Proposed Solution

Recent Covid-19 Pandemic has raised alarms over one of the most overlooked area to focus: Healthcare Management. While healthcare management has various use cases for using data science, patient length of stay is one critical parameter to observe and predict if one wants to improve the efficiency of the healthcare management in a hospital.

For analysis of data to be efficient we need to use the datatset in a significant manner.This is done by splitting the data into training and testing in order to analyze the accuracy of the model. In order to categorize efficiently we need to find the count of patients in each duration category.The major factors that determine length of stay are age and severity of illness.These should be utilized properly.

Data Preprocessing is the subsequent phase.Every parameter in the dataset such as Hospital type code ,Department ,Ward Type , Severity of Illness , etc are identified induvidually and their values are stored in a seperate list respectively.Subsequently a confusion matrix is plotted with the above data.The values in the confusion matrix should be analyzed and utilized efficiently.
Further, the training and test data can be used for implementation in several algorithms such as KNN(K nearest neighbour),Logistic Regression ,Gaussian NB in order to determine the accuaracy in these different kind of models.We can use this data to find the model with the best accuaracy.

Other evaluation metrics such as precision and recall can also be determined for that particular model.For validation, conversion of data from object data to numerical data is required and this can be done using label encoding.This process is useful for feature selection as data needs to be digitized and any label encoder can be used. Several different algorithms are used constantly using the selected features in order to accurately determine the length of stay without the slightest error.