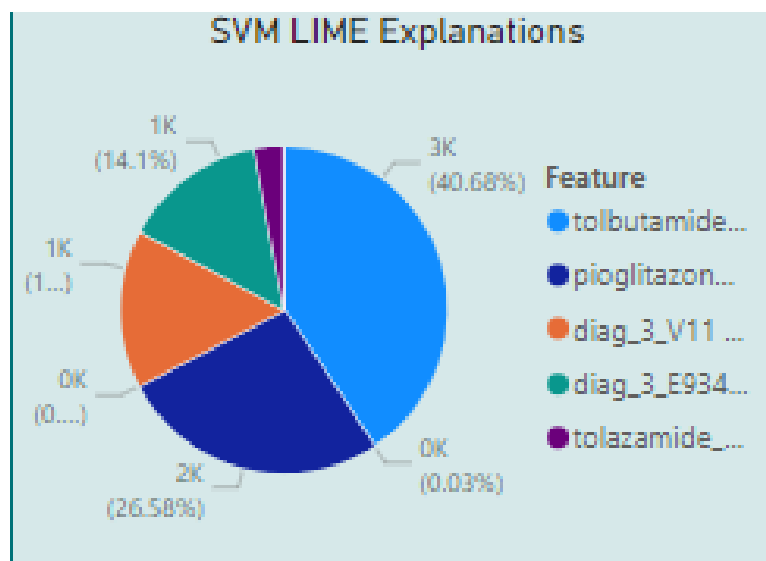


Figure 42: SVM LIME Explanation

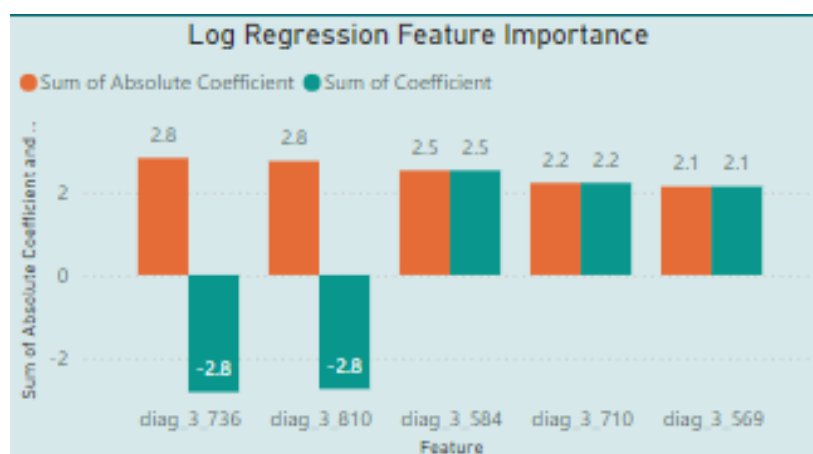


Figure 43: Log Regression Feature importance

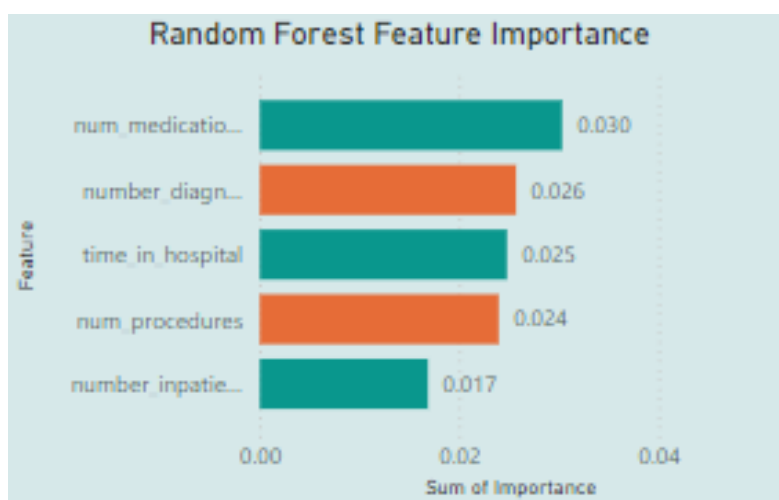


Figure 44: Random Forest Feature importance

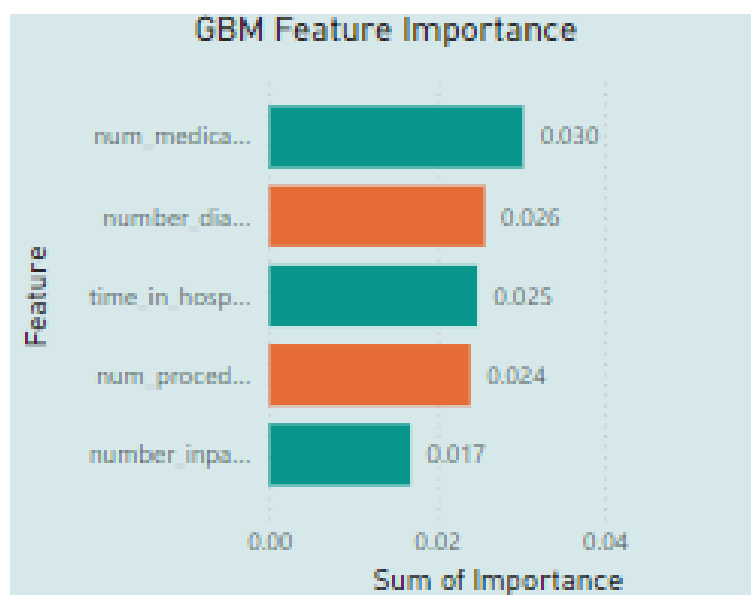


Figure 45: GBM Feature importance

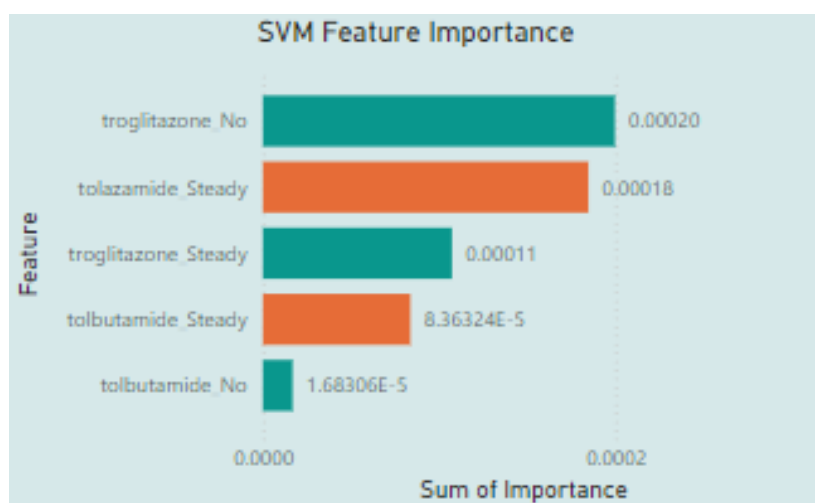


Figure 46: SVM Feature importance

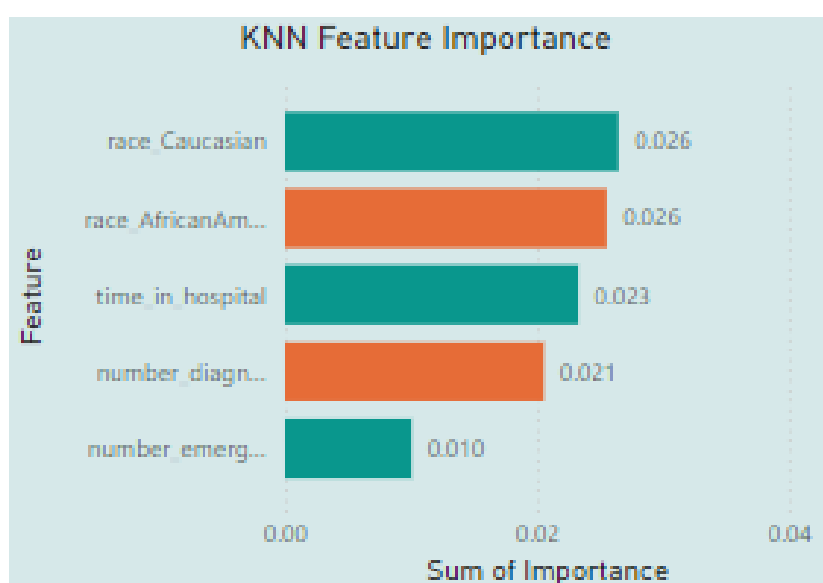


Figure 47: KNN Feature importance

### 3.6 Data Cleaning

The initial phase of data preparation involves data cleaning, a meticulous process designed to address and rectify issues, inconsistencies, and errors within a dataset. This essential step enhances data integrity, ensuring that model applications produce reliable and accurate results. To facilitate this, specialized modules were developed to systematically manage all facets of data cleaning throughout the model's implementation. According to IBM (2021), Data preparation is a vital and frequently labour-intensive phase in data mining, often accounting for 50-70% of the total time and effort in a project. This stage is essential for ensuring the success and accuracy of subsequent analyses.

A crucial aspect of the data cleaning process is the identification and removal of missing or null values. In the dataset used, several variables had missing values: payer\_code had 1,499 missing entries (100%), weight had 1,484 missing entries

(98.999%), medical\_specialty had 425 missing entries (28.35%), diag\_1 had 4 missing entries (0.27%), diag\_2 had 13 missing entries (0.87%), diag\_3 had 52 missing entries (3.47%), and race had 40 missing entries (2.67%), as illustrated in Figure 3. To address this, columns with a high percentage of missing values were removed, while the mode was used to fill in the gaps for columns with minimal missing data. The mode is often used as a simple imputation method to handle missing values in a dataset. After handling the missing values, the dataset was saved and later used for Power BI analytics as indicated below:

```

Missing Value Counts:
encounter_id          0
patient_nbr           0
race                  40
gender                0
age                   0
weight               1484
admission_type_id     0
discharge_disposition_id 0
admission_source_id   0
time_in_hospital      0
payer_code            1499
medical_specialty     425
num_lab_procedures    0
num_procedures        0
num_medications       0
number_outpatient     0
number_emergency      0
number_inpatient      0
diag_1                 4
diag_2                13
diag_3                52
number_diagnoses      0
max_glu_serum         0
A1Cresult             0
metformin             0
repaglinide           0
nateglinide           0
chlorpropamide        0
glimepiride           0
acetohexamide         0
glipizide             0
glyburide             0
tolbutamide           0
pioglitazone          0
rosiglitazone         0
acarbose              0
miglitol              0
troglitazone          0
tolazamide            0
examide               0
citoglipton           0
insulin               0
glyburide-metformin   0
glipizide-metformin   0
glimepiride-pioglitazone 0
metformin-rosiglitazone 0
metformin-pioglitazone 0
change               0
diabetesMed           0
readmitted            0
dtype: int64
Percentage of missing values:

```

Figure 48: Handling Missing Values

Cleaned data saved to C:\Users\D3684573\OneDrive\School\Project\without\_PCA\Cleaned\_Diabetes\_Data\_Without\_PCA.csv

Figure 49: Data Saved for PowerBI Analytics

### 3.7 Data Pre-processing

According to García et al. (2016), data preprocessing in Structural Health Monitoring (SHM) systems involves uncovering valuable information that is hidden by noise and other factors, thus establishing a solid foundation for further data analysis. Data preprocessing is a critical phase that significantly enhances data quality and

addresses complexities, as real-world datasets frequently contain noise, inconsistencies, and redundancies. Effective preprocessing is essential for transforming raw data into a format that is conducive to applying data mining models and achieving the best possible performance. For this study, the dataset underwent a series of comprehensive preprocessing steps to ensure its suitability for advanced analysis.

In the preprocessing phase, several critical steps were undertaken to enhance the quality and suitability of the dataset for subsequent analysis. Initially, column names were cleaned by removing any leading or trailing spaces to ensure consistency and avoid errors during data manipulation. The presence of the essential 'readmitted' column was then verified, as it is crucial for the target variable in the analysis. Age ranges were mapped to numerical values to simplify this categorical variable for modelling purposes. Missing values were addressed using the most frequent value strategy through the SimpleImputer from scikit-learn, ensuring completeness in key columns such as 'age', 'insulin', 'change', and 'diabetesMed'. Interaction terms, such as 'age\_time\_interaction', were created to capture the combined effects of age and hospital stay duration, which may enhance model performance.

### **3.7.1 Data Encoding and Transformation**

The preprocessing involved handling categorical and numerical data separately: categorical variables were one-hot encoded, and numerical variables were standardized using StandardScaler. This transformation ensured that all features were on a comparable scale, which is essential for many machine learning algorithms.

The dataset was then transformed using the defined ColumnTransformer, applying these preprocessing steps and preparing it for modeling. The transformed dataset contained 1,499 rows and 865 columns. The increase in columns was due to one-hot encoding, which expanded the dataset with binary representations of categorical variables, as well as the inclusion of interaction terms. The standardized data now includes features such as `race_AfricanAmerican`, `age_5`, and `admission_type_id`, all transformed to improve the effectiveness of the subsequent analysis. By normalizing numerical features and encoding categorical variables, the preprocessing ensured that the data is well-prepared for accurate and effective modelling.

In the preprocessing phase, an analysis of skewness for each numerical feature was conducted to evaluate the distribution of the data. Initially, the skewness values were computed for features including `patient_nbr`, `number_outpatient`, `number_inpatient`, `time_in_hospital`, `num_procedures`, and `num_medications`, revealing significant skew in several variables. To address this skewness, various transformations were applied.

Log transformation was performed on the data to reduce skewness, resulting in transformed skewness values for most features. This technique successfully mitigated skewness for several variables, although `number_outpatient` and `number_inpatient` retained considerable skewness. Square root transformation was also applied and showed a similar reduction in skewness, yet certain features still exhibited non-normal distributions.

Box-Cox transformation was employed to further normalize the data. However, due to non-positive values in the `number_outpatient`, `number_inpatient`, and `num_procedures` columns, this transformation could not be applied to these features. For the remaining features, Box-Cox transformation effectively reduced skewness, bringing them closer to a normal distribution. Overall, these preprocessing steps were crucial in transforming the dataset into a more suitable format for analysis. By applying log, square root, and Box-Cox transformations, the skewness in numerical features was significantly reduced, thereby enhancing the data's normality and improving the reliability of subsequent statistical and machine learning models.

### **3.7.2 Correlation Matrix and Feature Selection**

The correlation matrix reveals several insights into feature redundancy and the relationships between different transformations of the dataset features. Notably, the features derived from the `patient_nbr`, such as `patient_nbr_log`, `patient_nbr_sqrt`, and `patient_nbr_boxcox`, exhibit very high correlations with each other, demonstrating coefficients close to 1. For example, `patient_nbr_log` and `patient_nbr_sqrt` have a correlation of 0.92, indicating that these transformations capture similar information about the original feature. Similarly, `time_in_hospital_log`, `time_in_hospital_sqrt`, and `time_in_hospital_boxcox` show near-perfect correlations, with coefficients of 0.99 or 1.00, suggesting that these different transformations of `time_in_hospital` essentially represent the same underlying variable.

The features `num_procedures_log` and `num_procedures_sqrt` are perfectly correlated (1.00), indicating redundancy within these transformations. Likewise, `num_medications_log`, `num_medications_sqrt`, and `num_medications_boxcox` show very high correlations, with a coefficient of 0.99 among them, reflecting that they nearly represent the same information.

Moderate to low correlations were observed between features like `num_procedures_log` and `num_medications_log`, which have a moderate correlation of 0.47. This suggests a level of relationship without significant redundancy. On the other hand, features such as `number_outpatient_log` and `number_outpatient_sqrt`,

and number\_inpatient\_log and number\_inpatient\_sqrt, show high correlations of 1.00, indicating that they are redundant.

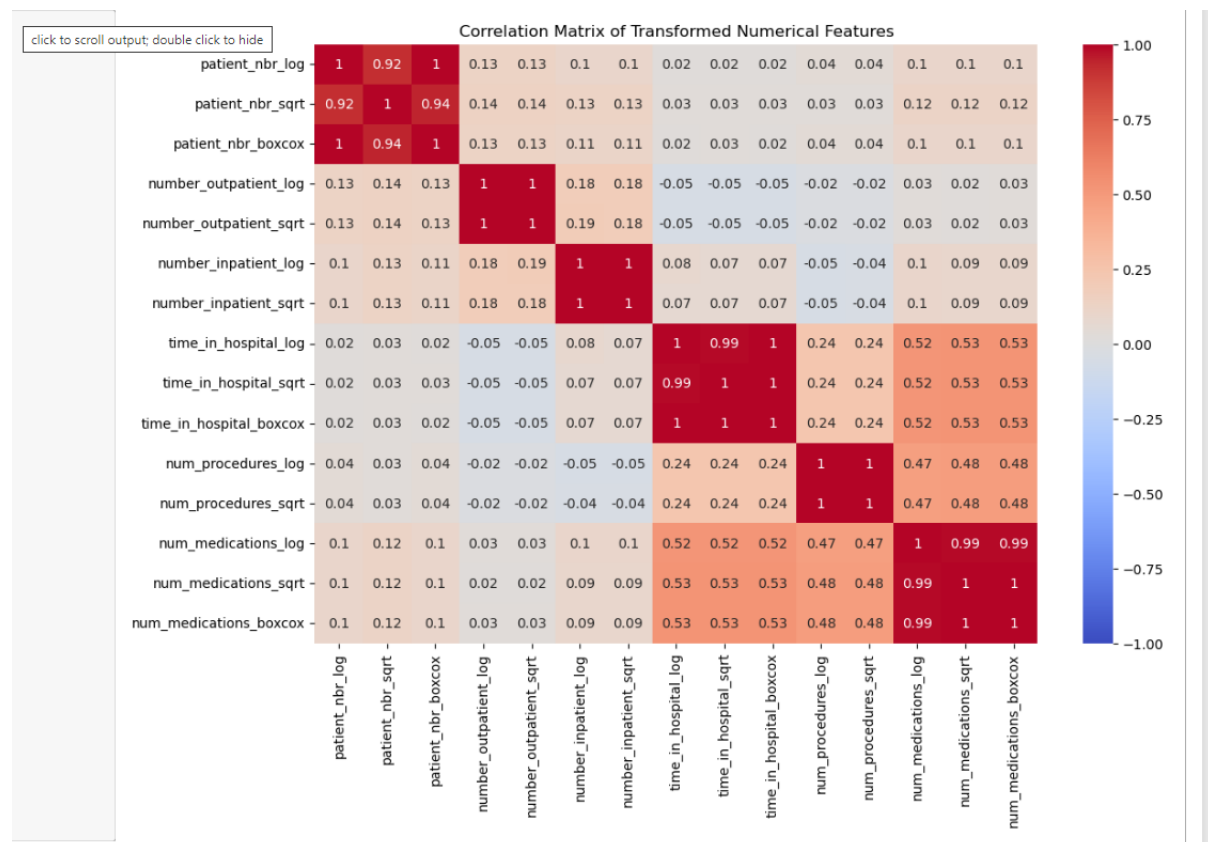


Figure 50: Numerical Features Correlation Matrix

The analysis indicates that many features are highly correlated with their transformed counterparts, which suggests that multiple transformations of the same feature do not provide new information. To address this, several features were identified for removal due to their high correlation with others in the dataset. Specifically, patient\_nbr\_sqrt, patient\_nbr\_boxcox, number\_outpatient\_sqrt, number\_inpatient\_sqrt, time\_in\_hospital\_sqrt, time\_in\_hospital\_boxcox, num\_procedures\_sqrt, num\_medications\_sqrt, and num\_medications\_boxcox were deemed redundant. Removing these features helps in reducing the risk of overfitting, simplifying the model, and improving computational efficiency by eliminating overlapping information and minimizing multicollinearity.

The high correlation among these features indicated that they were providing overlapping information, which could lead to redundancy in the dataset. Removing these features helps mitigate the risk of overfitting, simplifies the model, and reduces computational complexity by minimizing multicollinearity.

Furthermore, the dataset underwent an additional cleaning process where outliers were removed based on Z-scores. Specifically, any rows in the diabledf DataFrame

with numeric feature values having a Z-score greater than 3 were filtered out. This step aimed to enhance the quality of the dataset by eliminating extreme values that could skew analyses or negatively impact model performance. By addressing both feature redundancy and outlier contamination, the preprocessing steps ensured a more robust and reliable dataset for subsequent analysis and modelling.

### **3.7.3 Feature Engineering**

During the feature engineering phase, the analysis revealed that no features were removed based on correlation values, as indicated by the empty list for dropped features. This outcome suggests that all numeric features in the dataset exhibited correlation values below the specified threshold of 0.8. Consequently, there were no highly correlated features that needed to be eliminated during the correlation-based feature selection process.

This result implies that the dataset is well-balanced concerning numeric feature redundancy. The absence of highly correlated features indicates that multicollinearity among the numeric features is minimal, ensuring that the dataset is suitable for modeling. This lack of redundancy in numeric features allows for a more straightforward and effective modeling process, as it minimizes concerns about multicollinearity and enhances the reliability of the subsequent analyses.

### **3.7.4 Feature Transformation and normalisation**

In the preprocessing phase, various types of transformations were applied to the dataset to improve the quality of the data and enhance its suitability for machine learning algorithms. These transformations included log, square root, and Box-Cox transformations, each serving distinct purposes in addressing data skewness and stabilizing variance.

- a. Log Transformation:** This technique is used to reduce skewness, stabilize variance, and make relationships between variables more linear. It is particularly beneficial for features with exponential growth patterns. For example, the transformation applied to `patient_nbr_log`, `number_outpatient_log`, `number_inpatient_log`, `time_in_hospital_log`, `num_procedures_log`, and `num_medications_log` effectively compresses large values and normalizes the data distribution. This transformation converts highly variable or extreme values into a more manageable scale, thus improving the interpretability and performance of machine learning models.
- b. Square Root Transformation:** This transformation helps in reducing skewness, especially for count data, by moderating the range of values and



stabilizing variance. Applied to features such as `patient_nbr_sqrt`, `number_outpatient_sqrt`, `number_inpatient_sqrt`, `time_in_hospital_sqrt`, `num_procedures_sqrt`, and `num_medications_sqrt`, the square root transformation compresses the scale of the data, though less aggressively than the log transformation. It is particularly useful for features that represent counts or proportions, helping to moderate large values and bring them closer to a normal distribution.

- c. **Box-Cox Transformation:** This is a more flexible transformation that accommodates a range of distributions, including log and power transformations as special cases. For features like `patient_nbr_boxcox`, `time_in_hospital_boxcox`, and `num_medications_boxcox`, the Box-Cox transformation adjusts the data to approximate normality, thereby reducing skewness and improving the performance of machine learning algorithms. It is particularly effective for data that does not fit a normal distribution, enhancing model accuracy by stabilizing variance.

### 3.7.5 Standard Scaler

Following these transformations, the `StandardScaler` was employed to standardize the features. This step ensures that all transformed features have a consistent scale, which is crucial for many machine learning algorithms. By removing the mean and scaling to unit variance, the `StandardScaler` makes the data more comparable across features, preventing any single feature from dominating the model due to a larger range. This standardization enhances the performance and reliability of the models by ensuring that the features are on a comparable scale. The formula for the standard scaler is as follows:

$$y = \frac{x - \text{mean}}{\text{standard deviation}}$$

**3.7.6 Summary of Scaled Data:** The transformations applied aimed to normalize the data and stabilize variance across different features, making them more comparable. By applying log, square root, and Box-Cox transformations, the dataset was adjusted to improve its distribution and make it more suitable for machine learning algorithms. This pre-processing step not only addresses skewness but also enhances the effectiveness of the models by ensuring that features are on a comparable scale and conform more closely to the assumptions of normality and homogeneity of variance.

Overall, these transformations were essential in preparing the dataset for analysis, ensuring that the features are well-suited for machine learning algorithms that rely on normality and variance stabilization for optimal performance.

### **3.7.7 Dimensionality Reduction and PCA Impact**

To improve and refine model performance, dimensionality reduction proved to be a vital strategy. This method seeks to decrease the number of features in a dataset, thereby reducing the likelihood of overfitting and potentially enhancing the model's ability to generalize to new data. Principal Component Analysis (PCA) was utilized as the primary technique for this purpose, transforming the dataset into a lower-dimensional form while retaining a significant portion of the original variance. PCA is a well-established feature extraction method that converts original features into a set of uncorrelated components, arranged by the variance they capture. By focusing on the most informative components and discarding less significant ones, PCA simplifies the feature space, which helps in reducing overfitting and improving model interpretability. To assess the effectiveness of PCA, a comparative analysis was performed between the model's performance with the original feature set and after applying PCA. This comparison aimed to evaluate how dimensionality reduction affects model performance, particularly in terms of balancing complexity and accuracy.

Different variance thresholds were tested during the PCA process to find the optimal balance between reducing dataset complexity and maintaining model performance. The goal was to minimize the number of features while preserving essential information. The results showed how various variance thresholds influenced the number of components needed:

- **Variance Threshold 0.95:** Required 5 components to retain 95% of the variance. While this approach preserved nearly all of the variance, it involved retaining a higher number of components, which did not significantly reduce dimensionality.
- **Variance Threshold 0.85:** Required 4 components to retain 85% of the variance. This provided a practical balance, offering a significant reduction in dimensionality while still capturing a substantial proportion of the variance, making it suitable for most applications.
- **Variance Threshold 0.75:** Required 3 components to retain 75% of the variance. This approach achieved the greatest reduction in dimensionality, which can enhance computational efficiency, though it results in a higher loss of variance.

The final PCA configuration selected involved retaining 85% of the variance with 4 components. This choice led to a reduced dataset with a shape of (1499, 4), offering an effective balance between dimensionality reduction and preservation of critical data characteristics.

In summary, using a variance threshold of 85% was chosen as the optimal PCA approach. It strikes a practical compromise between retaining significant variance and simplifying the feature space, thus improving model performance while maintaining computational efficiency. This configuration provides an effective way to handle the dataset for diabetic hospital readmission, balancing the need for detail with the benefits of reduced complexity. This code, The cumulative explained variance code and the PCA application are shown below:

```
In [33]: import numpy as np
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA

# Define variance thresholds
variance_thresholds = [0.95, 0.85, 0.75]
explained_variances = {}
num_components = {}

# Apply PCA for each variance threshold and store results
for threshold in variance_thresholds:
    pca = PCA(n_components=threshold)
    X_pca = pca.fit_transform(X_transformed_df)

    # Store explained variance ratios and number of components
    explained_variances[threshold] = np.cumsum(pca.explained_variance_ratio_)
    num_components[threshold] = X_pca.shape[1]

# Plot cumulative explained variance for each setting
plt.figure(figsize=(12, 8))
for threshold in variance_thresholds:
    plt.plot(range(1, len(explained_variances[threshold]) + 1),
             explained_variances[threshold],
             marker='o', label=f'Variance Threshold: {threshold}')

plt.xlabel('Number of Components')
plt.ylabel('Cumulative Explained Variance')
plt.title('Cumulative Explained Variance for Different PCA Settings')
plt.legend()
plt.grid(True)
plt.show()

# Print the results in text for analysis
print("PCA Explained Variance Ratios and Number of Components")
for threshold in variance_thresholds:
    print(f"\nVariance Threshold: {threshold}")
    print(f"Number of Components: {num_components[threshold]}")

# Print cumulative explained variance for the first few components
print("Cumulative Explained Variance:")
for i, variance in enumerate(explained_variances[threshold]):
    print(f"    Component {i + 1}: {variance:.4f}")
```

Figure 51: The cumulative explained variance code

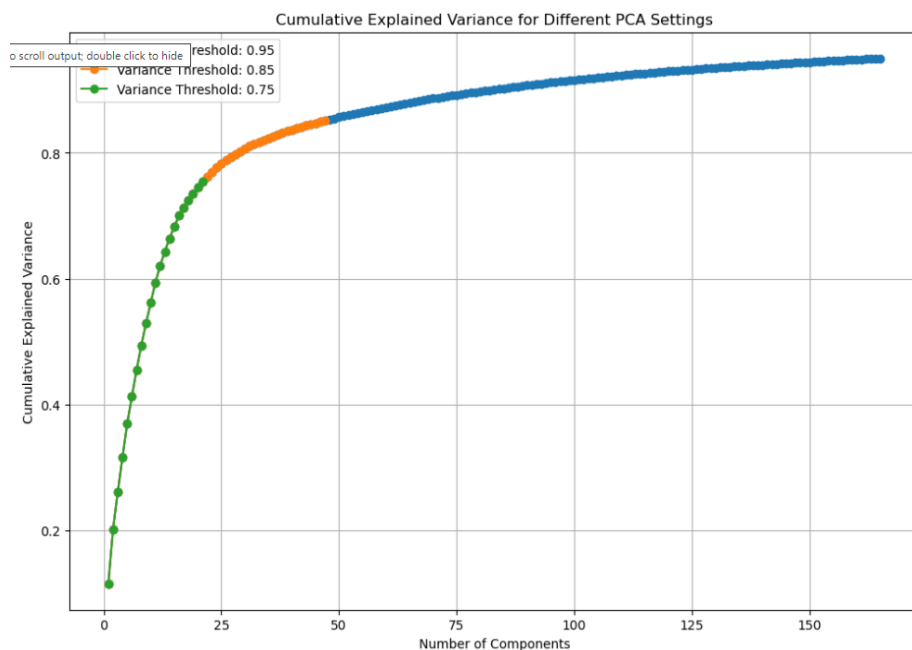


Figure 52: Cumulative Explained Variance for Different PCA settings

```
In [34]: from sklearn.decomposition import PCA

# Choose to keep 85% of the variance
pca = PCA(n_components=0.85)
X_pca_85 = pca.fit_transform(X_transformed_df)

# Check the shape of the data after PCA
print(f'Shape of the data after PCA (85% variance): {X_pca_85.shape}')

Shape of the data after PCA (85% variance): (1499, 53)
```

Figure 53: The PCA Application

### 3.7.8 Distribution of Classes Before and After Applying SMOTE

Initially, the distribution of the target variable, readmitted, exhibited a notable class imbalance. Class 0, representing patients not readmitted, comprised 58% of the dataset. In contrast, Class 1, indicating readmissions within 30 days, constituted only 12% of the dataset, while Class 2, representing readmissions after 30 days, made up 30%. This imbalance suggests that the model might be biased toward the majority class (Class 0) and may underperform in predicting the minority classes (Class 1 and Class 2). Addressing this imbalance is vital to guarantee that the model provides accurate predictions across all classes, particularly for the minority classes that are often of greater clinical interest. Such imbalances can skew model performance, potentially leading to a bias toward the majority class. To mitigate this issue, it is crucial to employ strategies such as re-sampling the minority classes, adjusting class weights, or utilizing algorithms designed to handle class imbalances effectively.

This is shown below:

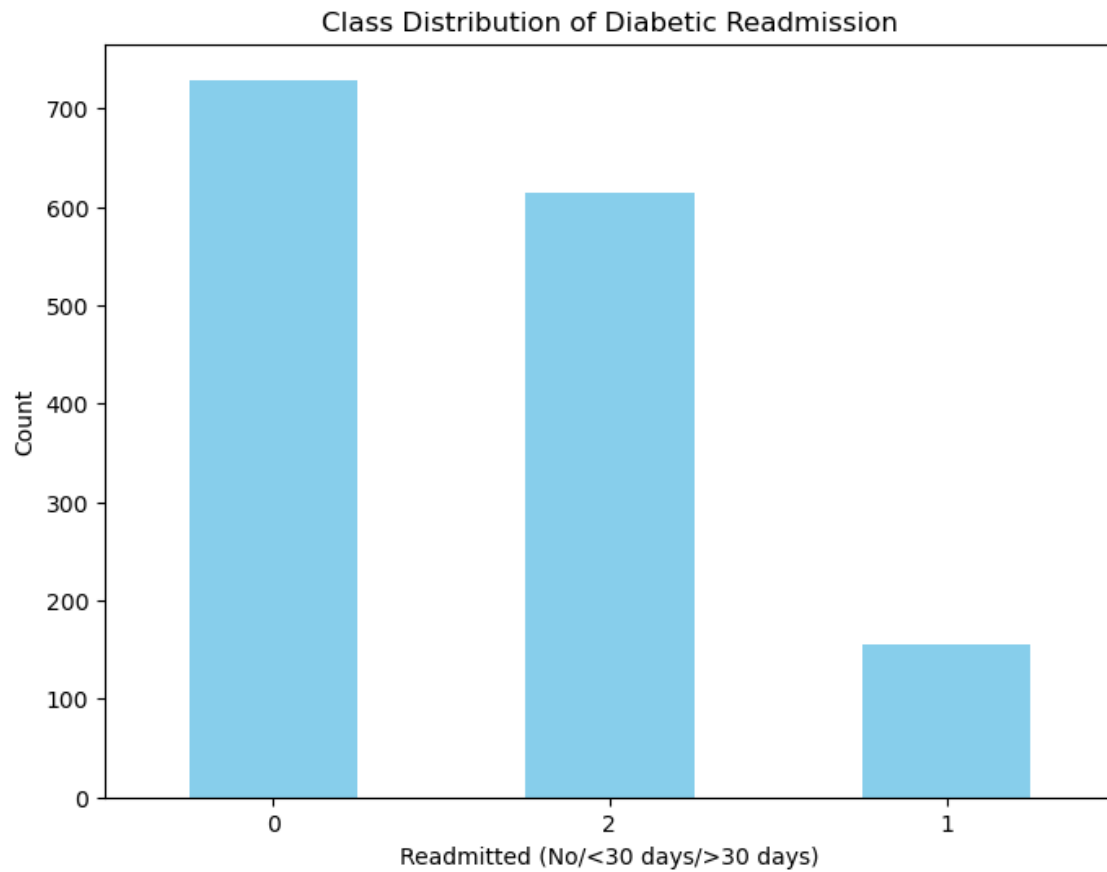


Figure 54: Class distribution of Diabetic Readmission

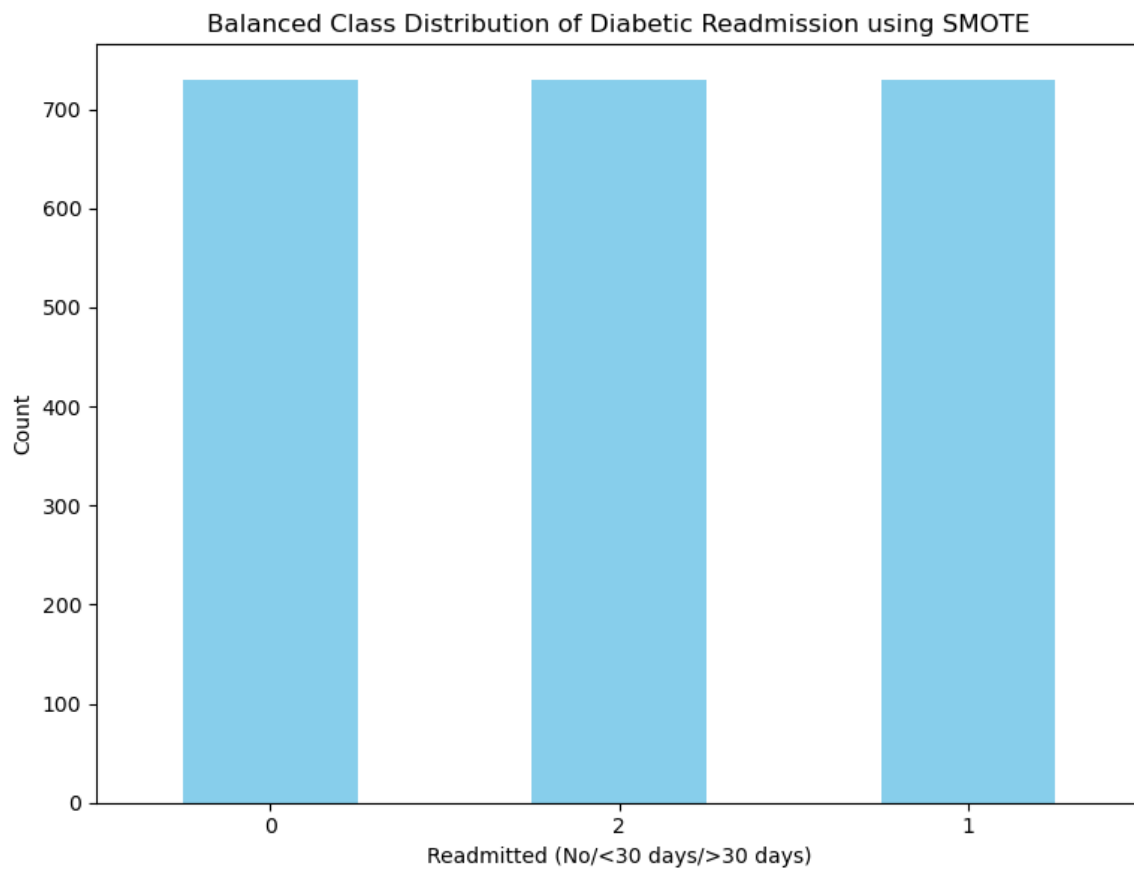


Figure 55: Class distribution of Diabetic Readmission using SMOTE

### 3.7.9 Class Distribution After SMOTE:

Following the application of the SMOTE, the dataset achieved a perfect balance among all classes. After SMOTE, each class contains 729 instances each. This balanced distribution corrects the initial imbalance, ensuring that the model training process is now evenly distributed among all classes. As a result, the model can be trained more effectively to recognize patterns in each class without being biased toward any one class.

### 3.8 Data Splitting

Before applying machine learning models, the data must be divided into two subsets: one for training the model and the other for testing its performance. This approach allows for an unbiased evaluation of the model on unseen data, which helps to assess its generalization ability.

```
# Split the upsampled data into training and testing sets (80% train, 20% test)
X_train, X_test, y_train, y_test = train_test_split(
    X_resampled, y_resampled, test_size=0.2, random_state=42
)
```

Figure 56: Data Splitting

### 3.9 Machine Learning Models

**3.9.1 Logistic Regression (LR):** LR is a widely used machine learning technique designed for classification tasks. The logistic function and its characteristics were initially presented by Pierre Franois Verhulst in a paper published in the Proceedings of the Belgian Royal Academy. In this work, Verhulst detailed the logistic function by defining three critical parameters, which in turn describe the characteristic curve that passes through them. LR is celebrated for its simplicity and effectiveness, making it a go-to method in various applications. Logistic Regression (LR) is a statistical model used to forecast binary outcomes where the dependent variable is distributed according to a Bernoulli distribution. This model utilizes the logistic or sigmoid function, which is an 'S'-shaped curve that produces output values between 0 and 1. The logistic function is structured such that as the input value increases towards positive infinity, the output approaches 1, reflecting a high likelihood of the positive class. Conversely, as the input value decreases towards negative infinity, the output approaches 0, indicating a high likelihood of the negative class (Majumder et al., 2021). This behaviour of the logistic function, which is visually represented in Figure 2, underpins its effectiveness in classification scenarios.

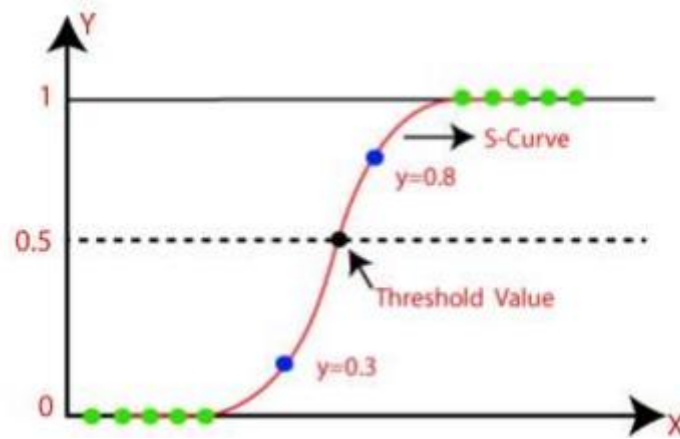


Figure 57: Logistic Regression Sigmoid Function

**3.9.2 KNN:** According to Cover and Hart (1967), in pattern classification, the Nearest Neighbour (NN) rule is among the simplest and most enduring methods. The core idea of NN is to find the closest training sample to a given query and assign the class label of this nearest neighbour to the query. This straightforward approach lays the foundation for classifying data based on proximity.

An enhanced variant of NN is the k-Nearest Neighbours (KNN) algorithm. Unlike NN, which relies on a single nearest neighbour, KNN classifies a query by taking into account the class labels of its k nearest neighbours from the training data. The class label for the query is determined through a majority vote among these k-nearest neighbours.

The KNN algorithm consists of two main steps:

- **Neighbour Identification:** Identify the k closest labelled neighbours to the query. This involves finding the nearest data points in the training set and ranking them based on their distance to the query. The set of these closest neighbours will be used for the classification decision.
- **Class Prediction:** Predict the class label of the query by performing a majority vote among the k nearest neighbours. The predicted class label is determined by counting which class label appears most frequently among these nearest neighbours and assigning that label to the query. Below graphical representation of KNN as well as the formula:

*Euclidean Distance Calculation:*

$$d(x', x_{NNi}) = \sqrt{(x' - x_{NNi})^T (x' - x_{NNi})}$$

*Class Prediction Using Majority Voting:*

$$y' = \arg \max_y \sum_{(x_{NNi}, y_{NNi}) \in T'} \delta(y = y_{NNi})$$

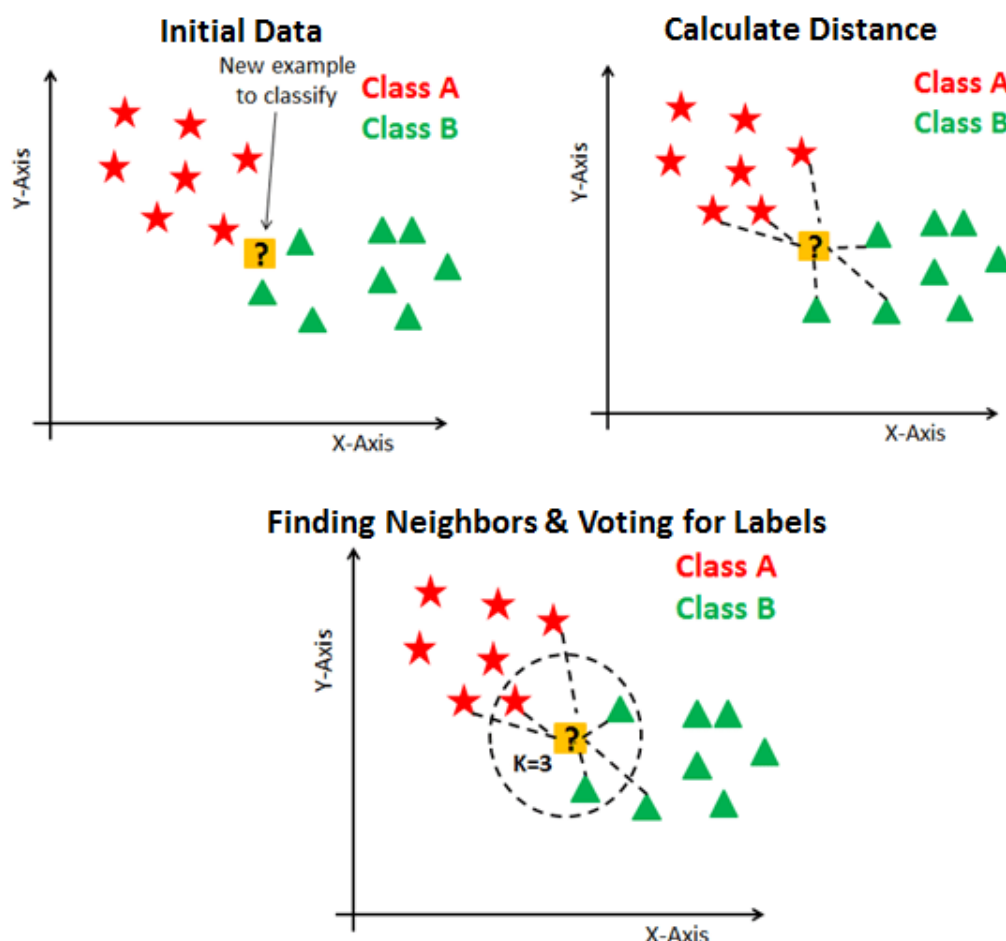


Figure 58: KNN Model

Source: <https://machinelearninggeek.com/knn-classification-using-scikit-learn/>

**3.9.3 Gradient Boosting Machines (GBMs):** To bridge the gap between boosting methods and statistical theory, Freund and Schapire (1997), along with Friedman et al. (2000) and Friedman (2001), formulated a gradient-descent-based approach that led to the development of GBMs. GBMs enhance predictive accuracy by sequentially adding new models that refine the estimate of the target variable. The underlying mechanism involves creating new base-learners that are highly aligned with the negative gradient of the loss function across the entire model ensemble. Although the choice of loss function can be flexible, using squared-error loss demonstrates how the method iteratively corrects errors. Researchers have the freedom to choose from various loss functions or develop custom ones tailored to specific problems.

This adaptability makes GBMs highly versatile for a wide range of data-driven applications, offering significant design flexibility. While this flexibility can require iterative experimentation to determine the most suitable loss function, GBMs are generally easy to implement and experiment with. Their success is well-documented across diverse practical and academic settings, highlighting their robustness in addressing several challenges related to machine learning and data mining (Bissacco



et al., 2007; Hutchinson et al., 2011; Pittman and Brown, 2011; Johnson and Zhang, 2012). The following is a flowchart of the Gradient Boosting Machine (GBM):

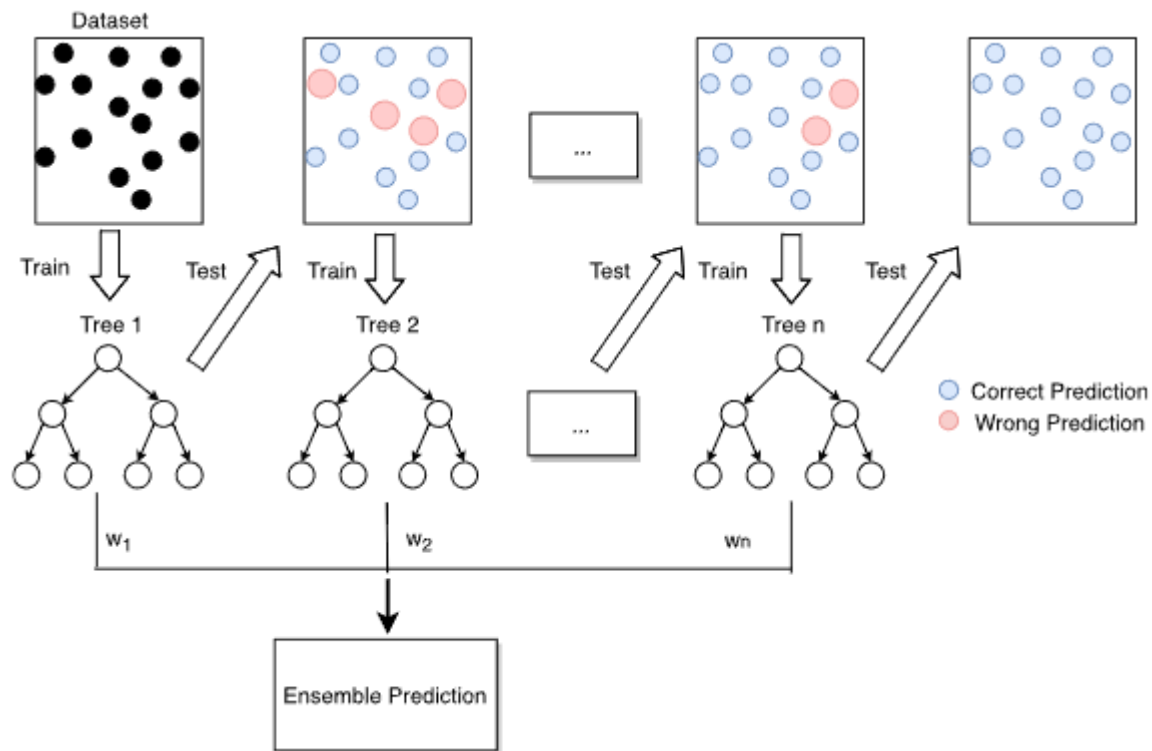


Figure 59: Flow diagram of GBM

Source: <https://agupubs.onlinelibrary.wiley.com/doi/epdf/10.1029/2020MS002365>

**3.9.4 Random Forest (RF):** RF is an ensemble classification method that constructs a collection of decision trees, each operating independently and differing from one another due to randomization. The fundamental idea of Random Forest is to generate numerous decision trees, each constructed from a randomly chosen subset of features and a random sample of the training data. This randomness in both feature selection and data sampling contributes to the diversity among the decision trees, enhancing the overall performance of the model.

The construction process for a Random Forest follows these steps:

- Bootstrap Sampling:** From the entire training dataset, multiple subsets are created by randomly sampling with replacement. Each subset is used to construct a separate decision tree, while the samples not selected in this process form what is known as out-of-bag (OOB) data.
- Random Feature Selection:** At each node of a decision tree, a random subset of features is chosen, and the best feature for splitting the node is selected from this subset.

- c. **Tree Growth:** Each decision tree in the forest is allowed to grow fully without any pruning, meaning that every tree can continue splitting until it perfectly classifies the training data within its subset.
- d. **Model Formation:** Once all the decision trees are constructed, they are combined to form the Random Forest model. This ensemble model is then used for identifying and classifying unknown data based on the majority vote of the individual trees.

#### 3.9.4.1 Key Calculations:

- **Bootstrap Sampling:** The process involves sampling the dataset multiple times to create subsets for training individual trees.
- **Out-of-Bag (OOB) Data:** These are the data points not included in a specific tree's training subset, used for validation and performance estimation.
- **Random Feature Selection:** A subset of features is randomly chosen at each node, and the best split is determined from this subset.

This method's strength lies in its ability to handle large datasets with higher dimensionality and its robustness against overfitting, making it a powerful tool for classification tasks. The Random Forest flow is shown below:

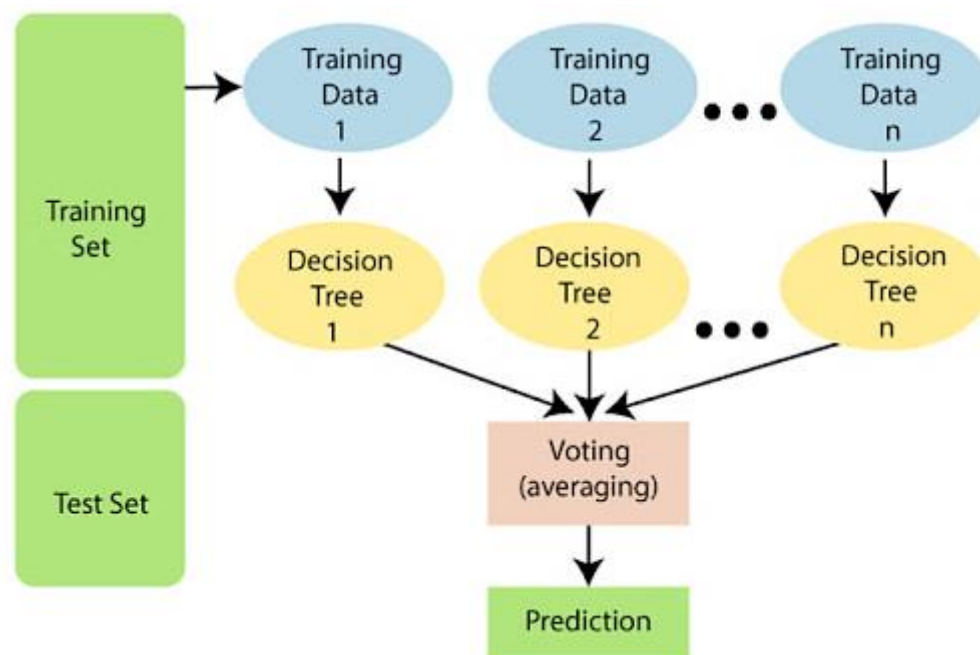


Figure 60: Flow diagram of Random Forest

Source: <https://www.simplilearn.com/tutorials/machine-learning-tutorial/random-forest-algorithm>

**3.9.5 Support Vector Machine (SVM):** According to Wendong, Zhengzheng, and Bo (2017), the Support Vector Machine (SVM) algorithm is a powerful supervised machine learning technique that is deeply rooted in the principles of statistical learning theory. Kareem and Abdulazeez (n.d.), along with Zeebaree et al. (2018), describe a process

where specific subsets of characteristics are selected from the training data. These subsets are designed to ensure that their classification mirrors the classification of the entire dataset. Support Vector Machines (SVM) have demonstrated effectiveness across a wide range of classification challenges, showcasing their versatility and robustness in various practical applications.

Support Vector Machines (SVMs) are adept at both linear and non-linear classification tasks. They achieve non-linear classification through a powerful technique known as the kernel trick, which allows them to implicitly project input data into higher-dimensional feature spaces. This transformation enables SVMs to create complex decision boundaries using a hyperplane that can more effectively separate different classes. The kernel trick essentially facilitates drawing decision margins in these high-dimensional spaces, ensuring that the distance between the margin and the nearest data points from each class is maximized. By optimizing these margins, SVMs reduce classification errors and enhance the overall accuracy of the model. Below is the image representation of SVM.

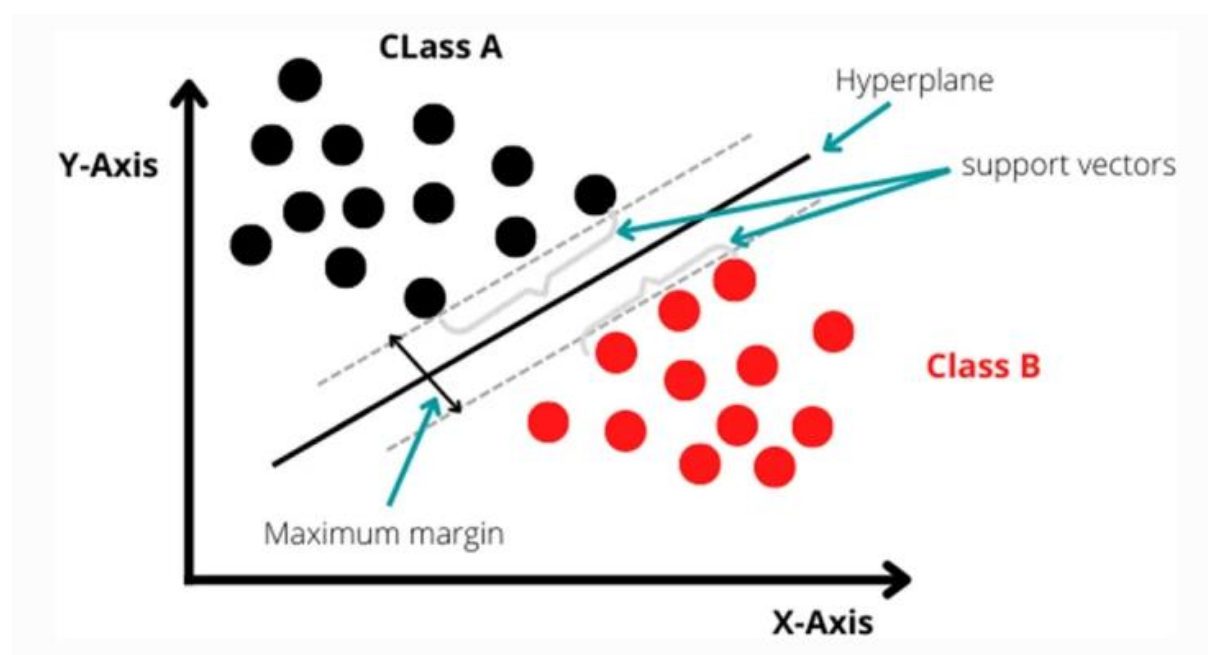


Figure 61: Support Vector Machine (SVM)

Source: <https://jesit.springeropen.com/articles/10.1186/s43067-023-00101-5>

The selection of Logistic Regression (LR), k-Nearest Neighbors (kNN), Gradient Boosting Machine (GBM), Random Forest (RF), and Support Vector Machine (SVM) for this study is strategically justified based on their unique strengths in handling various aspects of hospital readmission prediction. Logistic Regression is chosen for its high interpretability, providing clear insights into the relationships between input features and readmission risk, which is crucial for decision-making in healthcare. kNN

offers simplicity and intuition, making it valuable for exploring patient profiles and their similarities in relation to readmission. GBM is included for its exceptional predictive accuracy and ability to capture complex interactions among features, making it highly effective in dealing with structured healthcare data. Random Forest is selected for its robustness against overfitting and its capacity to identify key risk factors through feature importance analysis, enhancing both the model's reliability and its practical utility. Lastly, SVM is chosen for its effectiveness in high-dimensional spaces and its versatility in modeling complex, non-linear relationships through kernel functions, making it well-suited for the intricate patterns often found in patient data. Together, these algorithms provide a comprehensive and balanced approach, combining interpretability, accuracy, and the ability to manage complex data, thereby optimizing the prediction and management of hospital readmissions.

### 3.9.6 Evaluations Metrics

This section offers a detailed examination of the metrics used to assess the performance of the ML models developed in this study. The evaluation process involves comparing the predicted results of test samples with their actual outcomes, leading to four possible scenarios: false positives (FP), true negatives (TN), true positives (TP), and false negatives (FN). The assessment also includes several important factors such as feature importance, permutation importance, risk stratification, interpretability through LIME (Local Interpretable Model-agnostic Explanations), and the ROC AUC (Receiver Operating Characteristic Area Under the Curve) to gauge the models' overall accuracy and effectiveness.

In the context of predicting diabetic hospital readmissions, The subsequent metrics offer a thorough assessment of the model's performance:

- **TP:** This metric captures instances where the model correctly identifies patients who will indeed be readmitted due to diabetes-related complications. It reflects the model's capability to accurately foresee the need for further medical intervention.
- **TN:** This indicates the model's efficacy in correctly predicting that certain patients will not require readmission. It shows the model's capacity to accurately discern cases where additional hospitalization is unnecessary.
- **FP:** This metric represents cases where the model mistakenly predicts that a patient will be readmitted, even though they do not actually require it. Such errors might lead to pointless stress for patients and possibly unneeded medical procedures.

- **FN:** This is a critical measure of when the model fails to predict a necessary readmission, incorrectly concluding that a patient does not need to be readmitted despite the presence of complications. This type of error could result in insufficient care and missed opportunities for timely intervention.

The evaluation metrics are described in the following sections.

**3.9.6.1 Accuracy (ACC):** Accuracy (ACC) quantifies the percentage of correct predictions, combining true positives (TP) and true negatives (TN), relative to the total number of predictions made (P + N). This metric is particularly effective when the dataset features a balanced distribution of instances across all classes. The formula to compute accuracy is graphically represented as follows:

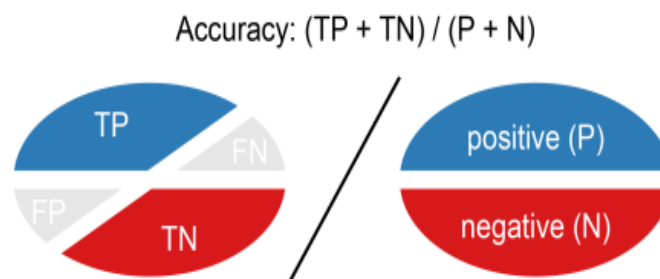


Figure 62: Accuracy (ACC) Formula

Source: <https://classeval.wordpress.com/introduction/basic-evaluation-measures/>

**3.9.6.2 Precision (PREC):** This metric is employed to evaluate the effectiveness of a classification model, especially in situations where false positives (FP) carry a significant cost. It measures the accuracy of the model's positive predictions. Precision is especially important in contexts where identifying true positives (TP) accurately is crucial, such as in medical diagnoses. The formula to calculate the precision is graphically represented as shown below:

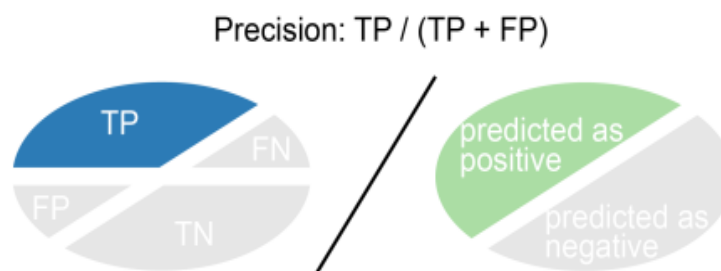


Figure 63: Precision (PREC) Formula

Source: <https://classeval.wordpress.com/introduction/basic-evaluation-measures/>

**3.9.6.3 Recall (REC):** Recall is a performance metric for classification models, especially useful when the cost of overlooking positive instances is significant. It assesses the model's ability to detect all relevant positive cases in the dataset.

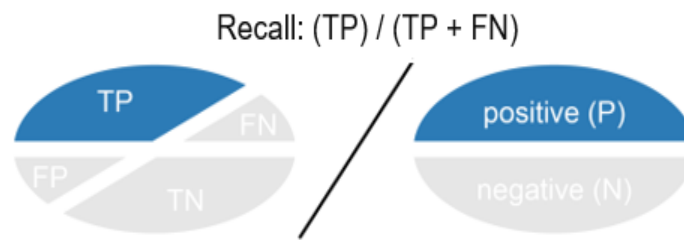


Figure 64: Recall (REC) Formula

**3.9.6.4 F1 Score:** The F1 Score is a metric that combines both precision and recall into a single value, providing a balanced measure of a model's accuracy. It is particularly useful in situations where the classes are imbalanced, meaning there are significantly more instances of one class than the other. The F1 Score is the harmonic mean of precision and recall, which helps to ensure that both metrics are considered when evaluating the model's performance.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

**3.9.6.5 Confusion Matrix:** A confusion matrix is an essential tool for assessing the effectiveness of a classification model. It offers a comprehensive summary of how the model's predictions align with the actual outcomes. By presenting a detailed view of where the model succeeds and where it fails, the confusion matrix helps in evaluating the model's performance. It is structured as a table displaying four distinct categories that compare predicted values with actual outcomes as shown below.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 65: Confusion Matrix

Source: <https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62>

In addition to these core metrics, the model's performance is further evaluated using:

**3.9.6.6 Feature Importance:** This analysis identifies the key factors driving the model's predictions, highlighting which variables most significantly impact the likelihood of diabetic readmission. Understanding feature importance helps refine the model and improve its predictive accuracy.

**3.9.6.7 Permutation Importance:** Permutation importance is a method used to assess the importance of each feature in a machine learning model. It measures how much a model's performance decreases when the values of a specific feature are shuffled, helping to identify which features are most crucial for making accurate predictions. The results rank features by their impact on the model's performance.

**3.9.6.8 Risk Stratification:** By categorizing patients into different risk levels based on their likelihood of readmission, this process enables healthcare providers to prioritize and tailor interventions for those at higher risk, potentially improving patient outcomes.

**3.9.6.9 LIME:** LIME is used to provide individualized explanations for the model's predictions, making it easier to understand why specific predictions were made. This interpretability is crucial for trust in the model's decisions, especially in a clinical setting.

**3.9.6.10 ROC AUC:** The ROC AUC metric evaluates the model's overall ability to distinguish between patients who will and will not be readmitted. A higher AUC score indicates superior model performance, demonstrating its effectiveness in predicting readmissions

### **3.10 Predictive Analytics for Diabetic Patient Readmission Using Business Intelligence Tools**

In Power BI, the data cleansing processes were meticulously handled within the 'Query Settings' pane of the Power Query Editor, with each action documented in the 'Applied Steps' section figure 60. Users can effortlessly remove or undo any step by clicking the cross next to the step name. The steps are listed in the precise order of their application, allowing for a clear and organized workflow.

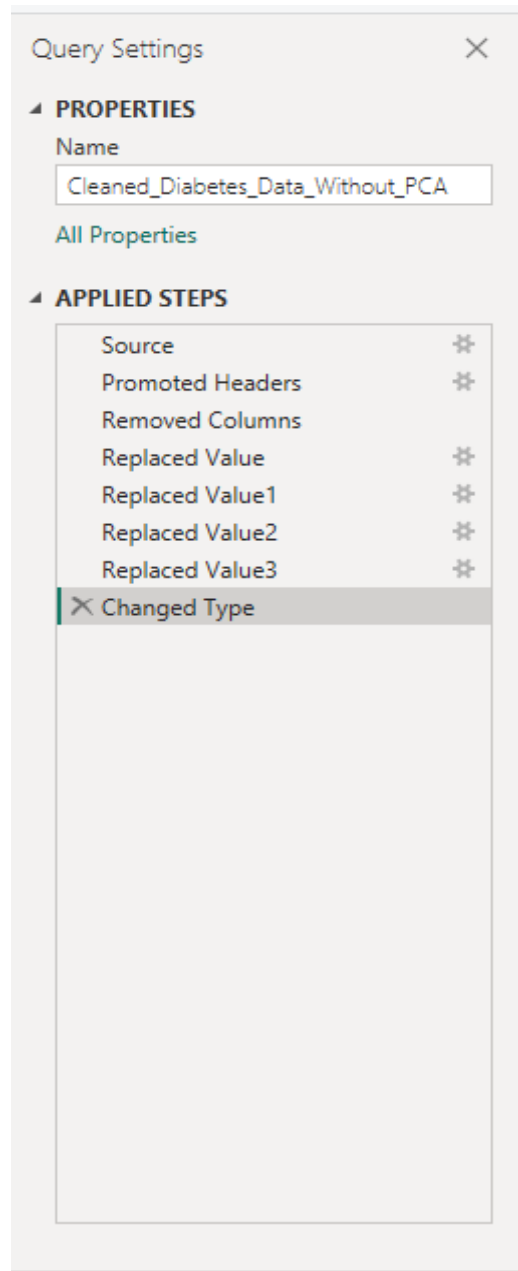


Figure 66: Applied Steps

### 3.11 Data Preprocessing in Power BI

The preprocessing of the dataset in Power BI was comprehensive, incorporating a range of techniques such as data modeling, advanced DAX calculations, value replacement, and the elimination of redundant columns, among other essential tasks. Notably, for the purposes of the Power BI analysis, the target variable was strategically configured to distinguish between readmission statuses, with "Yes" indicating a patient was readmitted and "No" indicating no readmission. Below shows some of the preprocessing taken for columns with the same values and would not offer any insight into the analysis. These columns are nateglinide', 'acetoexamide', 'miglitol', 'examide', 'citoglipton', 'glyburide-metformin', 'glipizide-metformin', 'glimepiride-pioglitazone', 'metformin-rosiglitazone', and 'metformin-pioglitazone. Additionally, one



column 'Peer Group' was created to be able to categorise group of people. These are all shown in figure 61 and 22 respectively.

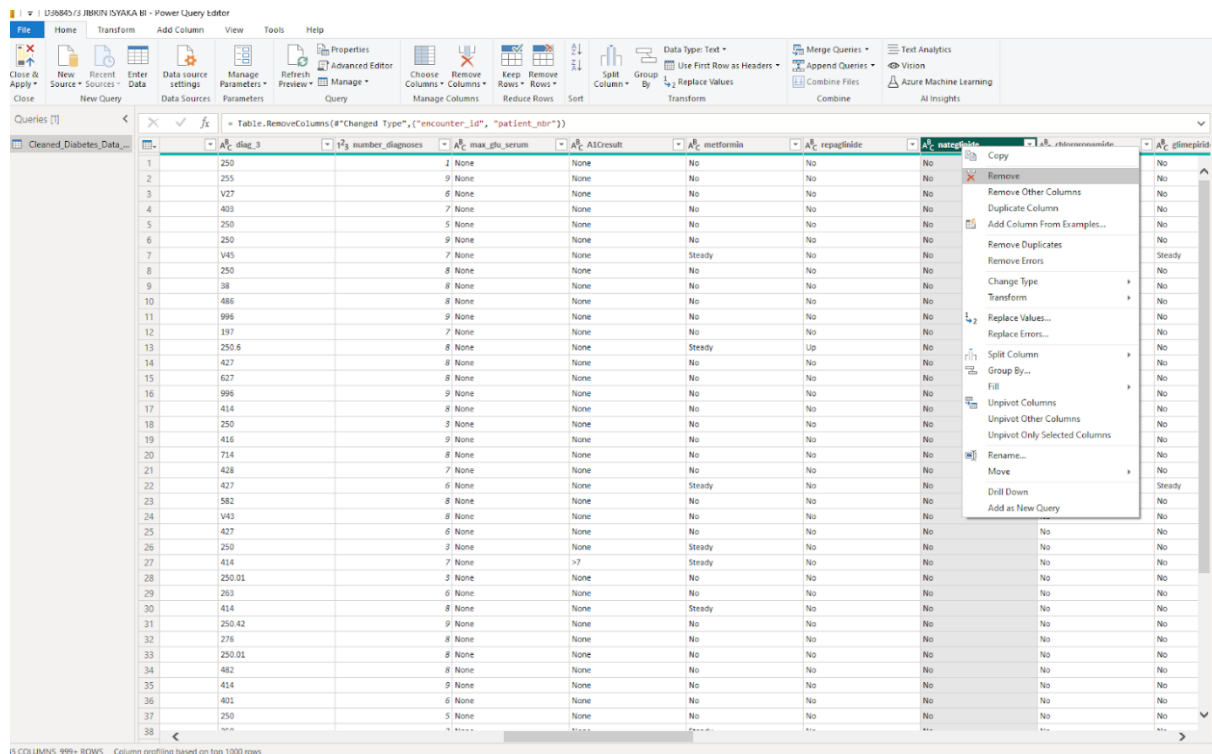


Figure 67: Data Pre-processing/Removing Redundant columns

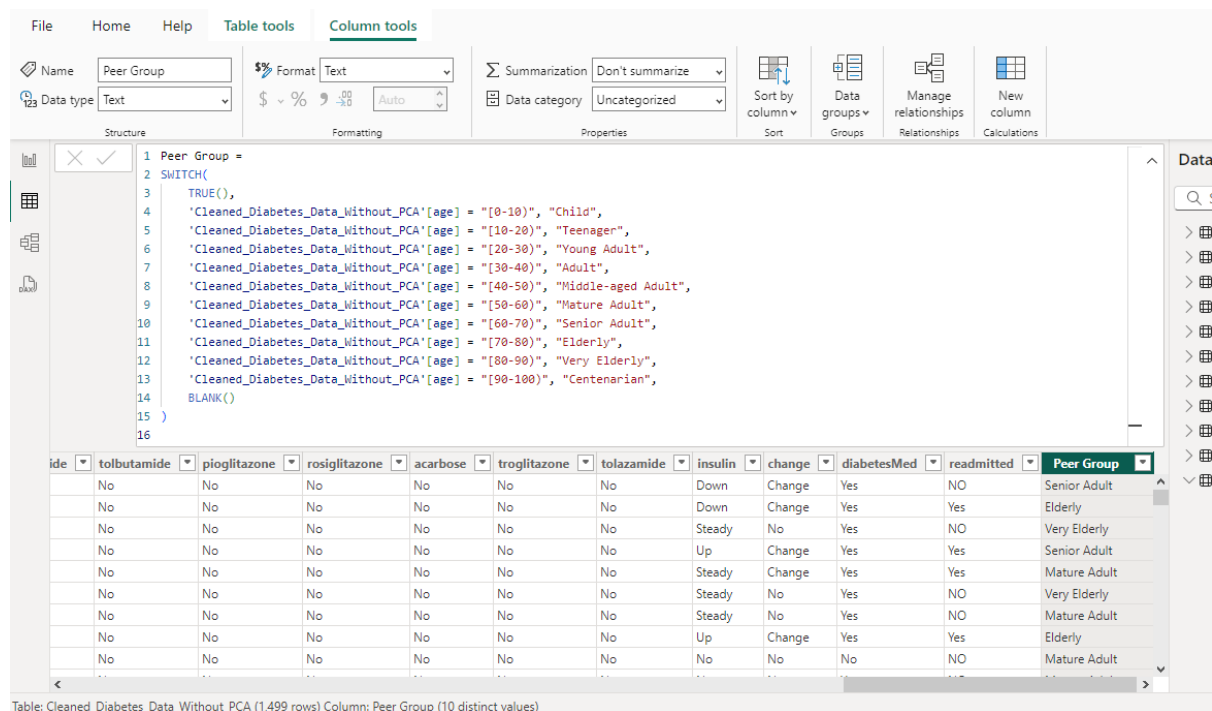


Figure 68: Data Preprocessing/creating New Column

### 3.12 Data Modelling

In Power BI, a Star Schema design was implemented for data modeling, featuring a central 'fact table' ('Cleaned\_Diabetes\_Data\_Without\_PCA') connected to multiple 'dimension' tables. This design was selected to enhance storage optimization, leading

to improved performance and more efficient data processing within the platform. Below are the Fact Table and the Dimension Tables:

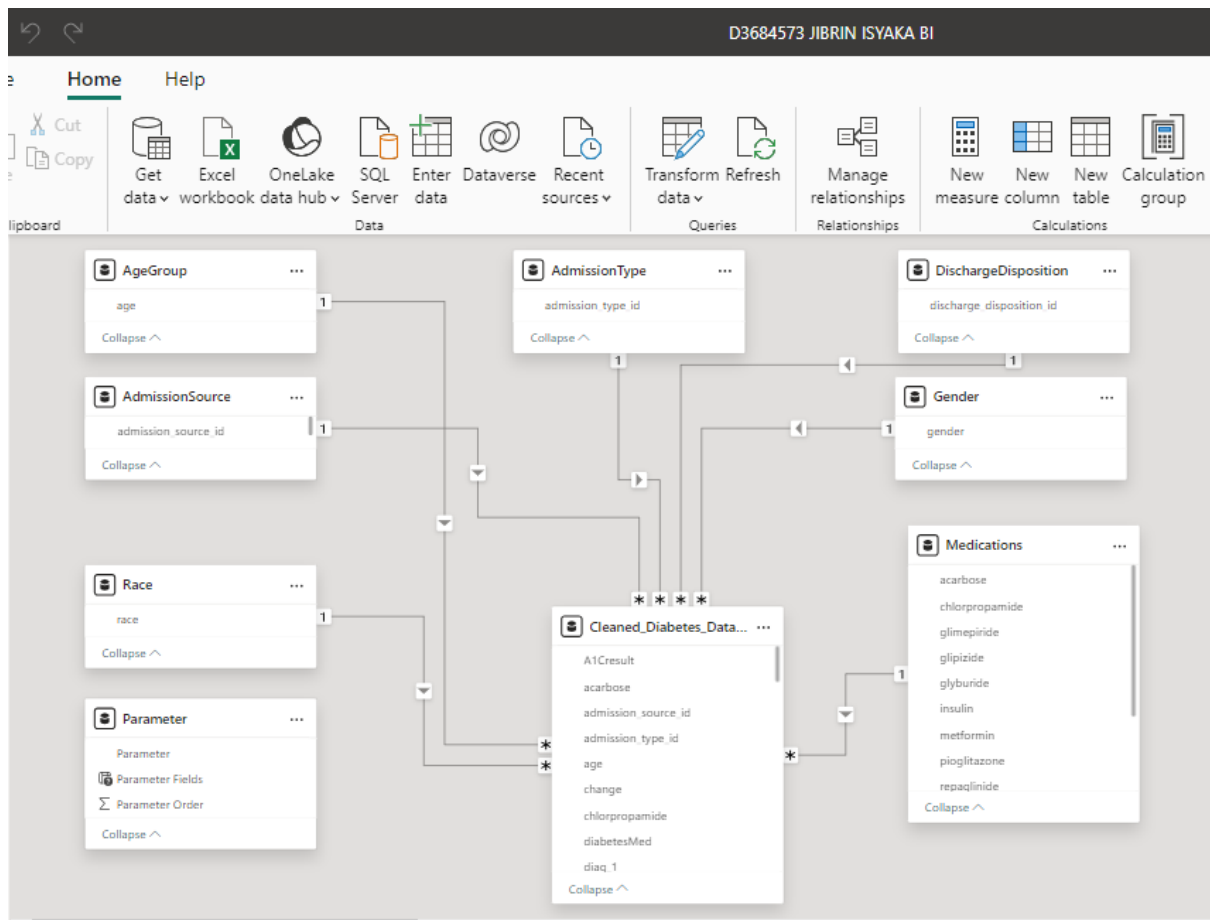


Figure 69: Modelling

Additionally, The 5 ML results were later imported as CSV files and pre-processed which later appear in the Data Pane, see figure 64.

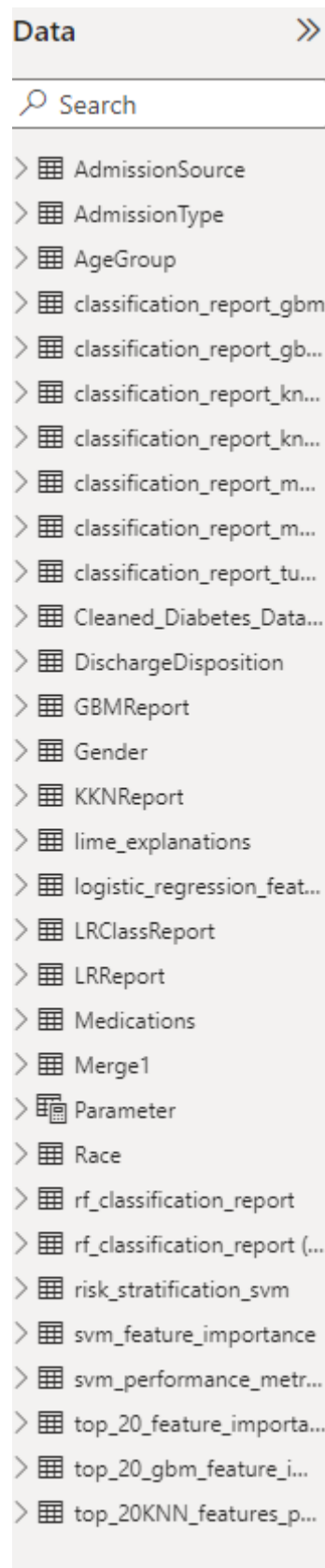


Figure 70: Modelling with ML results imported

### 3.13 Dax and Measures

An approach was employed to develop DAX columns and define measures, ensuring consistency and precision throughout the Power BI reports. This idea, involved the

careful selection of appropriate DAX functions, rigorous testing and validation of formulas, and the incorporation of dynamic visualization techniques to effectively communicate insights. The additional DAX column created in Power BI, with custom measures indicated by a calculator icon, specifically designed to elevate the analysis and achieve the project's objectives.

### 3.14 Data Visualisation

Dashboard organization (see Figure 65) was carefully designed to ensure consistency and facilitate seamless navigation across the application. This thoughtful arrangement enhances users' and stakeholders' ability to grasp and interpret the analytics with clarity.



Figure 71: Dashboard Organisation

## 4. IMPLETEMENTATION, RESUTS AND DISCUSSION

### 4.1 Implementation

The implementation of the proposed model began with the meticulous setup of all required libraries, modules, and functions within the Jupyter notebook environment. Data was then imported into the program using the Pandas library, setting the foundation for subsequent evaluation.

Model performance was evaluated using a comprehensive set of metrics, including accuracy, precision, recall, F1-score, confusion matrix, actual versus predicted data, risk stratification, feature importance, LIME, and ROC AUC. The most significant features identified for the best-performing algorithm were then leveraged for advanced visualization in Power BI. After resolving issues related to missing data, the cleaned dataset was saved and subsequently utilized in Power BI for detailed analytics. The following sections of this chapter present the results achieved at each stage of execution, offering a thorough overview of the model's performance and insights.

### 4.2 Model Performance Results and Discussion

This study implemented five distinct classification models to accurately predict diabetic hospital readmission, categorized into no readmission (0), readmission within 30 days (1), and readmission after 30 days (2). The performance of these models was rigorously evaluated using a comprehensive set of metrics, including accuracy (ACC), precision (PREC), recall, F1-score, confusion matrix, feature importance, risk stratification, LIME, and ROC curve analysis. To address the challenge of high dimensionality, Principal Component Analysis (PCA) was applied, while the SMOTE technique was used to correct for class imbalance in the dataset. Each model underwent meticulous hyperparameter tuning to maximize performance, with the optimal settings for all experimental models detailed in Table 2.

Table II: Hyperparameters Used for Different Classifiers

Hyperparameter Used		
Model Classifiers	With PCA	Without PCA
Logistic Regression	C = 1 penalty = l2 max_iter = 1000	C: 1 penalty: l2 max_iter: 1000
KNN	n_neighbors = 11 weights = distance	n_neighbors: 11 weights: distance
GBM	n_estimators: 100 learning_rate: 0.1 max_depth: 3	n_estimators: 100 learning_rate: 0.1 max_depth: 3

Random Forest	n_estimators: 100 max_depth: 10 min_samples_split: 5 min_samples_leaf: 2 max_features: sqrt	n_estimators: 100 max_depth: 10 min_samples_split: 5 min_samples_leaf: 2 max_features: sqrt
SVM	C: 10 gamma: 0.1 kernel: rbf	C: 10 gamma: 0.1 kernel: rbf

### 4.3 Model Performance With PCA

After applying Principal Component Analysis (PCA) to reduce dimensionality, the performance of various classification models was evaluated. The Logistic Regression model, tuned with optimal hyperparameters, achieved an accuracy of 41% and an F1-score of 0.34. Despite its decent precision and recall for Class 0 and Class 1, its performance on Class 2 was notably weaker. The K-Nearest Neighbors (KNN) model showed improvement with an accuracy of 49% and an F1-score of 0.48, particularly enhancing its performance on Class 1, though it still struggled with Class 2. The Gradient Boosting Machine (GBM) model achieved an accuracy of 51% and an F1-score of 0.51, demonstrating a balanced performance across classes with overall improved metrics. The Random Forest classifier excelled with an accuracy of 54% and an F1-score of 0.54, showcasing superior precision and recall across all classes. This model emerged as the most effective among those tested, reflecting the highest overall performance after dimensionality reduction through PCA.

#### 4.3.1 Confusion matrix

The confusion matrix analysis for the best model in predicting diabetic hospital readmission reveals varying levels of performance across the three classes: no readmission (Class 0), readmission within 30 days (Class 1), and readmission after more than 30 days (Class 2). The model demonstrates strong accuracy in predicting readmissions within 30 days, correctly identifying 99 cases, though it still confuses some instances with the other classes. For no readmission, the model also performs reasonably well, with 80 correct predictions, but it struggles with a significant number of false negatives, where patients readmitted after more than 30 days are misclassified. The model shows the greatest difficulty in accurately predicting readmissions after more than 30 days, with considerable misclassifications between

this and the other two categories. This is shown below:

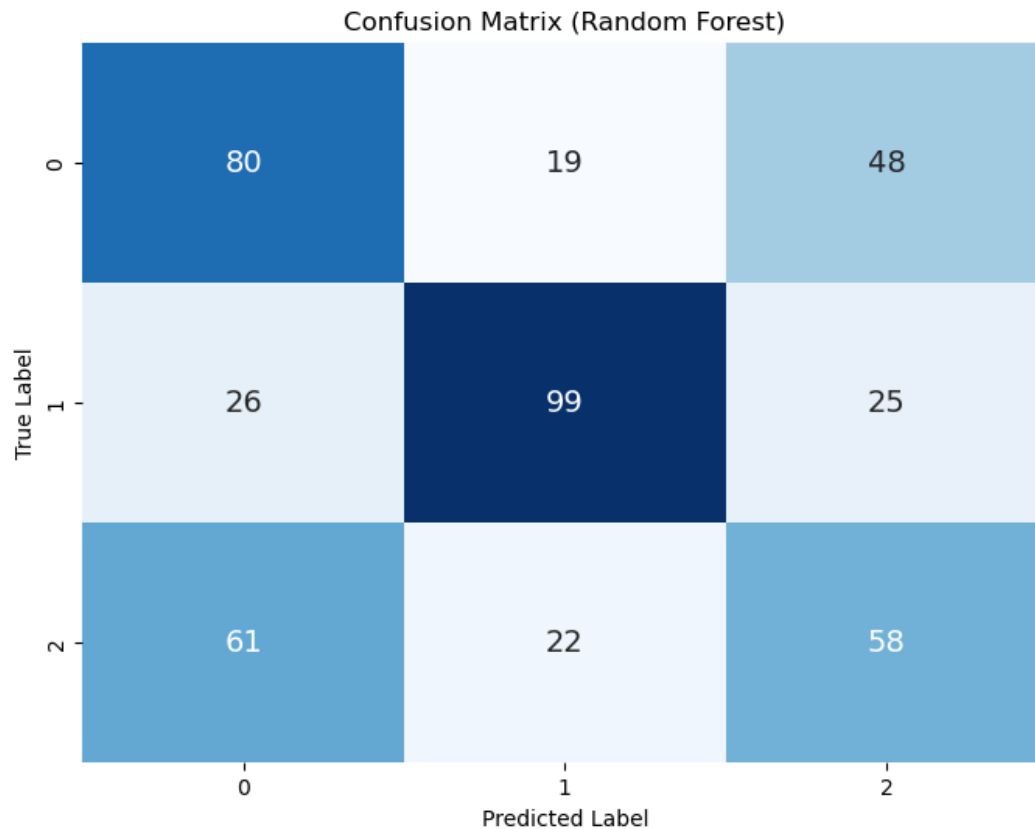


Figure 72: Confusion Matrix for Random Forest

#### 4.3.2 Feature Importance

The analysis of the feature importances highlights that Feature\_1 is the most influential feature in the model, with an importance score of 0.289898. This is followed by Feature\_2 at 0.252927, Feature\_3 at 0.234021, and Feature\_4 at 0.223153. These features are crucial in determining the model's predictions, with Feature\_1 having the most significant impact. The close importance scores of Feature\_2, Feature\_3, and Feature\_4 indicate that they also play substantial roles, though to a slightly lesser extent. The distribution of these scores reflects the varying degrees of influence each feature has on the model's decision-making process.

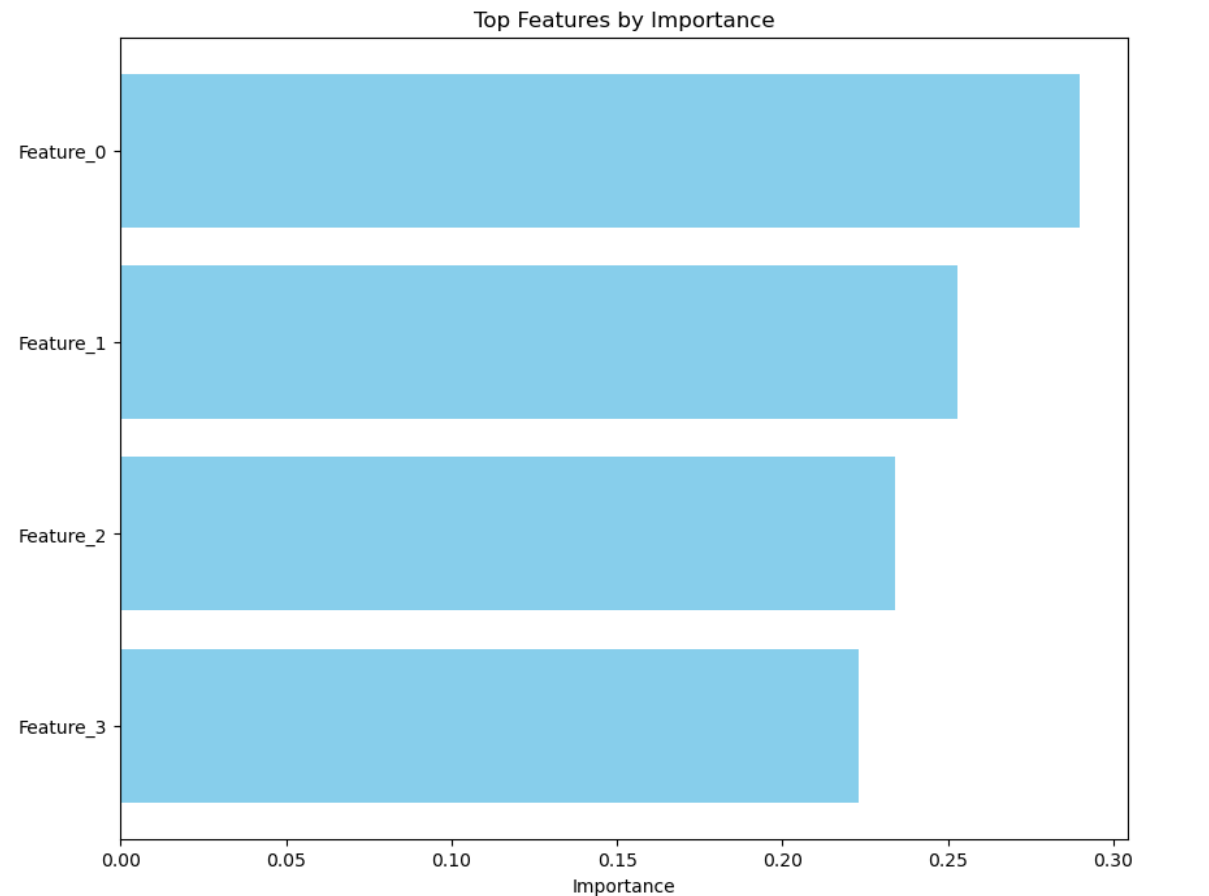


Figure 73: Random Forest Feature Importance

#### 4.3.3 Risk Stratification

The Risk Stratification provides an overview of model performance across different risk categories. For instance, instance 1846, which was correctly predicted as high risk with a probability of 0.690, falls into the Medium risk category. Similarly, instance 1316, correctly predicted as low risk with a probability of 0.099, is classified as Low risk. Instance 764, with a predicted probability of 0.511, is categorized as Medium risk, reflecting its moderate risk level. Instance 1550, accurately predicted as high risk with a high probability of 0.736, is classified as High risk, indicating a strong confidence in its high-risk status. Lastly, instance 993, predicted with a probability of 0.344, is also categorized as Medium risk. This is shown below:



Risk Stratification DataFrame:

	Actual	Predicted	Predicted_Probability	Risk_Category
1846	1	1	0.690455	Medium
1316	0	0	0.099448	Low
764	2	1	0.511526	Medium
1550	1	1	0.735776	High
993	2	2	0.343995	Medium

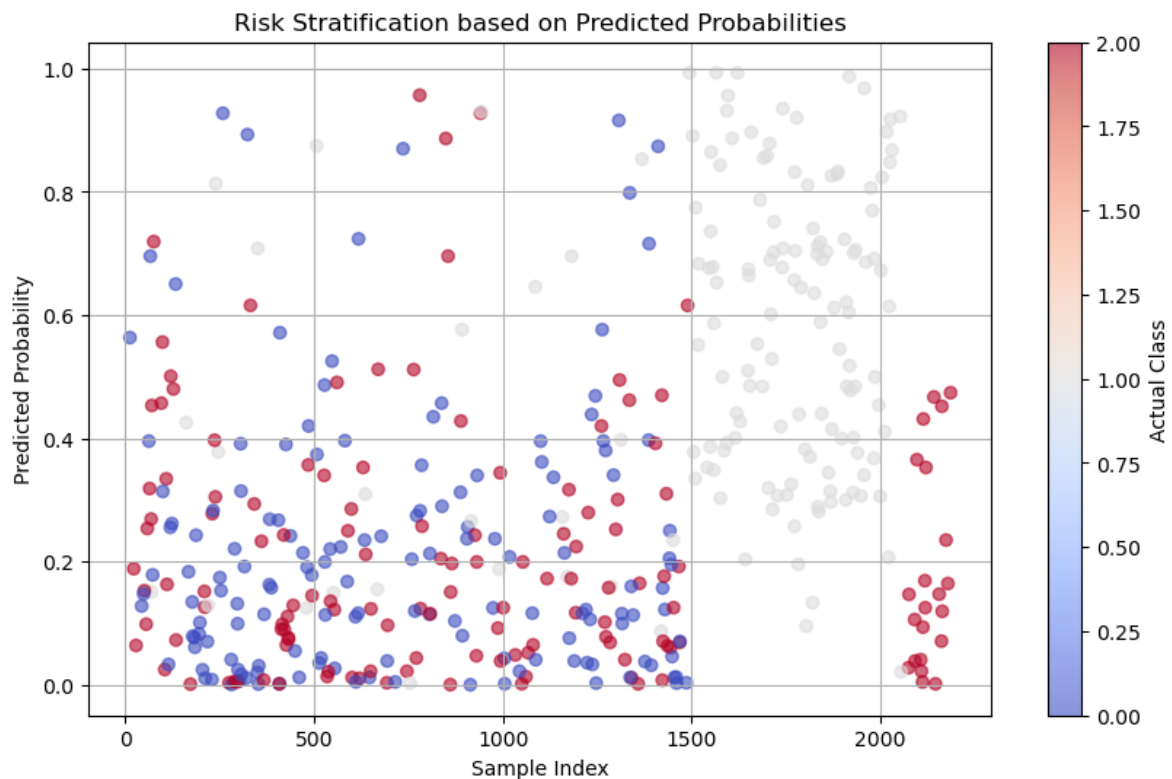


Figure 74: Risk Stratification

#### 4.3.4 LIME

In our LIME analysis of various instances, we identified several key features influencing predictions. For Instance 11, Feature 9 and Feature 10 were most significant. In Instance 12, Feature 9 and Feature 18 were crucial, with Feature 2 and Feature 6 having minor negative impacts. Instance 15 highlighted Feature 1 and Feature 12, while Instance 16 emphasized Feature 1 and Feature 18 with negative effects from Feature 4 and Feature 0. Instance 2 also showed the prominence of Feature 1 and Feature 10. Instance 3 highlighted Feature 9 and Feature 1, with minor negative influence from Feature 17. Instance 4's important features included Feature 11, Feature 9, and Feature 18. Instance 5 and Instance 6 both underscored the importance of Feature 9 and Feature 18, with Instance 6 also noting negative impacts from Feature 4 and Feature 7. This is shown below:

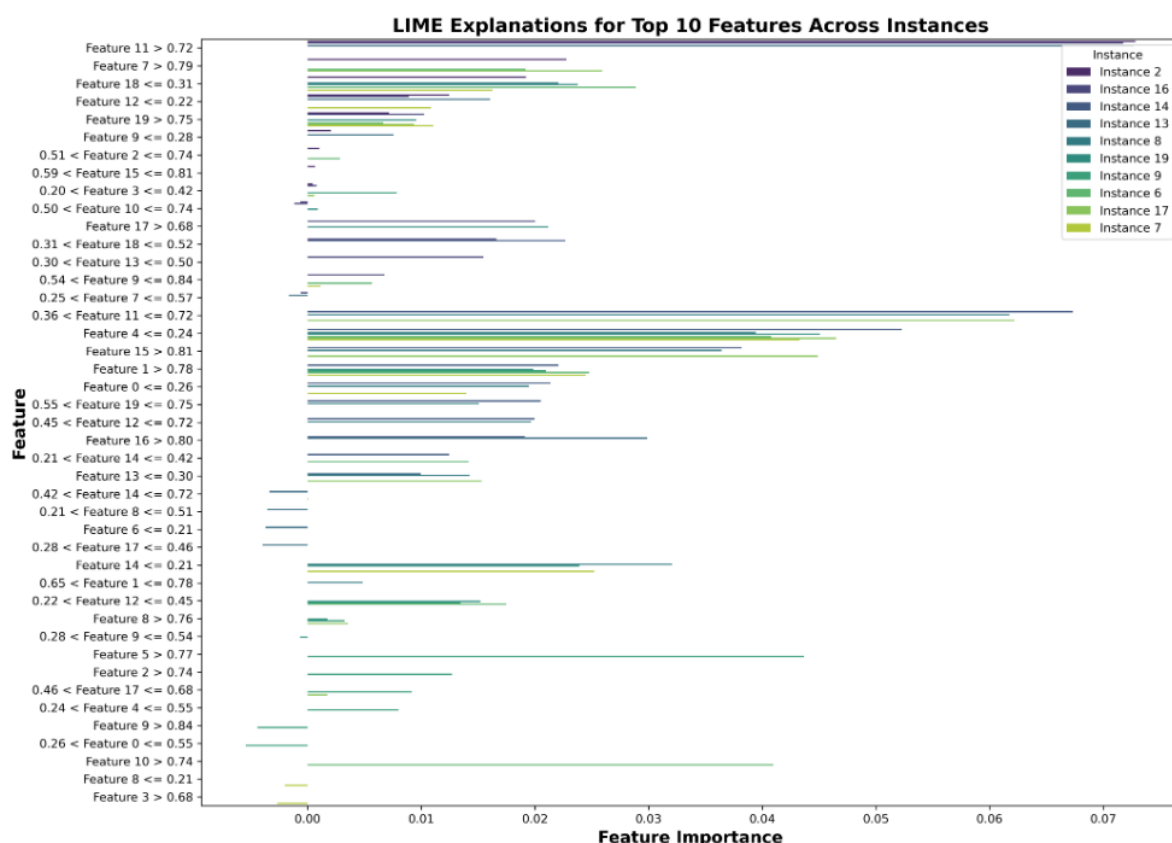


Figure 75: LIME Explanations

#### 4.4 Model Performance Without PCA

In the initial phase of the analysis, the performance of various classification models was evaluated on the imbalanced dataset. The Logistic Regression model, optimized with the best hyperparameters, achieved an accuracy of 69% and an F1-score of 0.68, demonstrating strong performance, particularly in Class 1. The K-Nearest Neighbors (KNN) model, after grid search, attained an accuracy of 56% and an F1-score of 0.51, with notable challenges in Class 0. The Gradient Boosting Machine (GBM) model performed with an accuracy of 67% and an F1-score of 0.67, providing balanced results across all classes. The Random Forest classifier excelled with an accuracy of 71% and an F1-score of 0.71, showing superior performance in all metrics. The Support Vector Machine (SVM) model also demonstrated high effectiveness, with an accuracy of 74% and an F1-score of 0.74, making it the top performer among the models evaluated.

##### 4.4.1 Confusion Matrix

The confusion matrix for the tuned Support Vector Machine (SVM) model illustrates its performance across the three classes in the diabetic hospital readmission dataset. The model correctly identified 90 cases as non-readmitted (Class 0), though it incorrectly classified 1 such case as readmitted within 30 days (Class 1) and 56 as readmitted after 30 days (Class 2). This indicates that while the model is proficient at

detecting non-readmitted cases, it occasionally misclassifies them as readmitted. For short-term readmissions (Class 1), the model performed exceptionally well, correctly identifying 147 instances with minimal errors, misclassifying only 3 as non-readmitted and none as long-term readmissions (Class 2). This reflects a strong capability in detecting short-term readmissions. Regarding long-term readmissions (Class 2), the model accurately identified 85 cases but confused 54 long-term readmissions with non-readmitted cases and 2 with short-term readmissions. While the model shows reasonable performance in predicting long-term readmissions, there remains a significant level of confusion with non-readmitted cases. This shown below:

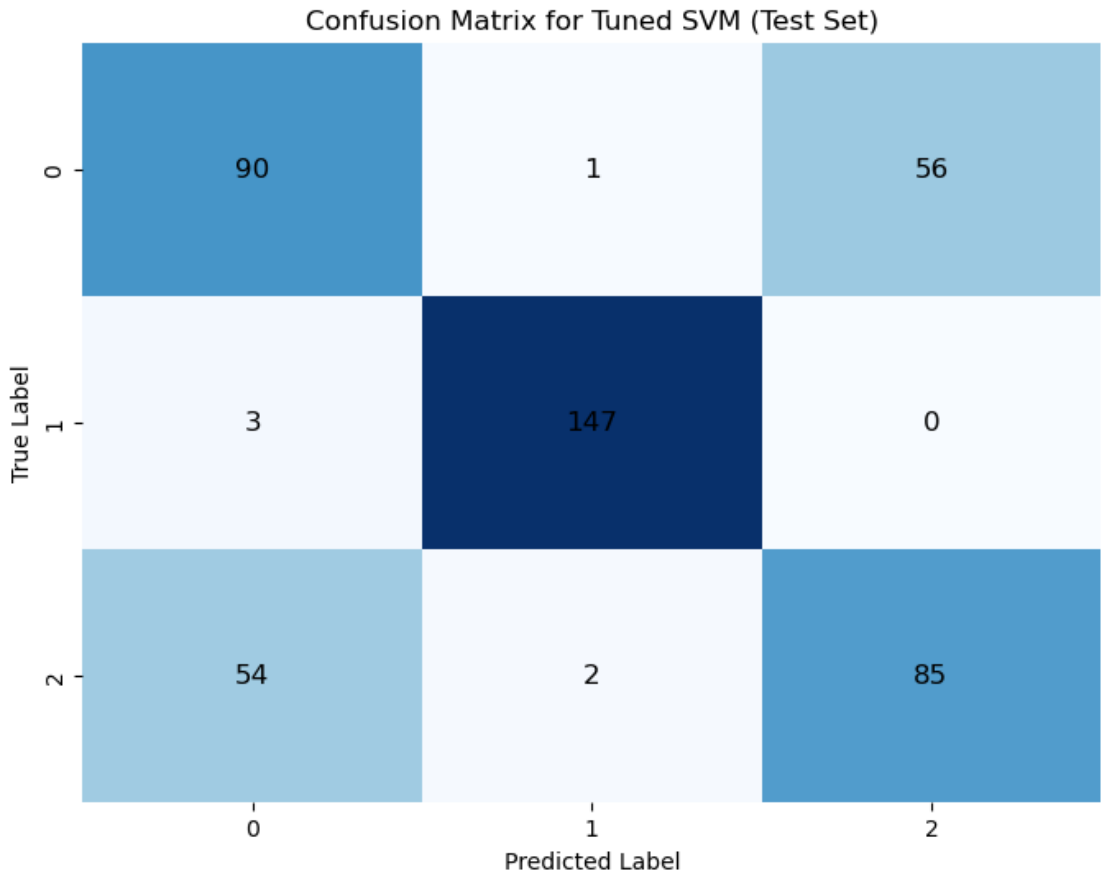


Figure 76: Confusion Matrix for SVM

4.4.2 Feature Importance

The permutation importance analysis for the Support Vector Machine (SVM) model reveals that features related to the number of lab procedures, medications, and diagnoses, as well as time in hospital, are the most influential in predicting patient readmissions. num\_lab\_procedures and num\_medications are the top contributors, indicating their significant role in the model’s decision-making process. Other important features include procedural and admission-related variables, such as num\_procedures and admission\_source\_id. Demographic factors like gender and age, along with clinical variables such as insulin usage and A1C result, have lower but

notable impacts. Overall, the model relies heavily on procedural and admission details, with demographic and clinical factors providing additional, albeit less critical, insights. This shown below:

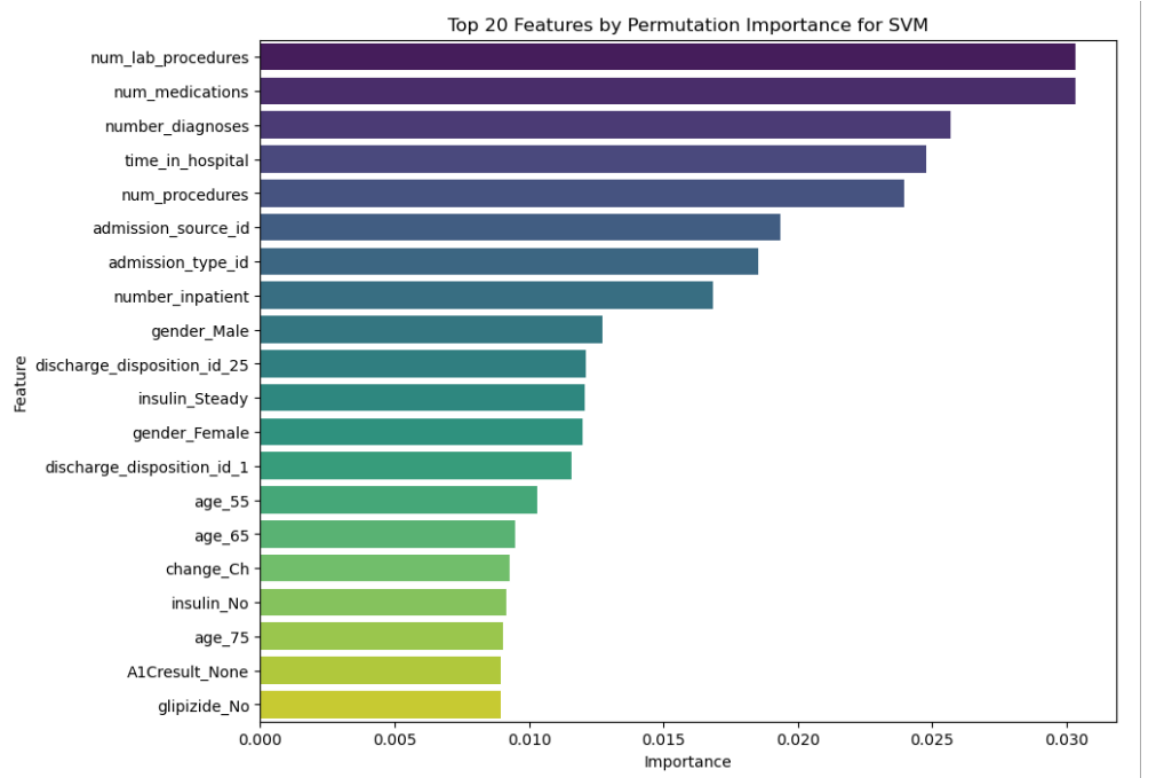


Figure 77: SVM Feature Importance

#### 4.4.3 Risk Stratification

The Risk Stratification DataFrame provides an insightful overview of how the model categorizes patients based on the predicted risk of readmission. The DataFrame showcases several cases where both the actual and predicted risk categories align, indicating accurate risk assessment by the model. For instance, patient entries with Actual and Predicted values of 1, 0, or 2 are consistently classified as "Low Risk," with high Max\_Decision\_Function\_Value scores reflecting robust model confidence in these predictions. This uniformity suggests that the model is effectively identifying low-risk patients across different classes, though this analysis is limited to instances

where risk is categorized as low. This is shown below:

Risk Stratification DataFrame:

	Actual	Predicted	Max_Decision_Function_Value	Risk_Category
1846	1	1	2.221111	Low Risk
1316	0	0	2.154857	Low Risk
764	2	2	2.228087	Low Risk
1550	1	1	2.243663	Low Risk
993	2	2	2.168173	Low Risk

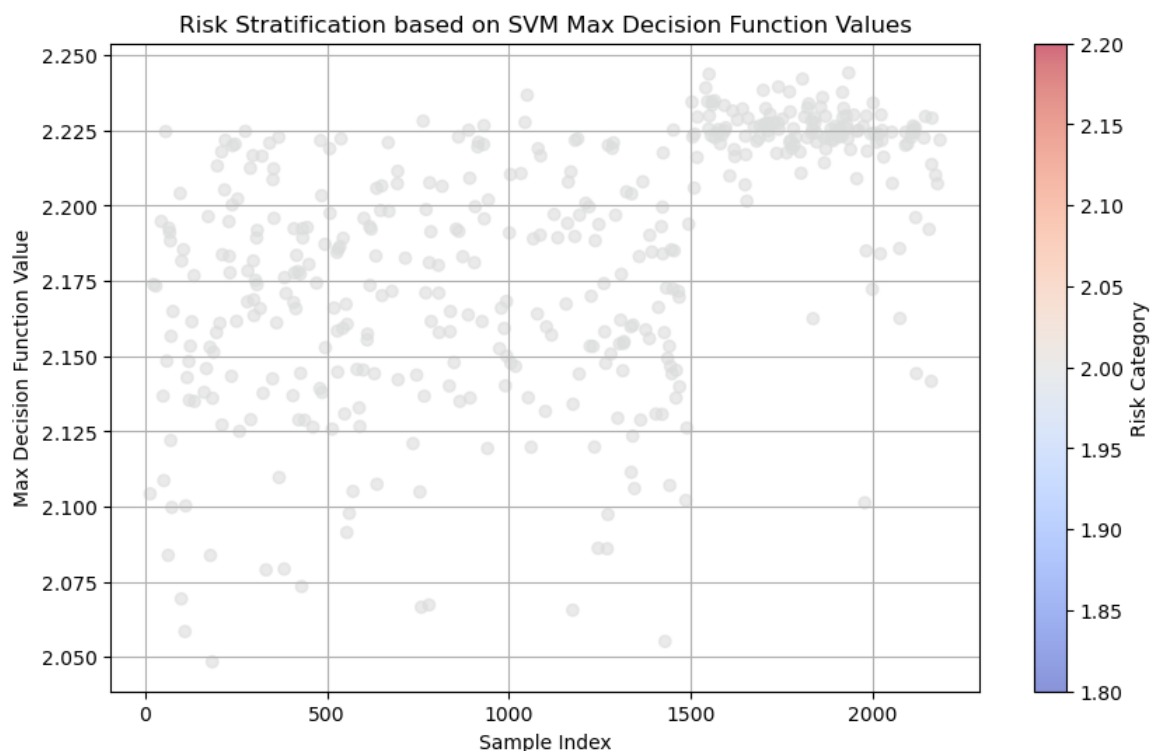


Figure 78: SVM Risk Stratification

#### 4.4.4 LIME Explanations

Analysing the LIME explanations for the top 10 features of 10 instances classified as Label 1 by the tuned SVM model reveals several key insights. Notably, certain features such as `diag_3_620`, `diag_2_435`, and `diag_1_555` consistently emerge with high importance across multiple instances, indicating their significant role in predicting Label 1. This recurring prominence suggests that these features are central to the model's decision-making process for this class, highlighting their critical role in distinguishing Label 1 from other classes. Additionally, features like `diag_1_250.93`, `diag_3_704`, and `diag_2_714` frequently appear, suggesting that their interactions with other features contribute notably to the model's predictions. A trend observed is that features often have a threshold value (e.g.,  $\leq 0.00$ ) which notably impacts their importance score, implying that the model may heavily rely on the presence or absence of specific diagnostic codes or categorical variables. This threshold sensitivity indicates that crossing certain feature value thresholds could significantly influence the model's predictions. Furthermore, while some features consistently appear among the top 10 across various instances, there is noticeable variation in the top features from

one instance to another. This variation suggests that although certain features are generally important, their specific impact can vary depending on individual instances, with some instances showing clusters of similar diagnostic codes that may be indicative of patterns associated with Label 1. This shown below:

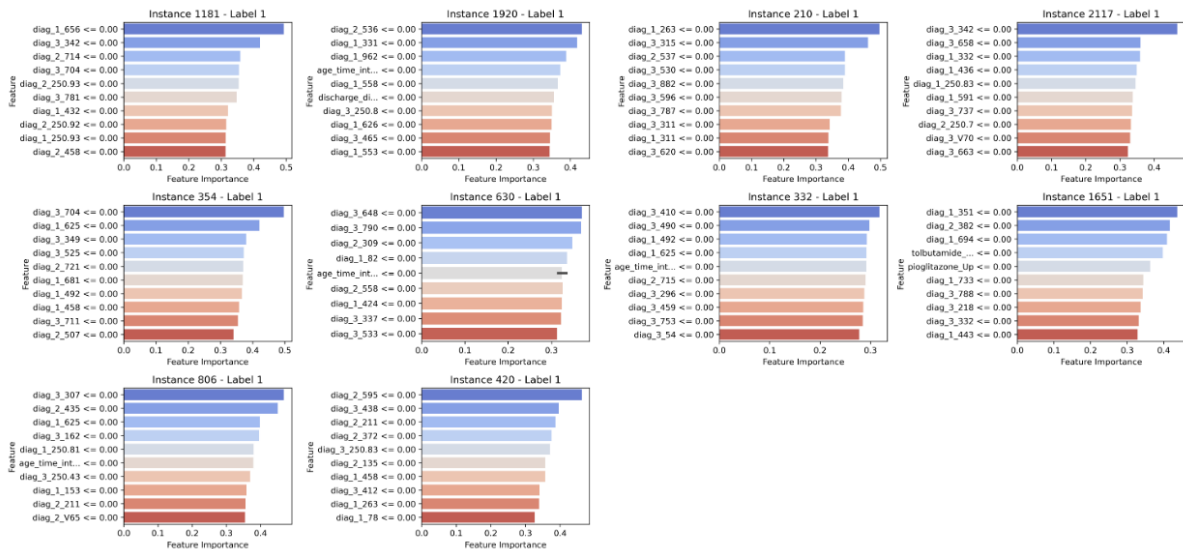


Figure 79: LIME Explanations for SVM

#### 4.4.5 ROC AUC

The model performance analysis indicates that the Support Vector Machine (SVM) is the top performer, achieving perfect classification for Class 1 (AUC = 1.00) and strong results for Classes 0 and 2. Random Forest also shows impressive performance, especially for Class 1 (AUC = 0.99), and maintains strong results across other classes. Logistic Regression performs well with Class 1 (AUC = 0.94) but is less effective for Classes 0 (AUC = 0.72) and 2 (AUC = 0.75). K-Nearest Neighbours (KNN) mirrors Class 1's strong performance but struggles with Class 2 (AUC = 0.79). Gradient Boosting provides balanced results but does not exceed Random Forest's performance. Class 1 is consistently well-differentiated, whereas Class 2 shows the most variation, particularly with KNN. Overall, SVM is the most effective model, with Random Forest also performing robustly across classes.

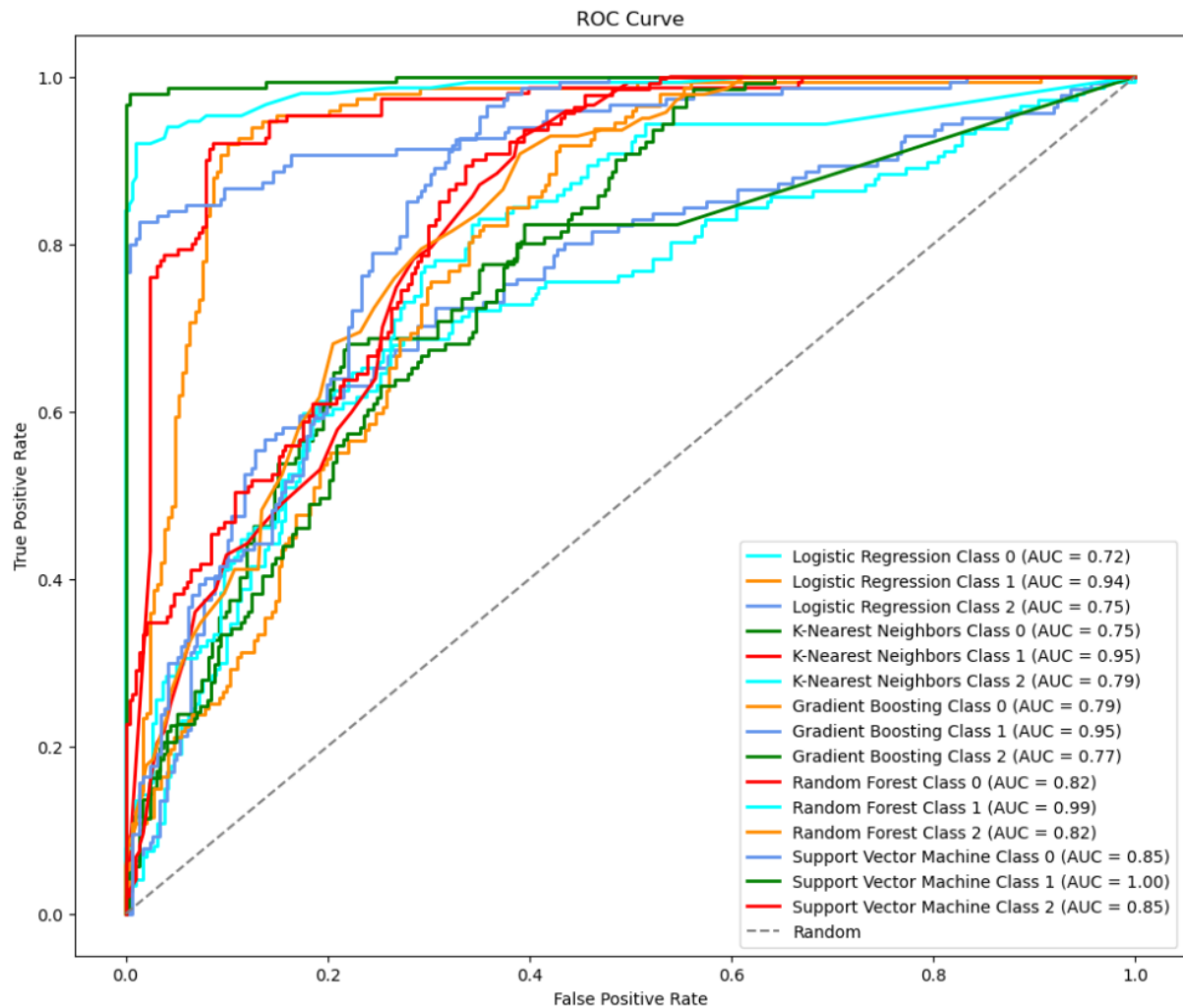


Figure 80: ROC AUC

#### 4.4.6 Performance Comparison

The comparative analysis of classification models with and without PCA reveals significant differences in their performance. With PCA, the dimensionality reduction technique was applied to address the challenges of high-dimensional data. The Logistic Regression model's performance improved slightly, achieving an accuracy of 41% and an F1-score of 0.34. The KNN model showed enhanced results with an accuracy of 49% and an F1-score of 0.48. The Gradient Boosting Machine (GBM) demonstrated a more significant boost with an accuracy of 51% and an F1-score of 0.51. The Random Forest classifier also benefitted from PCA, with an accuracy of 54% and an F1-score of 0.54. The SVM model, while still performing well, achieved an accuracy of 49% and an F1-score of 0.48.

Without PCA, the models were evaluated based on their raw feature sets. The Logistic Regression model achieved an accuracy of 69% and an F1-score of 0.68. The KNN model, while optimized, showed lower performance with an accuracy of 56% and an F1-score of 0.51. The GBM model yielded an accuracy of 67% and an F1-score of 0.67, providing balanced results across the classes. The RF classifier performed

notably well with an accuracy of 71% and an F1-score of 0.71, and the SVM emerged as the top performer with an accuracy of 74% and an F1-score of 0.74. Overall, applying PCA led to varied results across models, with some showing improved performance in accuracy and F1-scores, while others exhibited a decrease.

The classification results are illustrated as shown below:

Table III: Model Performance Comparison

Model	Accuracy (PCA)	F1-Score (PCA)	Accuracy (Without PCA)	F1-Score (Without PCA)
LR	0.41	0.34	0.69	0.68
KNN	0.49	0.48	0.56	0.51
GBM	0.51	0.51	0.67	0.67
RF	0.54	0.54	0.71	0.71
SVM	0.49	0.49	0.74	0.74

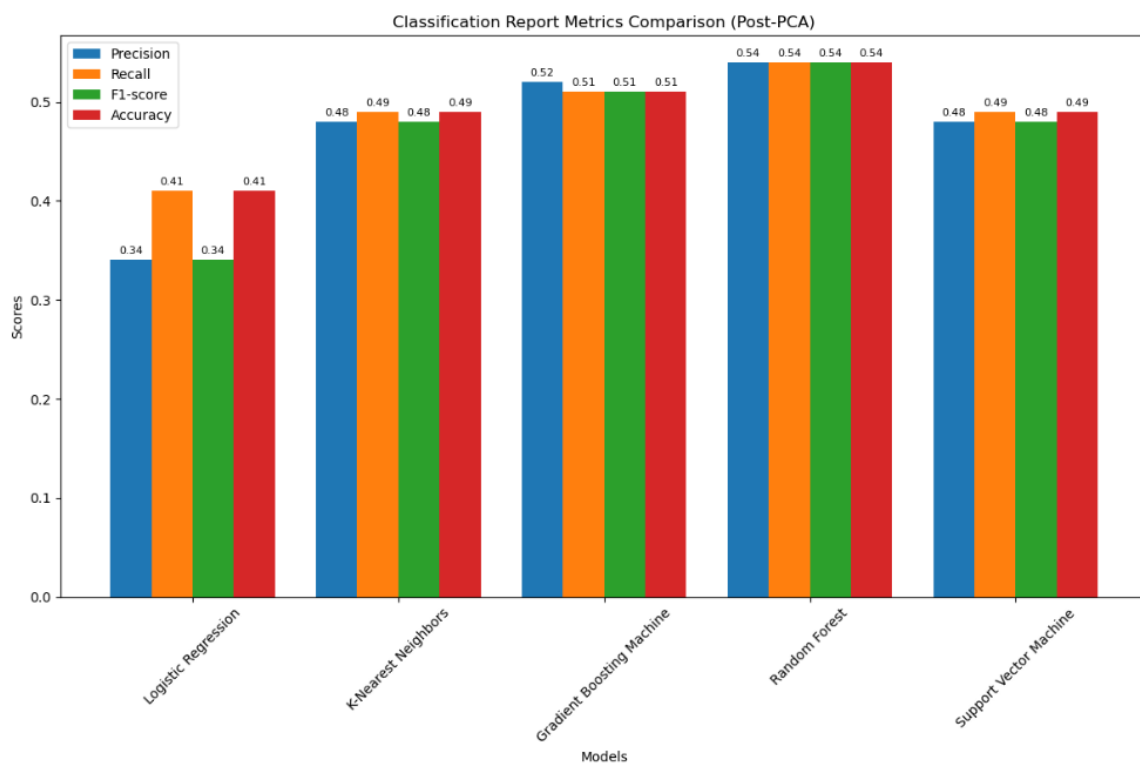


Figure 81: Comparison of Classification Report With PCA



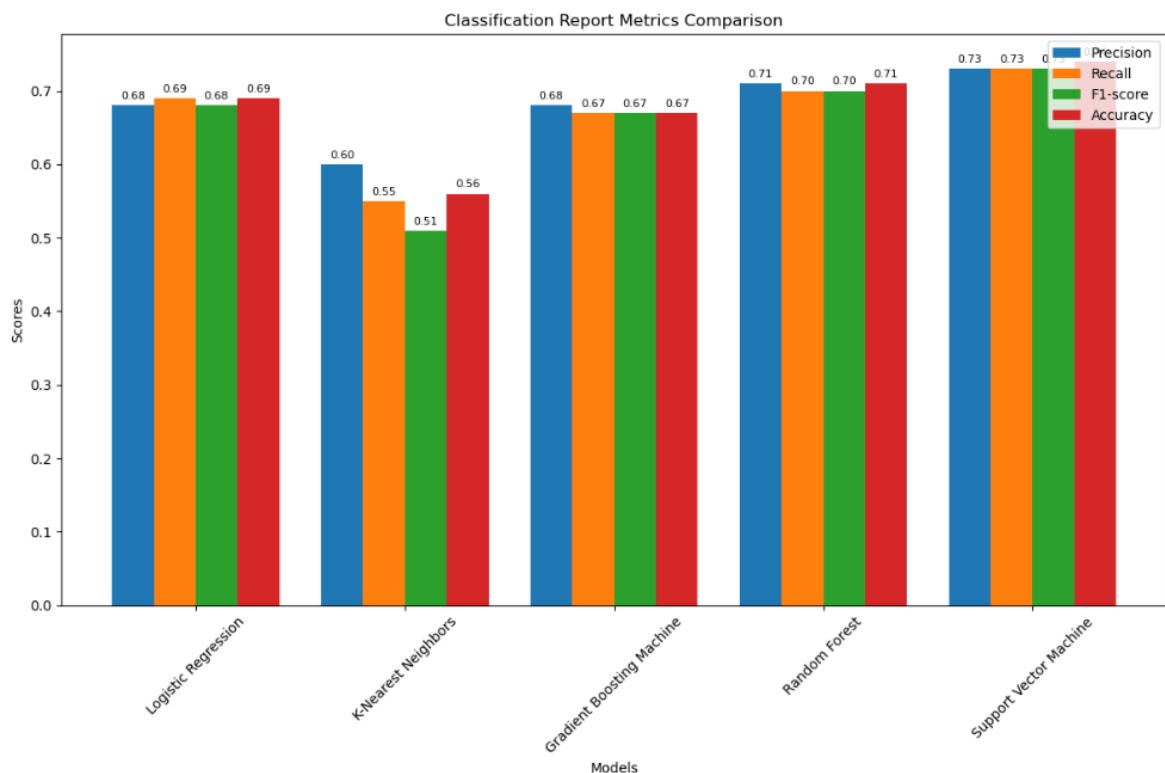


Figure 82: Comparison of Classification Report Without PCA

#### 4.5 Power BI

The dashboards deliver an in-depth analysis of diabetic hospital readmissions and related metrics from various angles. The main dashboard offers a high-level view of essential metrics, setting the stage for detailed exploration through additional dashboards. This structure is intended to provide actionable insights that support informed decision-making and planning.

The "Discharge Disposition Based on Admission and Race" visualization is an interactive decomposition tree that displays readmission data segmented by variables such as race, age, admission type, and discharge status. This dynamic tool updates all related dashboard metrics in response to selections made through the 'patient identification number slicer.' By choosing a specific patient, users can view real-time changes across all visualizations, which helps them analyze how individual patient data affects overall readmission rates. This feature also facilitates a more in-depth examination of medication use and changes, offering crucial insights for decision-making and strategic planning. It allows for a nuanced understanding of diabetic hospital readmissions by examining factors such as demographics, race, and medication variables.

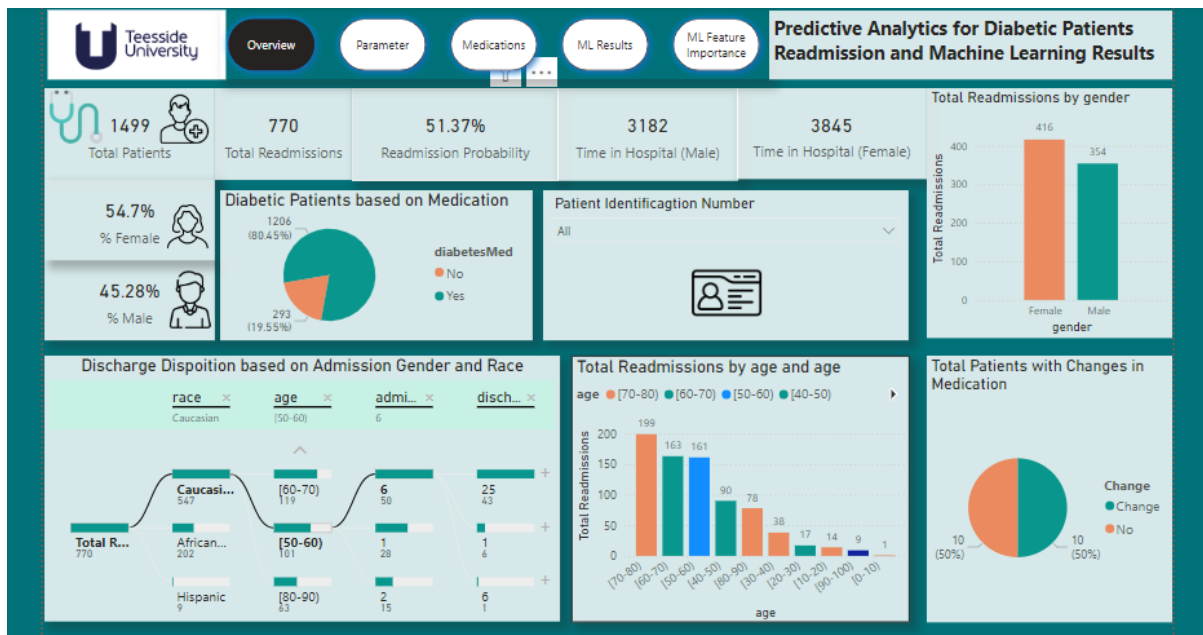


Figure 83: Power BI Overview

The parameter dashboard below features a heatmap where DAX measures were utilized to calculate metrics such as 'Average Time in Hospital,' 'Average Number of Medications per Age Group,' and 'Readmission Rate' segmented by race. This interactive heatmap dynamically integrates with other dashboard metrics, enabling users to analyse patterns in patient insulin usage and hospitalization time across different demographic groups. By providing a clear visualization of these key indicators, the dashboard supports healthcare professionals in identifying trends and disparities, ultimately facilitating more informed decisions regarding patient care and resource allocation. This capability enhances the medical profession's ability to tailor interventions, improve patient outcomes, and address potential disparities in treatment based on demographic factors.

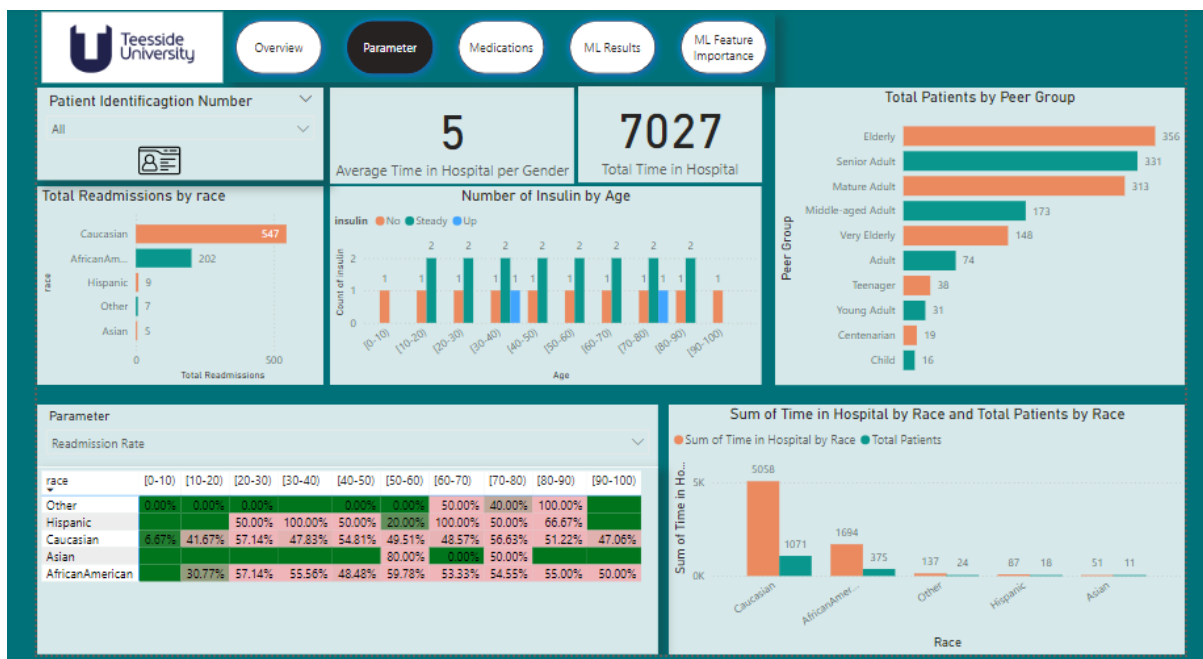


Figure 84: Parameter Dashboard

The medication dashboard provided below illustrates the medications prescribed to patients segmented by race which appears to be equal distribution. This table can be expanded to include additional details such as diagnoses, procedures, and lab tests, all categorized by race. Additionally, the dashboard features a risk stratification component derived from the top-performing machine learning model. This component classifies patients into risk categories as 0, 1, and 2 based on their predicted risk levels.

Analysing this detailed data allows healthcare professionals to uncover patterns in medication use and treatment efficacy among different demographic groups. This insight can reveal disparities in medication access and adherence, enabling the optimization of treatment protocols and the customization of interventions to address the needs of various patient populations. The risk stratification tool further enhances this analysis by predicting patient outcomes and identifying those at higher risk, which aids in prioritizing care and improving patient management. Class 1 appears to have 147 predicted risk stratification as against the actual 150. This suggest that class 1 have a higher rate of getting readmitted.

This comprehensive approach supports strategic planning and enhances decision-making within the medical field.



Figure 85: Medication Dashboard

The machine learning dashboard below effectively compares various ML models, highlighting SVM as the most accurate even when evaluated by macro averages. This comparison enables healthcare professionals to select the most reliable model for predicting patient readmission risks, facilitating more informed decision-making and strategic planning. The clear visualizations make it easy for stakeholders to understand model performance and apply these insights to optimize resource allocation and improve patient care. By leveraging the top-performing model, healthcare providers can enhance predictive accuracy, allocate resources more efficiently, and drive strategic improvements in patient management practices.



Figure 86: ML Results

The final dashboard presents ML feature importance alongside SVM LIME explanations, with a focus on SVM due to its superior performance compared to other models. The feature importance is filtered to the top 5, highlighting diagnoses, medications, and the number of medications, thereby offering clear insights into the key factors influencing patient outcomes. This emphasizes the importance for medical practitioners and stakeholders to identify the most critical variables driving patient readmission and treatment efficacy. By comprehending these visuals, they can make data-driven decisions to optimize treatment plans, improve patient care, and allocate resources more effectively, ultimately enhancing the quality of healthcare delivery.



Figure 87: ML Feature Importance

#### **4.6 Summary of Findings**

The comparative analysis of classification models for predicting diabetic hospital readmissions, both with and without PCA, reveals notable differences in performance. When PCA was applied, the models showed varied results: Logistic Regression achieved 41% accuracy and an F1-score of 0.34, KNN reached 49% accuracy and an F1-score of 0.48, GBM achieved 51% accuracy and an F1-score of 0.51, and Random Forest performed at 54% accuracy with an F1-score of 0.54. The SVM, while still effective, had an accuracy of 49% and an F1-score of 0.48 under PCA.

In contrast, without PCA, models performed better overall: Logistic Regression achieved 69% accuracy and an F1-score of 0.68, KNN showed 56% accuracy and an F1-score of 0.51, GBM reached 67% accuracy and an F1-score of 0.67, Random Forest excelled with 71% accuracy and an F1-score of 0.71, and SVM emerged as the top performer with 74% accuracy and an F1-score of 0.74. The application of PCA led to mixed results, improving performance for some models but decreasing effectiveness for others.

The Power BI dashboards provided a comprehensive analysis of diabetic hospital readmissions. The main dashboard highlighted key metrics and offered detailed insights through interactive visualizations. The "Discharge Disposition Based on Admission and Race" visualization allowed for real-time updates and in-depth examination of patient-specific data, facilitating a nuanced understanding of readmissions across demographics. The parameter dashboard, featuring a heatmap with DAX measures, enabled analysis of hospitalization time and medication use by age and race, aiding in identifying trends and disparities.

The medication dashboard displayed medication use across different races, with the potential to expand to include more details such as diagnoses and lab tests. The risk stratification tool, derived from the top-performing SVM model, classified patients into risk categories and highlighted potential disparities in medication access and adherence. The final dashboard emphasized the importance of feature analysis and model interpretability, supporting data-driven decision-making to optimize treatment plans and resource allocation.

Overall, these findings underscore the importance of selecting appropriate models and features for accurate predictions of readmissions and highlight the value of integrating diverse data sources and advanced analytics to improve healthcare delivery.

## 5. Discussions: Machine Learning and Power BI Findings

The comparative analysis of classification models for predicting diabetic hospital readmissions, both with and without Principal Component Analysis (PCA), reveals important findings. Models were tested to distinguish between no readmission, readmission within 30 days, and readmission after 30 days. The SVM model showed the highest accuracy at 74% and an F1-score of 0.74, while the Random Forest model also performed well with 71% accuracy and an F1-score of 0.71. PCA's impact was mixed: it slightly improved performance for RF and GBM but reduced effectiveness for KNN and SVM. Sharma et al.'s study demonstrated that Random Forest achieved the highest accuracy (94%) for predicting 30-day readmissions. The best PCA model excelled in predicting short-term readmissions but faced difficulties with long-term predictions. Feature importance analysis highlighted key features, and LIME analysis reinforced the value of raw feature models for practical insights.

Critically evaluating these results reveals important implications for healthcare practice. While models like SVM and Random Forest show promise for accurately predicting hospital readmissions, the inconsistent impact of PCA underscores the need for careful consideration when applying such techniques. The reduction in model effectiveness for long-term predictions, even with PCA, suggests that more advanced feature engineering or alternative techniques may be necessary to capture the nuances in patient data that influence readmissions.

The practical application of these findings is further enhanced by Power BI's dynamic visualization capabilities, which offer healthcare professionals an interactive platform to explore data in real-time. This tool is invaluable for predicting patient readmissions, helping medical professionals manage resources more effectively and make informed decisions. The ability to visualize key factors, such as feature importance and SVM LIME explanations, allows for a deeper understanding of what drives patient readmission, facilitating targeted interventions.

In the broader context of healthcare, these findings contribute to the ongoing efforts to reduce hospital readmissions, a critical issue given the high costs associated with repeat admissions, especially among diabetic patients. The insights derived from both machine learning models and Power BI analytics align with existing literature, reinforcing the importance of operational and clinical factors in predicting readmissions. However, this study also highlights the variability in model performance, particularly with PCA, and the need for continuous evaluation and adaptation of predictive models to ensure they meet the specific needs of healthcare settings.

In the context of existing literature, this study's findings on predicting diabetic hospital readmissions reveal several similarities and differences when compared to prior research. Howell et al. (2009) demonstrated the importance of operational factors alongside clinical variables in predicting readmissions, highlighting those features such as diagnosis codes and clinic occupancy rates were significant predictors (Howell et al., 2009). This aligns with our study's emphasis on feature importance, though our findings also indicate that models with PCA sometimes performed inconsistently, affecting predictive accuracy.

Similarly, Adhiya et al. (2024) highlighted the efficacy of machine learning models, specifically Random Forest and XGBoost, in predicting 30-day readmissions and identified significant predictors such as patient demographics and service types (Adhiya et al., 2024). Our results complement this by showing that Random Forest performed well, achieving 71% accuracy, while SVM excelled with 74% accuracy and better AUC scores, reinforcing the role of advanced models in predicting readmissions. Lu and Uddin (2022) utilized stacking-based models and feature selection techniques to improve predictive accuracy, which reflects our findings where feature importance and model interpretability played critical roles (Lu & Uddin, 2022). Their emphasis on model explainability through LIME supports our observation that raw feature models provide valuable insights, particularly in predicting short-term readmissions.

In relation to the research questions, the findings from the comparative analysis of machine learning models for predicting diabetic hospital readmissions offer several important implications.

In conclusion, while advanced machine learning models like SVM and Random Forest are effective tools for predicting hospital readmissions, their practical application in healthcare requires careful consideration of the data preprocessing techniques used. The integration of these models with business intelligence tools like Power BI enhances their utility by providing actionable insights that can improve patient outcomes and reduce healthcare costs.

**5.2 Factors Contributing to High Readmission Rates:** The performance of the machine learning models, particularly the SVM and Random Forest, highlights key factors influencing readmission rates. These models, which effectively distinguish between different classes of readmission, suggest that certain variables identified through feature importance analysis, significantly impact readmission rates. Understanding these factors can help pinpoint critical areas for intervention.

**5.3 Predictive Analytics for High-Risk Patients:** The ability of the SVM model to achieve high accuracy and F1-scores, especially in identifying short-term



readmissions, demonstrates the potential of predictive analytics to flag patients at high risk of readmission. This capability allows for early identification and targeted management of patients who are most likely to experience readmission within 30 days.

**5.4 Impact of Targeted Interventions:** By leveraging predictive models, healthcare providers can implement targeted interventions for patients identified as high-risk. This proactive approach can potentially reduce readmission rates by addressing the specific needs and risks associated with these patients before they are discharged.

**5.5 Effect of Machine Learning on Healthcare Costs:** The use of machine learning to predict readmission risks can lead to more efficient resource allocation and cost management. Accurate predictions enable healthcare providers to focus their resources on patients who need the most attention, thereby optimizing treatment costs and reducing unnecessary hospitalizations.

**5.6 Gaps in Diabetes Management and Post-Discharge Care:** The findings reveal gaps in managing diabetes and post-discharge care, as evidenced by the mixed results from PCA. These gaps, such as the challenges in predicting long-term readmissions, highlight the need for continuous model updates and integration of additional features to address these shortcomings.

**5.7 Improvement in Identifying High-Risk Patients:** The successful application of SVM and Random Forest models underscores their potential to enhance the identification of high-risk diabetic patients. By improving risk stratification, healthcare systems can better target interventions and improve patient outcomes.

**5.8 Effective Strategies for Resource Allocation:** The ability of machine learning models to provide actionable insights into patient readmission risks supports more strategic resource allocation. This targeted approach helps ensure that healthcare resources are used effectively to address the needs of the highest-risk patients.

**5.9 Impact of Predictive Analytics and BI Tools on Health Outcomes:** The integration of machine learning and Power BI into healthcare practices facilitates a deeper understanding of patient data and readmission patterns. This enhanced insight drives improvements in treatment protocols and overall patient care, contributing to better health outcomes and quality of life for diabetic patients.

Overall, the findings highlight the significant role of advanced analytics and machine learning in addressing critical challenges in diabetes management and hospital readmissions, offering valuable strategies for enhancing patient care and optimizing healthcare resources.

## 5.6 Limitations

The project faced several challenges that impacted its overall effectiveness. One significant limitation was the reduction of the dataset from 101,767 records to 1,500, driven by the high computational costs associated with training algorithms on the full dataset. This reduction may have compromised the depth and reliability of the findings. Moreover, the dataset, which covers the years 1999 to 2008, might not fully capture recent developments or emerging trends, emphasizing the need for ongoing model retraining to maintain accuracy and relevance in a rapidly evolving healthcare landscape.

While Power BI excelled in data visualization, it struggled with managing large datasets and performing advanced analytical tasks, limiting the scope of insights that could be derived. Additionally, the models' predictive performance could have been enhanced by including additional features, such as socio-economic and lifestyle factors, which were absent from the dataset. The application of Principal Component Analysis (PCA) also presented mixed outcomes—occasionally enhancing model performance but often at the expense of interpretability, which is crucial in a medical context.

Future work could address these limitations by integrating more diverse and up-to-date data sources, as well as leveraging advanced machine learning techniques to boost both the predictive accuracy and practical utility of the models. Despite these challenges, the project offers valuable insights that can guide strategies for reducing diabetic readmissions and optimizing healthcare resources, ultimately contributing to more effective and efficient patient care.

## 5.7 Ethical Consideration

This project places significant emphasis on ethical considerations, particularly in how patient data is handled, ensuring compliance with relevant data protection regulations and ethical standards. In the course of this project, the following was carried:

**5.7.1 Anonymization and Data Handling:** The dataset used in this project is open-source and fully anonymized, ensuring that no personally identifiable information (PII) is accessible. Anonymization procedures were meticulously followed, where all direct and indirect identifiers were removed or altered, making it impossible to trace data back to individual patients. This process is crucial for maintaining patient confidentiality and protecting individuals from potential misuse of their data.

**5.7.2 Informed Consent:** Although the data is anonymized and open-source, the project aligns with ethical guidelines regarding informed consent. The original data providers would have secured informed consent from the patients, ensuring that

participants were aware of how their data would be used in research. The use of this data in the project respects these consent agreements, adhering to the intended use stipulated by the data providers.

**5.7.3 Compliance with Data Protection Regulations:** The project strictly complies with the General Data Protection Regulation (GDPR) and other relevant data protection laws. GDPR mandates that any personal data used in research must be handled with the utmost care, ensuring privacy and data security. Even though the data used in this project is anonymized, the research process respects the principles of data minimization, purpose limitation, and data integrity as outlined by GDPR. Regular audits and checks were implemented to ensure continuous compliance with these regulations.

**5.7.4 Prevention of Bias and Fairness:** The ethical integrity of the project extends to the prevention of bias in the predictive models. The dataset was carefully balanced to avoid any disproportionate impact on specific patient groups, such as those based on race, gender, age, or socio-economic status. By doing so, the project aims to prevent algorithmic discrimination and promote fairness in the predictions made by the models. Each model was rigorously tested to ensure that it treats all patient groups equitably, avoiding any bias that could lead to unfair treatment or healthcare disparities.

**5.7.5 Transparency and Accountability:** Throughout the research process, transparency was maintained by clearly documenting all methodologies, data sources, and analytical techniques used. This documentation allows for the reproducibility of results and ensures that the research can be scrutinized and validated by others in the field. Transparency also supports accountability, ensuring that the research adheres to the highest ethical standards.

**5.7.6 Adherence to Ethical Standards:** The project adheres to the British Computer Society (BCS) Codes of Conduct and follows the guidelines set by the Information Commissioner's Office (ICO) for data governance. These standards provide a framework for ethical behavior in the handling and analysis of data, ensuring that the research is conducted with integrity and respect for the rights of individuals. Additionally, the project upholds the principles of the Equality Act 2010, which emphasizes the importance of equality and non-discrimination. This commitment ensures that all aspects of data handling, model development, and analysis are conducted in a manner that respects and promotes equality for all individuals involved.

## 6. Conclusion

This project successfully demonstrated the application of ML techniques and Power BI for predicting diabetic patient readmissions and visualizing key metrics. By leveraging various ML algorithms, particularly the SVM, the analysis provided valuable insights into the factors influencing readmissions. Power BI's interactive dashboards effectively highlighted these factors, facilitating a deeper understanding of patient data and supporting informed decision-making.

However, the project faced notable limitations, including the reduction of the dataset to manage computational demands and the use of outdated data, which may not reflect current trends. Additionally, while Power BI excelled in visualization, it struggled with handling large datasets and complex analytical tasks. The inclusion of PCA yielded mixed results, occasionally enhancing performance but also impacting interpretability.

### 6.1 Recommendations

To build upon the successes of this project and address its limitations, several recommendations are proposed:

- **Expand Data Scope:** Integrate more recent and comprehensive datasets to capture current trends and developments in diabetic care. Including additional features such as socio-economic and lifestyle factors could further enhance predictive accuracy and provide a more holistic view of patient readmission risks.
- **Enhance Analytical Tools:** Explore advanced machine learning techniques and tools beyond those used in this project. Techniques such as ensemble methods or deep learning could offer improved performance and insights. Additionally, consider alternative data visualization platforms that can handle larger datasets and more complex analyses effectively.
- **Continuous Model Retraining:** Implement a strategy for ongoing model retraining and validation to ensure that predictive models remain relevant and accurate over time. This will help adapt to changes in patient demographics, treatment protocols, and healthcare practices.
- **Improve Interpretability:** Given the critical nature of decision-making in healthcare, future work should focus on balancing model accuracy with interpretability. Refining feature selection and employing more interpretable models or visualization techniques will help ensure that the insights provided are both actionable and understandable for healthcare professionals.

By addressing these recommendations, future projects can enhance their predictive capabilities, provide deeper insights, and ultimately contribute to more effective and efficient management of diabetic patient care.

## 7. REFERENCES

- Adhiya, J., Barghi, B., and Azadeh-Fard, N., 2024. Predicting the risk of hospital readmissions using a machine learning approach: a case study on patients undergoing skin procedures. *Frontiers in Artificial Intelligence*, [online] 6. Available at: <https://www.frontiersin.org/articles/10.3389/frai.2023.1213378/full> [Accessed 30 June 2024].
- Alajmani, S. and Elazhary, H. (2019) 'Hospital Readmission Prediction using Machine Learning Techniques', *International Journal of Advanced Computer Science and Applications*, 10(4). doi: <https://doi.org/10.14569/ijacsa.2019.0100425>.
- Alloghani, M., Aljaaf, A., Hussain, A., Baker, T., Mustafina, J., Al-Jumeily, D. and Khalaf, American Diabetes Association. Economic costs of diabetes in the U.S. in 2007. *Diabetes Care*. 2008;31(3):596–615.
- Anon (n.d.) KNN Classification using Scikit-learn – Machine Learning Geek. [online] Available at: <https://machinelearninggeek.com/knn-classification-using-scikit-learn/> [Accessed 21 Aug. 2024].
- Bissacco, A., Yang, M.-H. and Soatto, S., 2007. Fast human pose estimation using appearance and motion via multi-dimensional boosting regression. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, MN. doi: 10.1109/CVPR.2007.383129.
- Calver J, Brameld KJ, Preen DB, Alexia SJ, Boldy DP, McCaul KA. High-cost users of hospital beds in Western Australia: a population-based record linkage study. *Med J Aust*. 2006;184(8):393–7.
- Camizuli, E. & Carranza, E.J.M., 2018. Exploratory Data Analysis (EDA). In: *Encyclopedia of Statistics in Applied Science*. pp.1-7. Wiley. doi:10.1002/9781119188230.saseas0271.
- Chin, D.L., Bang, H., Manickam, R.N., and Romano, P.S., 2016. Rethinking thirty-day hospital readmissions: shorter intervals might be better indicators of quality of care. *Health Affairs (Millwood, VA)*, 35(10), pp.1867-1875.
- Collins, J., Abbass, I.M., Harvey, R., et al., 2017. Predictors of all-cause 30-day readmission among Medicare patients with type 2 diabetes. *Current Medical Research and Opinion*, 33(8), pp.1517-1523.
- Cover, T.M. and Hart, P.E. (1967) 'Nearest neighbor pattern classification', *IEEE Transactions on Information Theory*, 13(1), pp. 21-27.

Eby, E., Hardwick, C., Yu, M., et al., 2015. Predictors of 30-day hospital readmission in patients with type 2 diabetes: a retrospective, case-control, database study. *Current Medical Research and Opinion*, 31(1), pp.107-114.

Fischer, C., Lingsma, H.F., Marang-van de Mheen, P.J., Kringos, D.S., Klazinga, N.S., and Steyerberg, E.W., 2014. Is the readmission rate a valid quality indicator? A review of the evidence. *PLoS One*, 9(11), p.e112282.

Freund, Y. and Schapire, R., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55, pp.119–139.

Friedman, J., Hastie, T. and Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 28, pp.337–407. doi: 10.1214/aos/1016218222.

Gao, X., Alam, S., Shi, P., Dexter, F. and Kong, N., 2023. Interpretable machine learning models for hospital readmission prediction: A two-step extracted regression tree approach. *BMC Medical Informatics and Decision Making*, 23(1). [online] Available at: <https://doi.org/10.1186/s12911-023-02193-5> [Accessed 3 July 2024].

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J., & Herrera, F. (2016) 'Big data preprocessing: methods and prospects', *Big Data Analytics*, 1, pp. 1-22. doi: 10.1186/s41044-016-0014-0.

Gerhardt, G., Yemane, A., Hickman, P., Oelschlaeger, A., Rollins, E., and Brennan, N., 2013. Medicare readmission rates showed meaningful decline in 2012. *Medicare Medicaid Research Review*, 3(2).

Goodreads.com. (2024). Data Analysis with Microsoft Power BI Quotes by Larson. [online] Available at: <https://www.goodreads.com/work/quotes/69234781-data-analysis-with-microsoft-power-bi> [Accessed 29 Aug. 2024].

Googleusercontent.com. (2020). الجمعية العربية لوقاية النباتات [The Arab Society for Plant Protection]. Retrieved from [https://scholar.googleusercontent.com/scholar?q=cache:gwQd6U9crclJ:scholar.google.com/+mathematical+representation+of+standard+scaler+in+machine+learning&hl=en&as\\_sdt=0](https://scholar.googleusercontent.com/scholar?q=cache:gwQd6U9crclJ:scholar.google.com/+mathematical+representation+of+standard+scaler+in+machine+learning&hl=en&as_sdt=0). (Accessed 21 Aug. 2024).

Habehh, H. and Gohel, S., 2021. Machine Learning In Healthcare. *Current Genomics*, 22(4). doi:10.2174/1389202922666210705124359.

Hutchinson, R.A., Liu, L.-P. and Dietterich, T.G., 2011. Incorporating boosted regression trees into ecological latent variable models. In: *AAAI Conference on Artificial Intelligence (AAAI'11)*, San Francisco, CA, pp.1343–1348. Available at:

<http://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/view/3711> [Accessed 21 Aug. 2024].

IBM (no date) Data Preparation Overview. Available at: <https://www.ibm.com/docs/en/spss-modeler/saas?topic=preparation-data-overview> (Accessed: 19 August 2024).

Howell, S., Coory, M., Martin, J., and Duckett, S., 2009. Using routine inpatient data to identify patients at risk of hospital readmission. *BMC Health Services Research*, 9(1), p.96.

Isik, O., Jones, M.C., and Sidorova, A., 2011. Business intelligence (BI) success and the role of BI capabilities. *Intelligent Systems in Accounting, Finance and Management*, 18(4), pp.161-176.

Jayatilake, S.M.D.A.C. and Ganegoda, G.U., 2021. Involvement of Machine Learning Tools in Healthcare Decision Making. *Journal of Healthcare Engineering*, 2021, pp.1-20. doi:10.1155/2021/6679512.

Johansen H, Nair C, Bond J. Who goes to the hospital? An investigation of high users of hospital days. *Health Rep*. 1994;6(2):253–77.

Johnson, R. and Zhang, T., 2012. Learning nonlinear functions using regularized greedy forest. Technical Report. arXiv:1109.0887. doi: 10.2172/1052139.

kaggle.com. (n.d.). *Predicting Hospital Readmission of Diabetics*. [online] Available at: <https://www.kaggle.com/code/chongchong33/predicting-hospital-readmission-of-diabetics/notebook>.

Kareem, F. Q., & Abdulazeez, A. M. (n.d.). Ultrasound Medical Images Classification Based on Deep Learning Algorithms: A Review.

Kassin MT, Owen RM, Perez SD, Leeds I, Cox JC, Schnier K, Sadiraj V, Sweeney JF. Risk factors for 30-day hospital readmission among general surgery patients. *J Am Coll Surg*. 2012;215(3):322–30.

Lu, H. and Uddin, S. (2022). Explainable Stacking-Based Model for Predicting Hospital Readmission for Diabetic Patients. *Information*, 13(9), p.436. doi: <https://doi.org/10.3390/info13090436>.

Low, Z.K., Liew, L., Chua, V., Chew, S. and Ti, L.K. (2023) 'Predictors of unplanned hospital readmission after non-cardiac surgery in Singapore: a 2-year retrospective review', *BMC Surgery*, 23, p.202. doi: <https://doi.org/10.1186/s12893-023-02102-7>.

M. (2019). Implementation of machine learning algorithms to create diabetic patient re-admission profiles. *BMC Medical Informatics and Decision Making*, 19(S9). doi:<https://doi.org/10.1186/s12911-019-0990-x>.



Majumder, A.B., Gupta, S., Singh, D. and Majumder, S., 2021. An intelligent system for prediction of COVID-19 case using machine learning framework-logistic regression. *Journal of Physics: Conference Series*, 1797(1). Available at: <https://doi.org/10.1088/1742-6596/1797/1/012011> [Accessed 21 August 2024].

Mettler, T. and Vimarlund, V., 2009. Understanding business intelligence in the context of healthcare. *Health Informatics Journal*, 15(3), pp.254-264.

Michailidis, P., Dimitriadou, A., Papadimitriou, T. and Gogas, P. (2022) 'Forecasting Hospital Readmissions with Machine Learning', *Healthcare*, 10(6), p. 981. doi: <https://doi.org/10.3390/healthcare10060981>.

Negash, S., 2004. Business intelligence. *Communications of the Association for Information Systems*, 13, pp.177-195.

Ostling, S., Wyckoff, J., Ciarkowski, S.L., et al., 2017. The relationship between diabetes mellitus and 30-day readmission rates. *Clinical Diabetes and Endocrinology*, 3(1), p.3.

Pittman, S.J. and Brown, K.A., 2011. Multi-scale approach for predicting fish species distributions across coral reef seascapes. *PLoS ONE*, 6(11), e20583. doi: 10.1371/journal.pone.0020583.

Png, M.E., Yoong, J., Chen, C., et al., 2018. Risk factors and direct medical cost of early versus late unplanned readmissions among diabetes patients at a tertiary hospital in Singapore. *Current Medical Research and Opinion*, 34(6), pp.1071-1080.

Ramamurthy, K., Sen, A., and Sinha, A.P., 2008. An empirical investigation of the key determinants of data warehouse adoption. *Decision Support Systems*, 44(4), pp.817-841.

Rohloff, R., 2011. Health-care BI: A tool for meaningful analysis. *Healthcare Financial Management*, 65(5), pp.100-108.

Rubin, D.J., Handorf, E.A., Golden, S.H., Nelson, D.B., McDonnell, M.E. and Zhao, H., 2016. Development and validation of a novel tool to predict hospital readmission risk among patients with diabetes. *Endocrine Practice*, 22(10), pp.1204-1215.

Schott, M. (2020). Random Forest Algorithm for Machine Learning. [online] Medium. Available at: <https://medium.com/capital-one-tech/random-forest-algorithm-for-machine-learning-c4b2c8cc9feb> [Accessed 21 Aug. 2024].

Seah, M., Hsieh, M.H., and Weng, P., 2010. A case analysis of Savecom: The role of indigenous leadership in implementing a business intelligence system. *International Journal of Information Management*, 30(4), pp.368-373.

Simplilearn (2023). Random Forest Algorithm. [online] Available at: <https://www.simplilearn.com/tutorials/machine-learning-tutorial/random-forest-algorithm> [Accessed 21 Aug. 2024].

Sharma, A., Agrawal, P., Madaan, V., & Goyal, S. (2019). Prediction on diabetes patient's hospital readmission rates. Proceedings of the Third International Conference on Advanced Informatics for Computing Research. ACM, 1(1), 1-5.

Soh, J.G.S., Wong, W.P., Mukhopadhyay, A., Quek, S.C., and Tai, B.C., 2020. Predictors of 30-day unplanned hospital readmission among adult patients with diabetes mellitus: a systematic review with meta-analysis. *BMJ Open Diabetes Research & Care*, 8(1), p.e001227.

Stefan MS, Pekow PS, Nsa W, Priya A, Miller LE, Bratzler DW, Rothberg MB, Goldberg RJ, Baus K, Lindenauer PK. Hospital performance measures and 30-day readmission rates. *J Gen Intern Med*. 2013;28(3):377–85.

Sultana, A. and Islam, R.M., 2023. Machine learning framework with feature selection approaches for thyroid disease classification and associated risk factors identification. *Journal of Electrical Systems and Information Technology*, 10(1). doi: 10.1186/s43067-023-00101-5.

Peter-Myers. (2023). BI solution architecture in the Center of Excellence - Power BI. Microsoft Learn. Retrieved from <https://learn.microsoft.com/en-us/power-bi/guidance/center-of-excellence-business-intelligence-solution-architecture>

Tremblay, M., Hevner, A., and Berndt, D., 2012. Design of an information volatility measure for health care decision making. *Decision Support Systems*, 52(2), pp.331-341.

Wixom, B.H. and Watson, H.J., 2001. An empirical investigation of the factors affecting data warehousing success. *MIS Quarterly*, 25(1), pp.17-41.

Wendong, Y., Zhengzheng, L., & Bo, J. (2017, June). A multi-factor analysis model of quantitative investment based on GA and SVM. In 2017 2nd International Conference on Image, Vision and Computing (ICIVC) (pp. 1152-1155). IEEE.

Zeebaree, D.Q., Haron, H., & Abdulazeez, A.M. (2018). Gene selection and classification of microarray data using convolutional neural network. In 2018 International Conference on Advanced Science and Engineering (ICOASE) (pp. 145-150). IEEE.

Zhang, T., Lin, W., Vogelmann, A.M., Zhang, M., Xie, S., Qin, Y. and Golaz, J., 2021. Improving convection trigger functions in deep convective parameterization schemes using machine learning. *Journal of Advances in Modeling Earth Systems*, 13(5). Available at: <https://doi.org/10.1029/2020ms002365> [Accessed 21 Aug. 2024].

## 8. APPENDIX A

**Distribution of time\_in\_hospital Before and After Log Transformation:** The original data for several features exhibited high right-skewness, characterized by a concentration of lower values and few high-value outliers. Various transformations were applied to address this skewness and normalize the distributions. The log transformation was effective for most features but less so for highly skewed ones like number\_outpatient. The square root transformation further improved balance for most features, while the Box-Cox transformation was particularly successful in normalizing features with moderate skewness. These transformations are essential for improving model performance by stabilizing variance and aligning with algorithmic assumptions. The Box-Cox transformation, in particular, proved highly effective for achieving near-normal distributions in moderately skewed data.

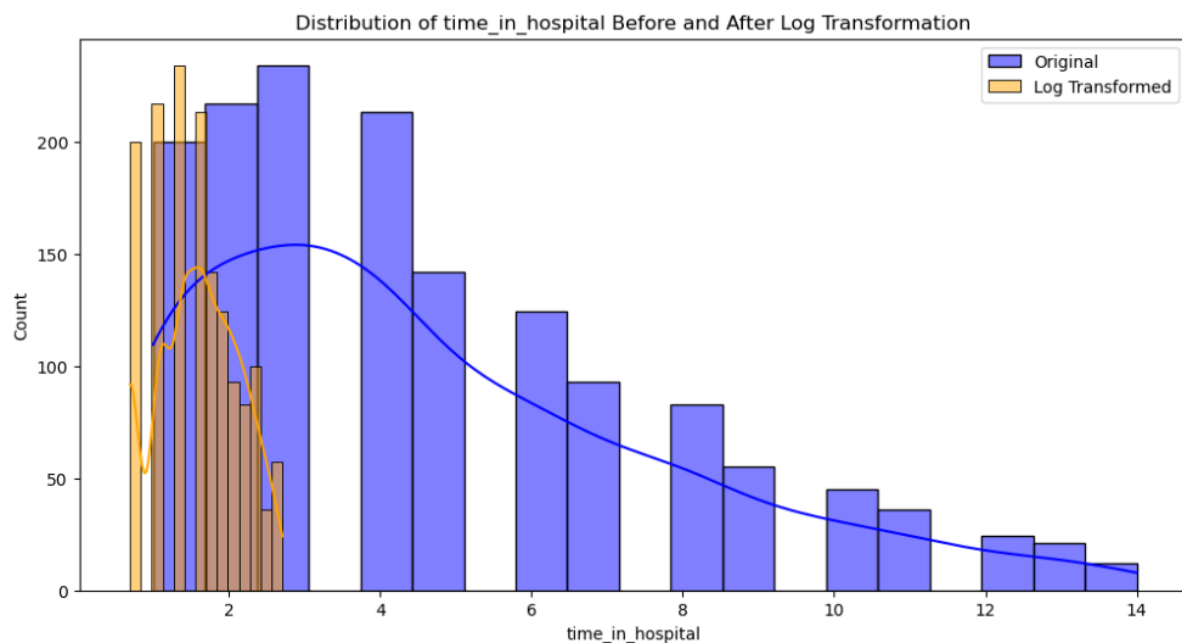


Figure 1: Distribution of Time in Hospital Before and After Log Transformation