Length of Patients' Stay Prediction

Jack Lin, Jianing Pei, Qixuan Chen

Introduction

During the Coronavirus pandemic, the question of how to allocate medical resources efficiently becomes one of the pressing concerns. In our project, we attempt to predict the length of each patient's stay at hospital based on the features about hospitals and patients. The length of stay is divided into 11 classes. The first 10 classes correspond to 0-10 days, 11-20 days...91-100 days respectively, and the last class is more than 100 days.

Analysis plan

AV : Healthcare Analytics II dataset contains 318438 training samples and 137058 test samples. Each individual sample consists of 11 relevant features (17 features total), including severity of symptoms and hospital region, and one label for the length of the patient's stay at the hospital.

We will apply three different ML models to our data, namely, K-nearest Neighbors Algorithm, Logistic Regression, and Decision Tree. We will be tuning different hyperparameters for each model to fit our training dataset. We expect the training accuracy to be at least 75% in order to be useful for bed allocation planning.

Link to Dataset: https://www.kaggle.com/nehaprabhavalkar/av-healthcare-analytics-ii

Contributions

KNN: Qixuan Chen
Logistic Regression: Jianing Pei
Decision Tree: Jack Lin

In the end, all three group members will compare the test accuracies of the three ML models and choose the one that fits our dataset best. We will then collectively further optimize the chosen model's hyperparameters to improve model performance.