Name :Swarna Hemanth
UbNumber:50559715

MGS670 – Healthcare Analytics Electronic Health

1. Dataset description:

In this project, I worked with a dataset called *canada_covid_merge*, which combines two sheets: *Data-at-admission* and *Hospital-length-of-stay*. The merged dataset contains 494 rows and 65 columns, capturing essential patient information like demographics, physiological measurements, and comorbidities, along with the target variable—Hospital Length of Stay (HLOS).

The features in the dataset are quite diverse, including age, sex, height, weight, and various health indicators such as blood pressure, heart rate, and laboratory results. Notably, I identified key features that significantly correlate with HLOS, like age, sex, and certain laboratory measurements. These insights are crucial for understanding factors that influence hospital stays, ultimately helping healthcare professionals make informed decisions.

During the data cleaning phase, I focused on ensuring high-quality data for analysis. I tackled missing values, eliminated irrelevant columns, and applied imputation techniques to fill gaps in the data. Specifically, I removed features with all-zero values and filled numerical columns using median imputation. For categorical columns, I used the mode to replace missing values. This thorough cleaning process helped refine the dataset, making it ready for modeling. And I have used knn imputation technique to fill the null values in the data set

2. Features Importance and Ideas

In the final model, I utilized 15 key features selected for their strong correlations with Hospital Length of Stay (HLOS). These features include age, sex, height, weight, systolic and diastolic blood pressure, heart rate, respiratory rate, oxygen saturation, temperature, WBC (white blood cells) count, RBC (red blood cells) count, hemoglobin, platelet count, and whether the patient was intubated (binary). The decision to drop other features was based on their weaker relationships with HLOS or the presence of excessive missing values. By focusing on these 15 features, the model effectively captures the most relevant physiological and demographic factors influencing hospital stays, which enhances its predictive accuracy and overall performance. This strategic feature selection helps ensure that the model remains robust while minimizing noise from less significant data.

Comorbidities, such as hypertension and diabetes, significantly impact hospital length of stay due to their influence on treatment complications. Additionally, vital signs like systolic blood pressure, oxygen saturation, and respiratory rate are crucial for assessing a patient's condition during hospitalization. These measurements provide insights into the patient's medical status, helping to predict their length of stay by indicating the severity of their health decline.

3. Model Implementation

Multilayer Perceptron Model : Data: Data Admission and Hospital Length of the stay is the dataset for the current study, the dataset was divided ina way that 80% of it was used in training while 20% in testing.

After preprocessing the data, we selected the most important features to predict hospital length of stay. We then split the dataset into training and testing sets to evaluate the model's performance.

Next, we built a neural network model with multiple layers to learn patterns from the training data. This model was trained on the selected features, allowing it to understand the relationships between patient characteristics and their hospital stay duration.

We used metrics like Mean Squared Error (MSE) and R-squared (R^2) to assess how well the model performed. Finally, we validated the model's predictions by comparing them with the actual lengths of stay in the testing set, ensuring it could accurately estimate hospital stays based on the input data.

Name :Swarna Hemanth
UbNumber:50559715

.

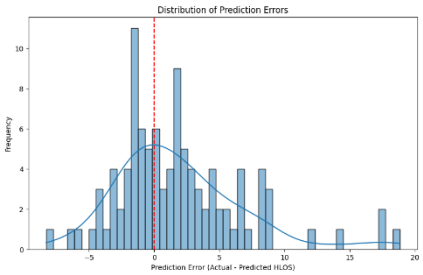
3.1 Recurrent Neural Network (RNN): For the RNN model, we first merged and preprocessed the Days Breakdown data with the Hospital Length of Stay using the parent_id. Since RNNs work best with continuous data, we created a variable called Adjusted_HLOS, which is calculated as Total HLOS minus the current day. After preprocessing, we ended up with 23 features plus the target feature, Adjusted_HLOS**.**

We used this time-series data of patients for training the model. Through experimentation, we identified the best hyperparameters for our model: hidden size of 256, two layers, a batch size of 32, and a total of 200 epochs.

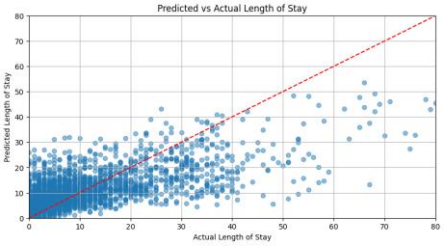
We used Mean Squared Error (MSE) as the performance metric. After running five MSE trials with different parameters, we calculated the mean MSE and its standard deviation. The results are summarized below.

4. Model Performance

MLP:
Mean MSE: 31.4348
Standard Deviation of MSE: 6.6895
Mean R^2: 0.7718
Standard Deviation of R^2: 0.0417



RNN:
Mean MSE: 138.2673
R^2 Score: 0.2165
Std Dev Test MSE: 26.4851



Scope of Future Work

Feature Engineering**:** In the future, we could improve the model by creating new features that combine comorbidities and vital signs. These combinations might reveal important relationships that aren't clear when looking at each variable on its own.

Hyperparameter Optimization**:** We can also enhance the model by fine-tuning its hyperparameters. Previous adjustments in the RNN model showed better results. By optimizing the learning rate, dropout rates, and the depth of the network, we could boost both the model's accuracy and overall performance.

For future work, we can enhance the model by incorporating additional features, such as interaction terms between comorbidities and vital signs. Additionally, exploring more advanced algorithms and optimizing hyperparameters further could improve accuracy and performance, providing deeper insights into hospital length of stay predictions.

Conclusion:

This project aimed to predict the Hospital Length of Stay (HLOS) for COVID-19 patients by utilizing machine learning (MLP) and deep learning (RNN) techniques. We successfully implemented both models, with the RNN showing promising results due to its ability to capture time series trends in the data. However, while the RNN's accuracy was commendable, there's still potential for improvement to reach optimal performance. Future efforts could focus on refining the model, enhancing feature engineering, and optimizing hyperparameters to further increase prediction accuracy and provide better insights into patient care and hospital resource management.

Name :Swarna Hemanth
UbNumber:50559715