Vanderbilt University Department of Biostatistics Dataset for SUPPORT2

PROJECT ANALYSIS PLAN

PROJECT BY:
Muhammad Ahmer: b00087698
Abdullah Abdullah: b00088619
Koushal Parupudi: b00087520

SUPERVISED BY:
Dr. Hana Sulieman

Spring 2024

Step #1: Data Pre-processing

Data pre-processing consists of data cleaning, transformation, normalization, and encoding. The dataset selected contains missing values, hence they must be accounted for either required to be dropped or replaced depending on the number of missing entries. Outliers will need to be handled if they affect the model heavily. The data contains a homogeneous mix of categorical and numerical data, hence we will be transforming, normalizing, and encoding all the data. Normalization will allow us to scale data to put all the numerical data on an equal scale. Lastly, we will use encoding methods such as one-hot encoding to transform categorical data into numerical data, for certain procedures.

Step #2: Exploratory Data Analysis (EDA)

For the EDA process, we will be exploring the dataset's shape, contents, trends, and relevant statistical data, such as mean, and variance. Plots will be used to visualize the data, such as scatterplots, histograms, and correlation matrix/heatmaps, giving us a better understanding of the data, and implicit trends or dependencies.

Step #3: Feature Engineering & Analysis

The feature engineering process will consist of selecting features that provide the most effect on the model, without increasing model complexity by adding redundant or irrelevant features. To do this, we will utilize machine learning dimensionality reduction techniques such as principal component analysis (PCA), as well as domain knowledge by removing columns that are irrelevant in the effect on the model. After building the model, we may perform the best subset feature selection method and compare the selection results. Chi-squared hypothesis testing will also be used to assess association between categorical attributes

Step #4: Linear Regression of Medical Charges

Linear regression-based model building and analysis will be employed on the predictor selected after removing irrelevant features, focused on the target variable medical charges. The procedure will consist of using OLS regression to fit the model and estimate the coefficient values. Residual analysis will be employed to test any assumptions, and remedial transformations will be applied as necessary. Further, tests for multi-collinearity will also be employed, to test correlation. Hypothesis testing will be used as needed, as well as F-tests for multiple regression.

Step #5: Binary Classification of Patient Death

Logistical regression-based model building and analysis will be employed. The selected predictors will be used to build the model, and the target variable will be a binary attribute of patient death, hence our model will be predicting patient death. We will be utilizing traditional machine learning techniques such as train_test_split to train and fit the model, as well as hyper-parameterization to prevent model overfitting or underfitting. Model performance will be evaluated using traditional evaluation metrics such as accuracy, recall, F1-score, and ROC curve analysis.