Proposal

Research Title: Estimation and prediction of patient length of stay

Project Overview:

The aim of this research project is to develop predictive models to estimate length of stay (LOS) by analysing a variety of factors such as patient demographic characteristics, admission information and healthcare costs. We will utilise a large dataset collected from a regional hospital management system that contains admission information for 28,109 patients. The dataset addresses several key variables including age, gender, race, type of admission, total charges, and total costs. Accurately predicting length of stay will help hospitals better plan bed scheduling, staffing, and treatment planning, which in turn will help with cost forecasting and rational pricing, while understanding the factors that influence length of stay is critical for healthcare organisations to manage resources, optimise hospital operations, and improve the quality of patient care.

Study Objective:

1. To analyse the relationship between demographic characteristics and length of stay.
2. 2.Examine how various admission types and ED metrics affect the duration of stay..
3. Develop linear regression and categorical regression tree (CART) models to predict length of stay.
4. Analyse the relationship between total healthcare costs (charges and costs) and length of stay.

Methodology:

1. Feature Engineering
   a) Dealing with missing data: analysing missing values and filling in missing data using appropriate methods
   b) One-hot coding: Convert categorical variables into binary variables for regression models.
   c) Logarithmic transformation: Logarithmic transformation of continuous variables to ensure that the data is suitable for model modelling.

2. Model Development

   a) Linear Regression: Use stepwise regression to select important variables and assess multicollinearity using VIF.
   b) CART: Build a classification and regression tree model to capture non-linear relationships.

Model Evaluation:

1. 70-30 Train-Test Split: The dataset will be split into 70% for training and 30% for testing, ensuring model generalization and reducing overfitting.

2. Evaluation Metrics:

   a) RMSE: Measures the difference between predicted and actual values; lower RMSE indicates better accuracy.

   b) R^2: Assesses the model's goodness of fit; higher R^2 indicates a better explanation of data variability.

3. Model Comparison: Compare linear regression and CART models to determine which better captures non-linear relationships and complex interactions.

Expected Outcome:

   a) Identify significant factors affecting length of stay and compare predictions to select the best model
   b) Provide recommendations for optimising hospital resource allocation, cost management and patient care.