Pune Institute of Computer Technology, Pune

A PROJECT REPORT ON

"Patient Readmission Prediction"

SUBMITTED TO THE UNIVERSITY OF PUNE, PUNE
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE

BACHELOR OF ENGINEERING
Computer Engineering

BY

SAMARTH MALI **ROLL NO:41146**
SARTHAK NIRGUDE **ROLL NO:41153**
KEDAR PAWAR **ROLL NO:41158**

UNDER THE GUIDANCE OF

PROF. P. S. JOSHI

DEPARTMENT OF COMPUTER ENGINEERING
PUNE INSTITUTE OF COMPUTER TECHNOLOGY, PUNE
2024-25

C E R T I F I C A T E

This is to certify that the Project Report entitled

"Patient Readmission Prediction"

Submitted by

SAMARTH MALI	**ROLL NO:41146**
SARTHAK NIRGUDE	**ROLL NO:41153**
KEDAR PAWAR	**ROLL NO:41158**

is a bonafide work carried out under the supervision of Prof. P. S. JOSHI and it is submitted towards the partial fulfillment of the requirement of Savitribai Phule Pune University, Pune for the award of the degree of Bachelor of Engineering(Computer Engineering).

Prof. P. S. JOSHI
(Guide)

Dr. Geetanjali Kale
(Head of Department
of Computer Engineering)

Dr. S.T. Gandhe
(Principal,Pune Institute of Computer
Technology, Pune)

Internal Examiner

External Examiner

Place: Pune

Date:

ABSTRACT

The mini-project on Business Intelligence (BI) presents a case study focusing on patient readmission prediction. The objective of this study is to explore the potential of BI techniques in improving healthcare outcomes by identifying factors that contribute to patient readmissions. In recent years, healthcare organizations have faced significant challenges in managing patient readmissions, which not only impact patient health but also pose financial burdens. By leveraging BI tools and methodologies, healthcare providers can gain insights from large volumes of patient data to predict and prevent readmissions. This mini-project analyzes a dataset comprising various patient attributes, including demographic information, medical history, and clinical factors. Using a combination of statistical analysis and machine learning algorithms, the study aims to identify patterns and correlations that influence the likelihood of patient readmission. The research methodology involves data preprocessing, feature selection, and model development. Different techniques such as logistic regression, decision trees, and ensemble methods are applied to build predictive models. Evaluation metrics, such as accuracy, precision, recall, and F1 score, are utilized to assess the performance of the models. The findings from this study can assist healthcare providers in developing targeted interventions and personalized care plans to reduce readmission rates. By understanding the key factors associated with readmission, healthcare organizations can allocate resources efficiently and implement preventive measures to improve patient outcomes and reduce healthcare costs. Overall, this mini-project showcases the potential of BI in the healthcare sector and demonstrates how data-driven approaches can enhance decision-making processes, specifically in the context of patient readmission prediction. The results provide valuable insights for healthcare professionals, researchers, and policymakers striving to optimize patient care and ensure better healthcare outcomes. .

Contents

Chapter 1
SYNOPSIS

1.1 Project Title

Patient Readmission Prediction

1.2 Objective

To predict the readmission status of patient using Data mining techniques

1.3 Technical Keywords

Patient readmission prediction, business intelligence, machine learning.

1.4 Problem definition

The problem addressed in this mini-project is the prediction of patient readmissions in healthcare settings. Patient readmissions pose significant challenges for healthcare providers, impacting patient health outcomes and incurring substantial financial costs. The objective is to leverage business intelligence techniques and machine learning algorithms to identify factors and patterns that contribute to patient readmissions. By accurately predicting readmissions, healthcare organizations can proactively intervene, allocate resources effectively, and implement targeted interventions to reduce readmission rates and improve patient care. The project aims to develop a predictive model that can assist healthcare providers in making data-driven decisions to prevent unnecessary readmissions and enhance overall healthcare outcomes.

Chapter 2

TECHNICAL KEYWORDS

2.1 Technical Keywords

Patient readmission prediction, business intelligence, machine learning, Patient readmission prediction, , Business intelligence, Machine learning , Healthcare analytics , Data preprocessing

2.2 Area of Project

“This project falls within the intersection of healthcare and data analytics, specifically focusing on patient readmission prediction. It encompasses the fields of business intelligence, machine learning, and healthcare analytics. The project involves the analysis and exploration of large healthcare datasets, including patient demographic information, medical history, and clinical factors. Techniques such as data preprocessing, feature selection, and model development are utilized to build predictive models. The project also includes the evaluation of model performance using various metrics. Overall, the project aims to leverage data-driven approaches to improve healthcare outcomes and reduce patient readmissions.”

Chapter 3

INTRODUCTION

3.1 Introduction

The case study presented here delves into the domain of healthcare analytics and focuses on the critical issue of patient readmission prediction. Patient readmissions pose significant challenges for healthcare providers, affecting both patient well-being and healthcare costs. By harnessing the power of business intelligence techniques and machine learning algorithms, healthcare organizations can gain valuable insights from vast amounts of patient data and predict the likelihood of readmission.

In recent years, advancements in data analytics and machine learning have opened up new possibilities for improving healthcare outcomes. By analyzing various patient attributes, including demographic information, medical history, and clinical factors, it becomes possible to uncover patterns and correlations that contribute to readmissions. The application of predictive models can aid in identifying high-risk patients, allowing healthcare providers to intervene proactively and provide targeted interventions to prevent unnecessary readmissions.

Overall, this case study showcases the potential of business intelligence and machine learning in the healthcare sector and demonstrates how data-driven approaches can empower healthcare providers to make informed decisions, ultimately leading to improved patient care and reduced readmission rates.

3.2 Scope

The project ”Patient Readmission Prediction” aims to develop a predictive model that can accurately forecast the likelihood of patient readmission in a healthcare setting.The primary focus is on identifying key factors and patterns that contribute to readmissions, which will be leveraged to build a robust and accurate prediction model. The project will involve data collection, preprocessing, feature selection/engineering, model development, and evaluation. In the data collection phase, relevant patient data will be gathered, including demographics, medical history, diagnoses, procedures, medications,

and other pertinent information. This data will be obtained from healthcare institutions or databases, ensuring a diverse representation of patients. The collected data will undergo preprocessing steps to clean and handle missing values, outliers, and inconsistencies, ensuring data integrity.

Next, feature selection and engineering techniques will be applied to extract meaningful features from the dataset. This step involves identifying the most relevant features based on domain knowledge and statistical analysis, as well as creating new features to capture specific patterns or relationships. Feature selection aims to optimize the model's performance and interpretability

Chapter 4

REQUIREMENTS

To effectively work on the patient readmission prediction project, the following system requirements are recommended:

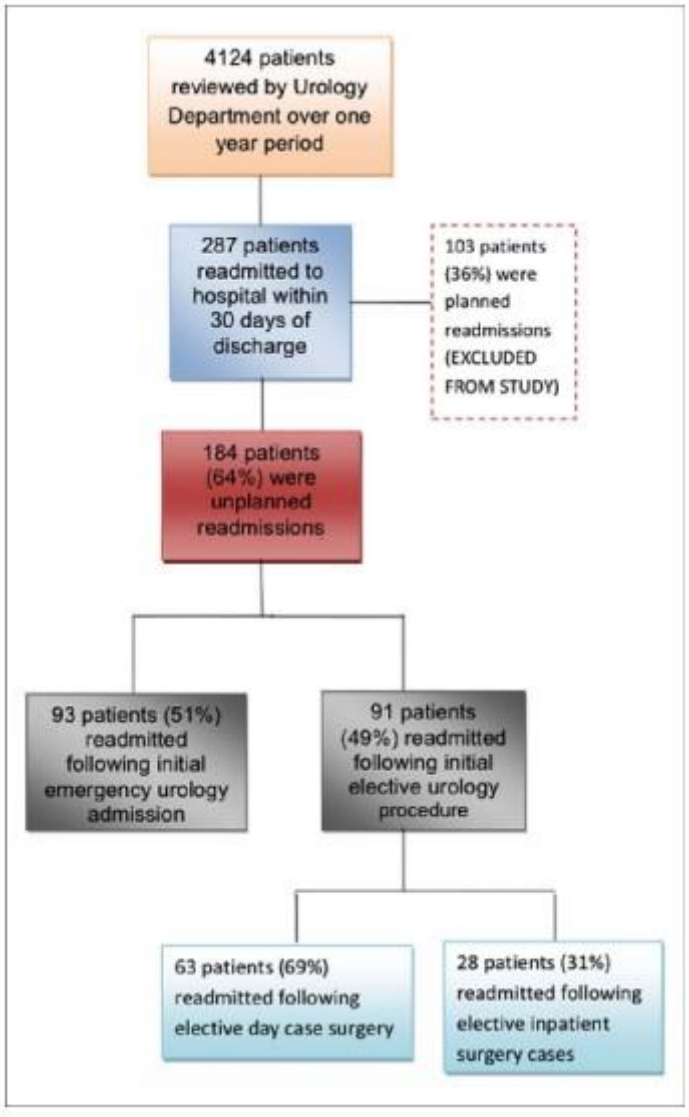
4.1 System Requirements

- Processor: Intel Core i5 or higher (or equivalent AMD processor)
- RAM: Minimum 8GB (16GB or higher recommended for larger datasets)
- Storage: At least 250GB hard disk drive (SSD recommended for improved performance)
- Operating System: Windows 10, macOS, or Linux Software: Python programming language (version 3.6 or higher)
- Development Environment: Jupyter Notebook, Anaconda, or any preferred Python IDE (Integrated Development Environment)
- Libraries: Required libraries such as Pandas, NumPy, Scikit-learn, TensorFlow, or other machine learning frameworks Internet Connection: Stable internet connection for downloading datasets, libraries, and accessing additional resources
- Graphics Processing Unit (GPU): Optional but recommended for faster training and inference in machine learning models

Chapter 5

IMPLEMENTATION

5.1 Flowchart



5.2 Algorithm

1. Start.

2. Collect patient data including demographics, medical history, diagnoses, procedures, medications, and other relevant information.

3. Preprocess the data by handling missing values, outliers, and inconsistencies.

4. Perform feature selection and engineering to identify relevant features and createnew ones if necessary.

5. Split the dataset into training and testing sets for model development and evaluation.

6. Choose a suitable machine learning algorithm (e.g., logistic regression, decisiontree, random forest, support vector machine, or neural network) for prediction.

7. Train the chosen model using the training dataset.

8. Validate the model using the testing dataset and evaluate its performance usingappropriate metrics such as accuracy, precision, recall, and F1-score.

9. If the model's performance is unsatisfactory, consider adjusting hyperparametersor trying alternative models.

10. Once a satisfactory model is obtained, use it to predict the likelihood of patientreadmission for new/unseen data.

11. End

Chapter 6

METHODOLOGY

6.1 Data Collection

Gather a comprehensive dataset containing relevant information related to patient readmissions, including demographic data, medical history, clinical factors, and outcome labels indicating readmission status.

6.2 Data Preprocessing

Cleanse and preprocess the collected data to ensure its quality and usability. Handle missing values, outliers, and inconsistencies. Perform data transformations, normalization, and feature scaling as necessary.

6.3 Feature Selection

Analyze the dataset to identify the most significant features that contribute to patient readmissions. Utilize statistical methods, correlation analysis, and domain knowledge to select the most relevant features for model development.

6.4 Model Development

Implement various machine learning algorithms such as logistic regression, decision trees, random forests, or support vector machines. Split the preprocessed dataset into training and testing sets. Train the models on the training set using the selected features.

6.5 Model Evaluation

Evaluate the performance of the trained models using appropriate evaluation metrics such as accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC-ROC). Compare the performance of different models to identify the most accurate and reliable predictor of patient readmissions.

6.6 Model Optimization

Fine-tune the selected model to improve its performance. Perform hyperparameter tuning, such as adjusting regularization parameters, tree depth, or ensemble size, using techniques like cross-validation or grid search.

6.7 Validation and Testing

Validate the optimized model on a separate validation dataset to assess its generalization capabilities. Evaluate the model's performance on the testing dataset to ensure its reliability and robustness.

6.8 Interpretation and Insights

Analyze the trained model to interpret the importance of features and identify factors influencing patient readmissions. Derive actionable insights and recommendations based on the model's predictions and feature contributions.

6.9 Documentation and Reporting

Document the entire methodology, including data collection procedures, preprocessing steps, model development, and evaluation processes. Present the findings, insights, and recommendations in a comprehensive report or presentation.

6.10 Iterative Improvement

Iterate through the methodology, refining and enhancing the models based on feedback and additional insights. Consider incorporating more advanced techniques, such as deep learning or ensemble methods, to further improve the accuracy and predictive power of the models.

Chapter 7

WORKING

Predicting patient readmission involves developing a model that can estimate the likelihood of a patient being readmitted to a healthcare facility within a specific time period after their initial discharge. The objective is to identify key factors and patterns that contribute to readmissions and leverage this information to build an accurate prediction model.

To train a predictive model, relevant patient data needs to be collected. This may include demographic information (age, gender), medical history, diagnoses, procedures, medications, laboratory results, and other relevant factors. Data should be obtained from diverse sources to ensure generalizability. Preprocessing steps involve handling missing values, outliers, and inconsistencies in the data, as well as normalizing numerical data and encoding categorical variables. Feature selection is an important step in developing a prediction model. It involves identifying the most relevant features that contribute to patient readmission. This can be achieved through domain knowledge, statistical analysis, or feature importance techniques such as correlation analysis or recursive feature elimination. Additionally, feature engineering techniques can be employed to create new features that capture specific patterns or relationships within the data.

Machine learning algorithms are commonly used for patient readmission prediction. Logistic regression, decision trees, random forests, support vector machines, and neural networks are popular choices. Logistic regression models the relationship between the independent variables and the probability of readmission. Decision trees and random forests provide interpretable rules for readmission prediction. Support vector machines maximize the separation between readmitted and non-readmitted patients. Neural networks offer flexibility and the ability to capture complex interactions.

The dataset is typically divided into training and testing sets. The training set is used to train the predictive model, while the testing set is used to evaluate its performance. The model is trained by optimizing its parameters using techniques such as gradient descent or maximum likelihood estimation. Evaluation metrics, such as accuracy, precision, recall, and F1-score, are used to assess the model's performance. Cross-validation techniques,

such as k-fold cross-validation, can be applied to obtain more robust performance estimates. Once a satisfactory predictive model is obtained, it can be deployed for predicting patient readmission on new/unseen data. The model's performance and generalizability should be validated on real-world scenarios and compared against established baselines or clinical guidelines. Regular monitoring and updating of the model may be necessary to ensure its continued effectiveness as new data becomes available.

Chapter 8

CODE OUTPUTS

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
```

```
[ ] # Load the dataset
    data = pd.read_csv('diabetic_data.csv')
```

```
[ ] # Data preprocessing
    # Replace missing values with NaN
    data = data.replace('?', np.nan)
```

```
    # Drop columns with a high percentage of missing values
    data = data.drop(['weight', 'payer_code', 'medical_specialty'], axis=1)
```

```
[ ] # Drop rows with missing values in certain columns
    data = data.dropna(subset=['race', 'diag_1', 'diag_2', 'diag_3'])
```

```
[ ] # Convert categorical variables to numeric
    data['gender'] = data['gender'].replace({'Male': 0, 'Female': 1})
    data['change'] = data['change'].replace({'No': 0, 'Ch': 1})
    data['diabetesMed'] = data['diabetesMed'].replace({'No': 0, 'Yes': 1})
```

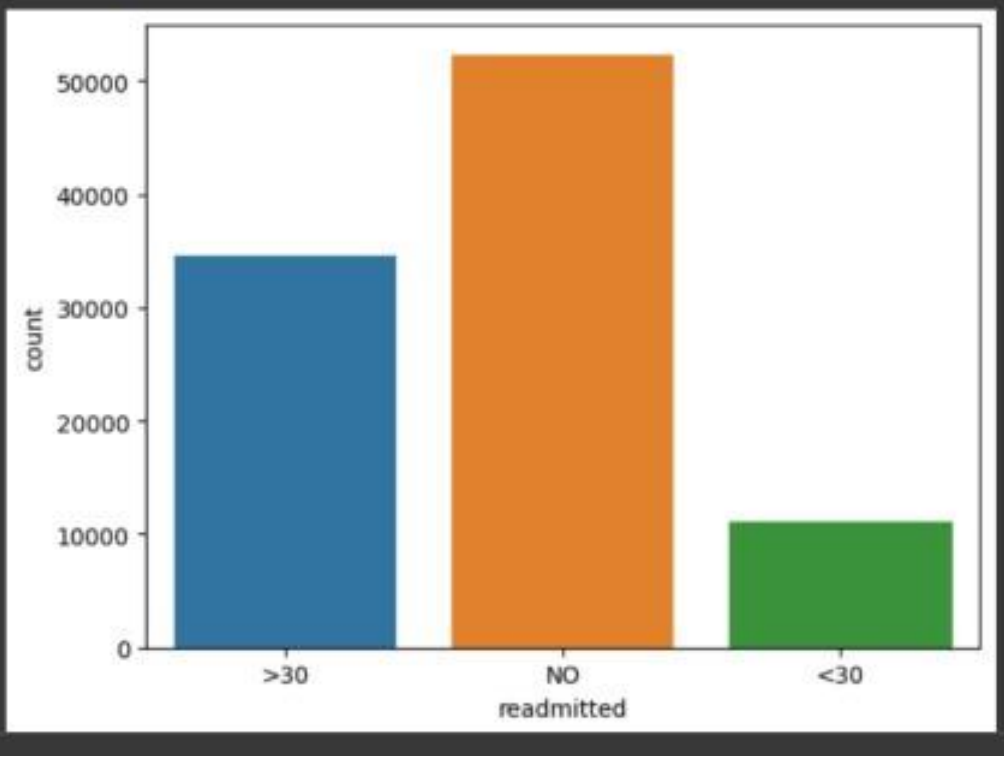
```
[ ] # Exploratory data analysis
    # Calculate summary statistics
    print(data.describe())
```

```
# Exploratory data analysis
# Calculate summary statistics
print(data.describe())
```

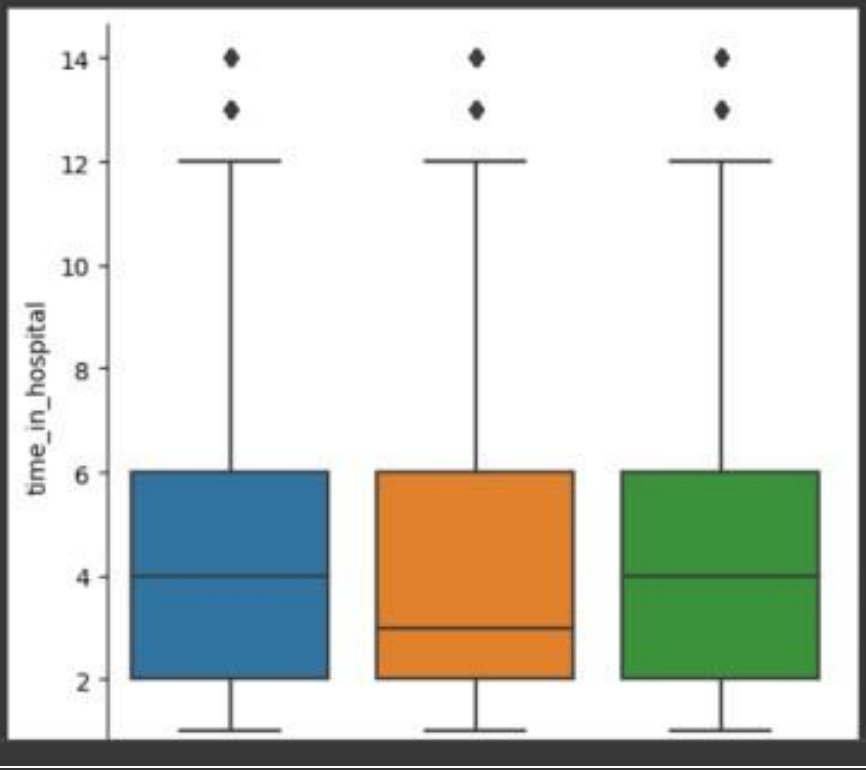
```
       encounter_id   patient_nbr  admission_type_id  \
count  9.805300e+04  9.805300e+04       98053.000000
mean   1.658294e+08  5.484792e+07           2.025813
std    1.024322e+08  3.866175e+07           1.450117
min    1.252200e+04  1.350000e+02           1.000000
25%    8.528566e+07  2.350234e+07           1.000000
50%    1.533019e+08  4.687790e+07           1.000000
75%    2.305007e+08  8.800306e+07           3.000000
max    4.438672e+08  1.895026e+08           8.000000

       discharge_disposition_id  admission_source_id  time_in_hospital  \
count              98053.000000         98053.000000      98053.000000
mean                   3.753368             5.776692          4.421976
std                    5.309392             4.071640          2.993074
min                    1.000000             1.000000          1.000000
25%                    1.000000             1.000000          2.000000
50%                    1.000000             7.000000          4.000000
75%                    4.000000             7.000000          6.000000
max                   28.000000            25.000000         14.000000
```

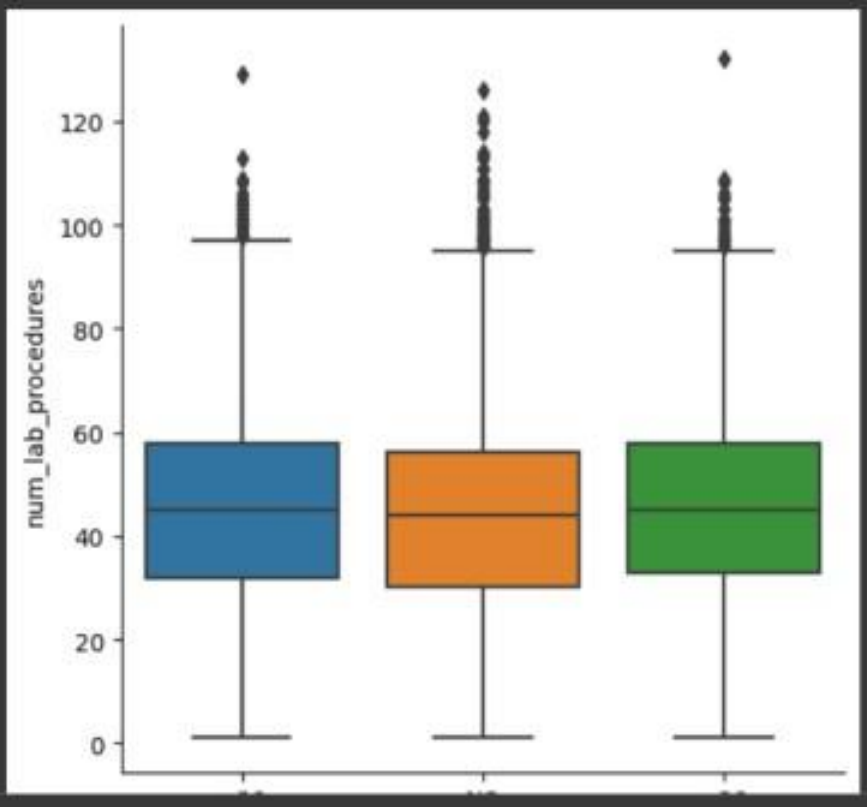
```
# Visualize the distribution of the target variable
sns.countplot(x='readmitted', data=data)
plt.show()
```



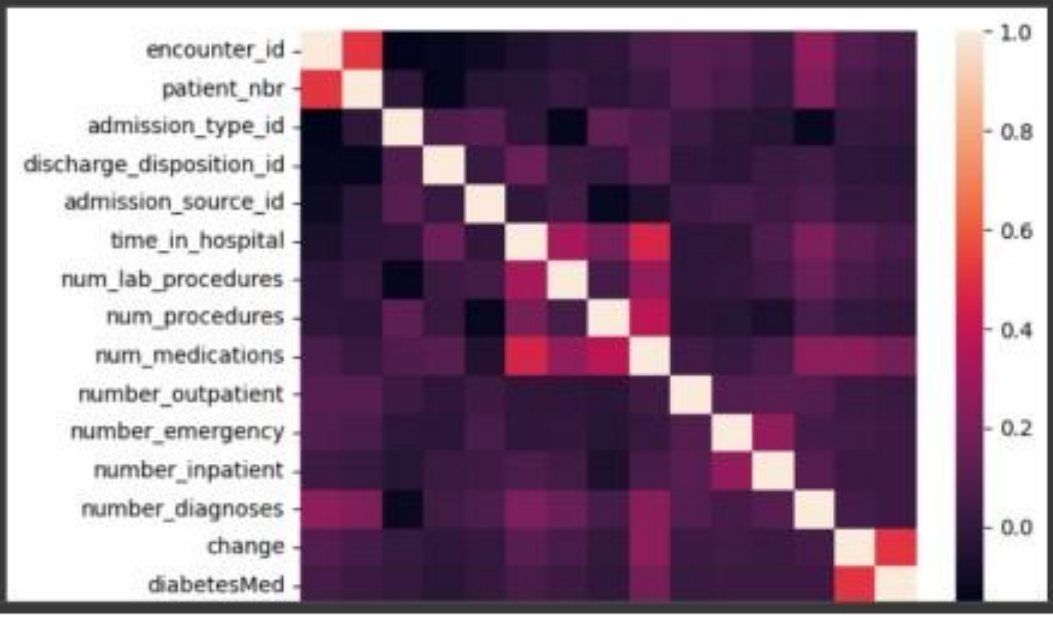
```
# Visualize the relationship between the target variable and other variables
sns.catplot(x='readmitted', y='time_in_hospital', kind='box', data=data)
plt.show()
```




```
[ ] sns.catplot(x='readmitted', y='num_lab_procedures', kind='box', data=data)
plt.show()
```



```
# Visualize the correlation matrix
sns.heatmap(corr_matrix)
plt.show()
```




```
print(features)
```

```
['time_in_hospital', 'num_lab_procedures', 'num_medications']
```

```
# Define the target variable and the features
target = 'readmitted'
features = ['time_in_hospital', 'num_lab_procedures', 'num_medications']
```

```
# Split the data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(data[features], data[target], test_size=0.2, random_state=0)
```

```
# Decision tree
# Initialize the model
dt = DecisionTreeClassifier()
```

```
# Train the model on the training data
dt.fit(X_train, y_train)
```

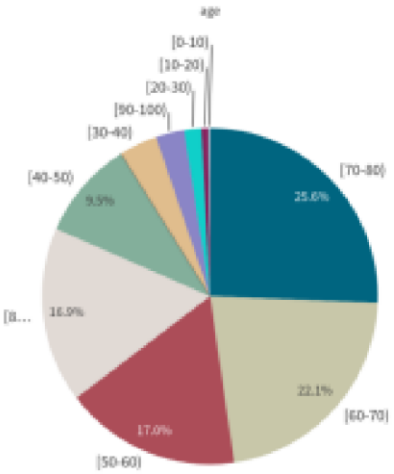
```
DecisionTreeClassifier()
```

```
[ ]  # Evaluate the performance of the model
     accuracy_dt = accuracy_score(y_test, y_pred_dt)
     precision_dt = precision_score(y_test, y_pred_dt, average='macro')
     recall_dt = recall_score(y_test, y_pred_dt, average='macro')
     f1_dt = f1_score(y_test, y_pred_dt, average='macro')
```

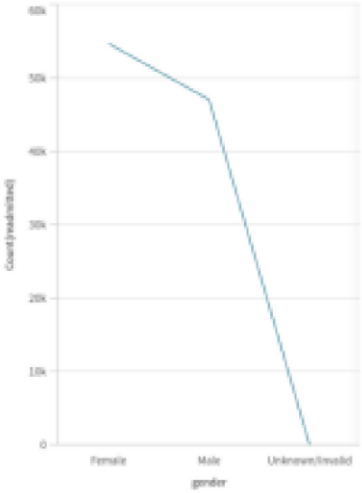
```
[ ]
     print('Decision Tree:')
     print('Accuracy:', accuracy_dt)
     print('Precision:', precision_dt)
     print('Recall:', recall_dt)
     print('F1-score:', f1_dt)
```

```
Decision Tree:
Accuracy: 0.44786089439600224
Precision: 0.34314149182748926
Recall: 0.3423547689913098
F1-score: 0.3412968227236801
```

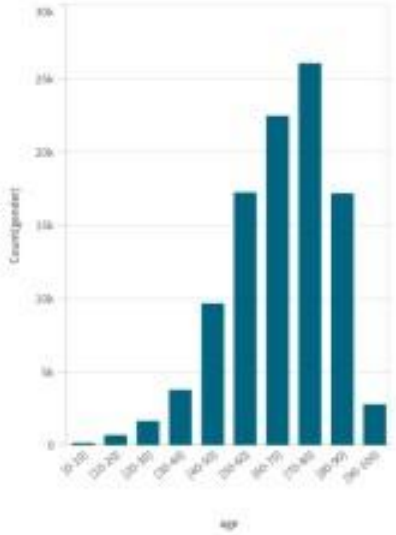
BI Report - Qlicksense Analytics



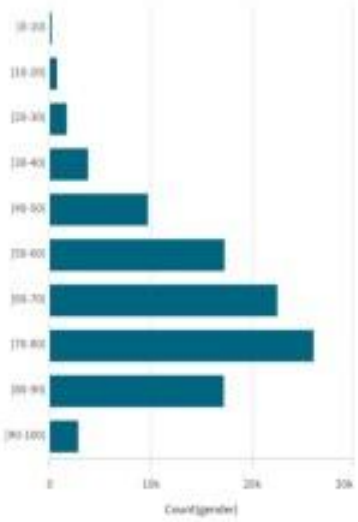
Pie Chart



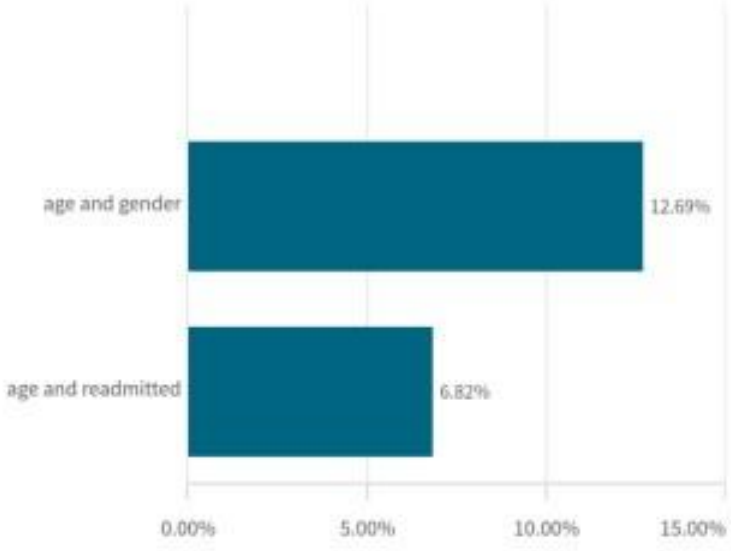
Line Chart



Vertical Bar Graph



Horizontal Bar Graph



Mutual Dependecies between - Age & Gender as well as Age & Readmitted

CONCLUSION

In conclusion, patient readmission prediction is a valuable approach that leverages data and machine learning techniques to estimate the likelihood of a patient being readmitted to a healthcare facility after initial discharge. By accurately predicting patient readmission, healthcare providers can proactively intervene and allocate resources effectively to improve patient outcomes and reduce readmission rates.

Through the collection and preprocessing of relevant patient data, including demographics, medical history, diagnoses, procedures, and medications, a comprehensive understanding of factors contributing to readmission can be obtained. Feature selection and engineering techniques further enhance the predictive model's ability to identify critical factors and patterns associated with readmission