Group 3: Patient Risk Assessment & Resource Optimization

Lee Jun Xian	23202396	junxian010729@student.usm.my
Lee Zi Ting	23202363	ziting.lee@student.usm.my
Liew Yu Xuan	23202456	yuxuan98@student.usm.my

DataSet

Dataset Source: Diabetes 130-US hospitals for years 1999-2008 dataset collected from UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008)

Dataset Collected: The dataset was collected from 130 hospitals in the US between 1999 and 2008, primarily from diabetic patient visits. The data include basic patient demographics, diagnostic information, medical procedures, medications, and admission and discharge details. The main target variable is "readmitted" status, which refers to whether the patient was readmitted within 30 days, after 30 days, or not.

Data Attribute and Type: Referred to Appendix A

Problem Definition and Business Application

Machine Learning Problem

This topic is framed as a categorization task aimed at determining the probability of patient readmission. Our objective is to classify each patient visit into one of three readmission outcomes: readmission within 30 days, readmission after 30 days, or no readmission. The target variable "readmitted" status represents these three possible outcomes.

To predict this, we used multiple predictor variables, including:

- Demographic information
- Past admissions
- Medical history
- Treatment details

By accurately categorizing patients, the model will support healthcare providers in identifying high-risk individuals for proactive interventions. This will not only help reduce readmissions, but also improve patient health outcomes.

Business Application

The application focuses on leveraging predictive analytics to assist healthcare providers in identifying high-risk patients who are susceptible to readmission. The objective is to implement proactive measures to reduce overall readmission rates, improve and optimize the resources allocations and strategize patient follow-ups. There are several key features and benefits of the application. First and foremost, the application will implement a prediction model to support the decision making for targeted intervention by analyzing patient data and predicting those at high risk of readmission with prediction model. Furthermore, it will enable healthcare providers to focus on preventative care by initiating timely follow-ups to the patients who are at higher risk. Last but not least, it will allocate the healthcare providers resources' such as treatment resources, staff and beds in efficient ways as workforce and infrastructure can be strategized and utilized based on risk prediction.

Preliminary Planning

Data Preparation

In the data preparation phase, the exploratory data analysis (EDA) will be performed to understand the data. It includes basic data description, missing data handling, demographic analysis as well as categorical and numerical exploration. The preprocessing phase will focus on cleaning and transforming the data for model training such as normalizing numerical data, encoding categorical data and addressing missing values.

Experiment Setup

Each group member is responsible for an experiment.

Experiment 1: Comparing Machine Learning Algorithms

Lee Jun Xian is assigned to handle the first experiment which is comparing the machine learning algorithms. In this experiment, we are using classification which is supervised machine learning. The initial plan is using 3 classification algorithms (Decision Tree, Random Forest and Support Vector Machine) to identify the best classification model for predicting the result based on patient health data.

Experiment 2: Features Selection

Lee Zi Ting is assigned to handle the second experiment which is selecting the features. The initial plan for selected features are Chi-Squared and ANOVA, Tree-Based Feature Importance and Correlation Analysis. These methods will help to identify useful features for algorithms selected in the first experiment.

Experiment 3: Ensemble Learning

Liew Yu Xuan is assigned to handle the third experiment which is ensemble learning. There are various ensemble learning methods that will be used to improve the model performance. The methods such as Bagging, Boosting and Voting will be used for ensemble learning based on the models identified in the first experiment in the initial plan.

Appendix A

Data	Description	Target / Feature	Attribute Type
encounter_id	Unique identifier of an encounter	ID	-
patient_nbr	Unique identifier of a patient	ID	-
race	Race	Feature	Categorical
gender	Gender	Feature	Categorical
age	Age	Feature	Categorical
weight	Weight	Feature	Categorical
admission_type_id	Label for admission types (e.g., emergency, urgent, elective, etc.)	Feature	Categorical
discharge_disposition_id	Label for discharge type (e.g., home, expired, unavailable)	Feature	Categorical
admission_source_id	Label for admission status (e.g., physician referral, ER, hospital transfer)	Feature	Categorical
time_in_hospital	Number of days between admission and discharge	Feature	Numerical
payer_code	Label for payment methods (e.g., Blue Cross, Medicare, self-pay)	Feature	Categorical
medical_specialty	Type of medical specialty (e.g., cardiology, internal medicine, general practice, surgery)	Feature	Categorical
readmitted	Days to readmission (<30, >30, or No readmission)	Target	Categorical