Prediction of the Cost of Hospital Inpatient People based on their Treatment done by the Hospital

DATASET INFORMATION:

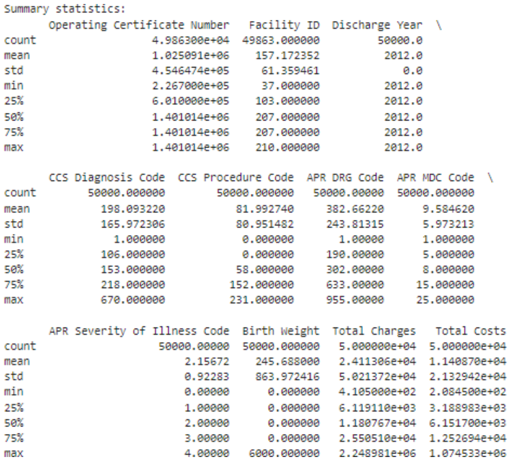
Dataset Name: **Hospital Inpatient Discharges (SPARCS De-Identified): 2012**
This data contains basic record level detail regarding the discharge of the patients and the cost and charges that are applied to them after being discharged from the hospitals. The data is provided by the New York State Department of Health.
Our aim is to predict the cost that is applied to them after being discharged based on the treatment which they have undergone and medicines/drugs which are used by them. It is a Regression Problem to predict the Total Cost based upon the treatment and other related features.
We have numerical and categorical features in our dataset. As it is a large dataset which has 2544543 rows, for our convenience I have sampled the dataset to 50000 rows.

Statistics about the Sample dataset:

Number of entries: 50000

Number of features: 34

Following is the Summary Statistics about the dataset:



```
Summary statistics:
      Operating Certificate Number   Facility ID  Discharge Year  \
count                 4.986300e+04  49863.000000         50000.0
mean                  1.025091e+06    157.172352          2012.0
std                   4.546474e+05     61.359461             0.0
min                   2.267000e+05     37.000000          2012.0
25%                   6.010000e+05    103.000000          2012.0
50%                   1.401014e+06    207.000000          2012.0
75%                   1.401014e+06    207.000000          2012.0
max                   1.401014e+06    210.000000          2012.0

      CCS Diagnosis Code  CCS Procedure Code  APR DRG Code  APR MDC Code  \
count       50000.000000        50000.000000   50000.00000  50000.000000
mean          198.093220           81.992740     382.66220      9.584620
std           165.972306           80.951482     243.81315      5.973213
min             1.000000            0.000000       1.00000      1.000000
25%           106.000000            0.000000     190.00000      5.000000
50%           153.000000           58.000000     302.00000      8.000000
75%           218.000000          152.000000     633.00000     15.000000
max           670.000000          231.000000     955.00000     25.000000

      APR Severity of Illness Code  Birth Weight  Total Charges   Total Costs
count                  50000.00000  50000.000000   5.000000e+04  5.000000e+04
mean                       2.15672    245.688000   2.411306e+04  1.140870e+04
std                        0.92283    863.972416   5.021372e+04  2.132942e+04
min                        0.00000      0.000000   4.105000e+02  2.084500e+02
25%                        1.00000      0.000000   6.119110e+03  3.188983e+03
50%                        2.00000      0.000000   1.180767e+04  6.151700e+03
75%                        3.00000      0.000000   2.550510e+04  1.252694e+04
max                        4.00000   6000.000000   2.248981e+06  1.074533e+06
```

There are some missing values in the column of APR Mortality Rate. As it has many rows, the rows are dropped as there are very few missing values.

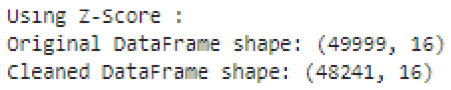
1. What kind of preprocessing techniques have you applied to this dataset?

   The preprocessing Techniques which i have applied are:

   a. Handle missing entries

   b. I have checked for handling mismatched string formats, but there aren’t any such in the dataset.

   c. Handle outliers: Using Z-Score method I have handled outliers and removed them from the dataset.



```
Using Z-Score :
Original DataFrame shape: (49999, 16)
Cleaned DataFrame shape: (48241, 16)
```

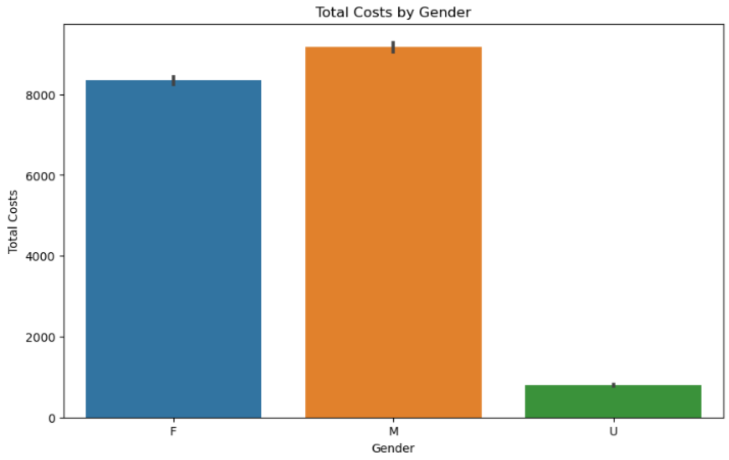
   d. Converting the Age Group into Numerical Values: I have performed this as they are ordered in ranges, and I have made them numerical such that it will be convenient to access the Age Group column easily.

   e. One Hot Encoding, creating binary columns for each category, denoting their presence or absence.

VISUALIZATIONS:

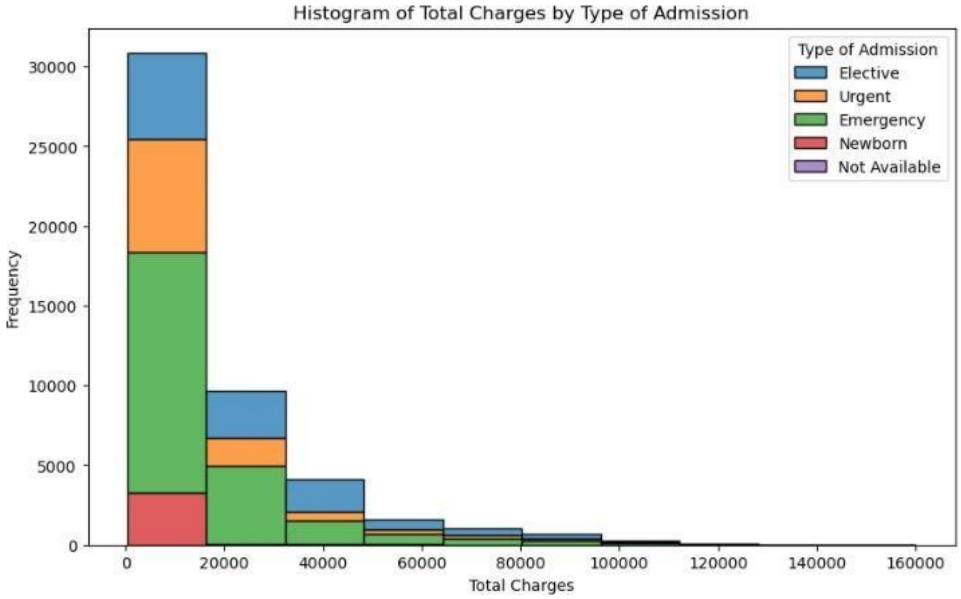
i. Total Costs by Gender:

From this graph we can observe the charges applied by the hospitals based on Gender. This graph also tells us from which Gender the hospitals are earning more. From the sample dataset we can observe that Male contribute more when compared with other Genders.



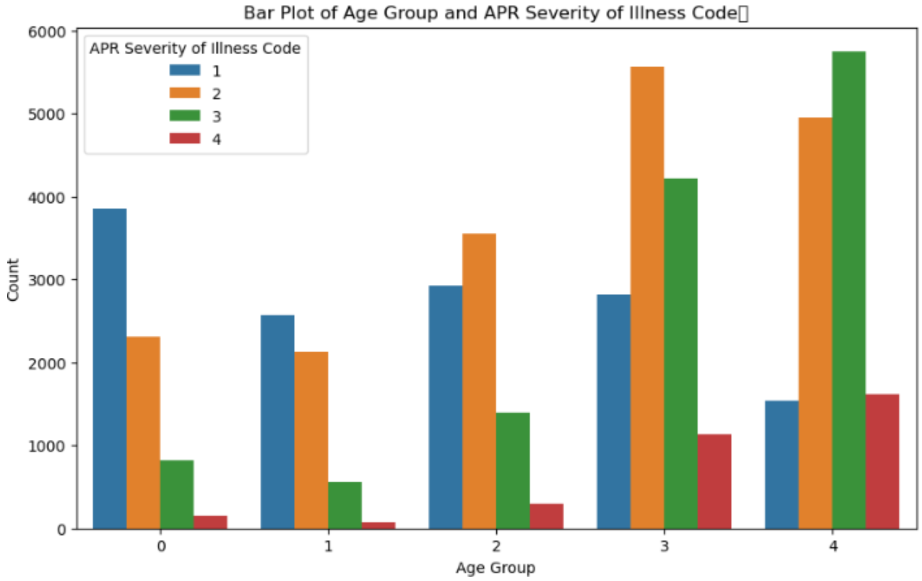
ii. Total Charges by Type of Admission:

From this graph we can observe that based on type of admissions, how much the hospitals are charging on their patients. From the below sample dataset graph, we can observe that Emergency category patients are being charged more when compared with other categories.



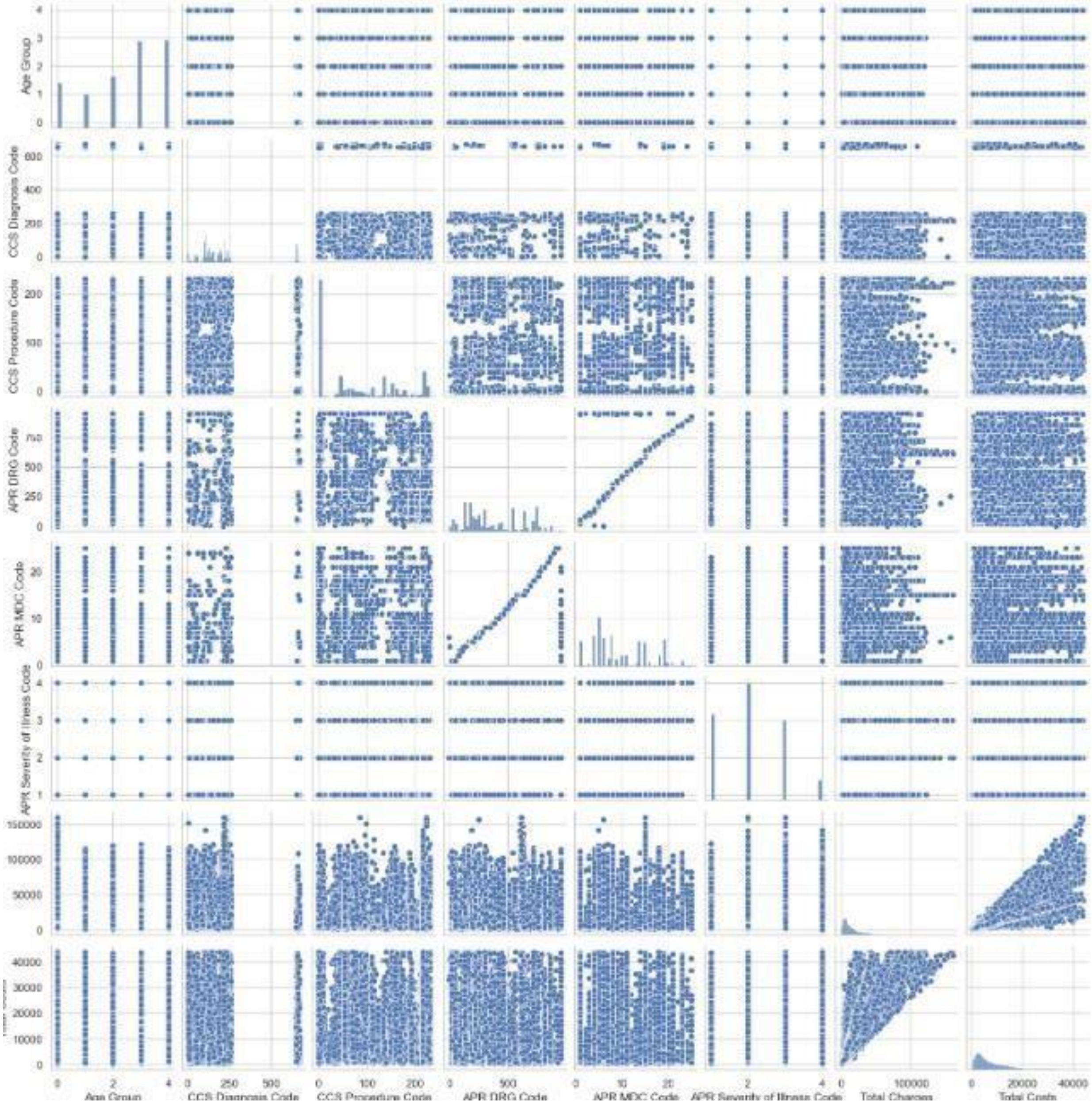
iii. Age group and APR Severity Illness Code

From this graph we can observe which age group people are facing which kind of illness. From this it is easy to understand which age group people are being affected to which disease. **The codes 0, 1, 2, 3 are mentioned in the Dataset Information.**



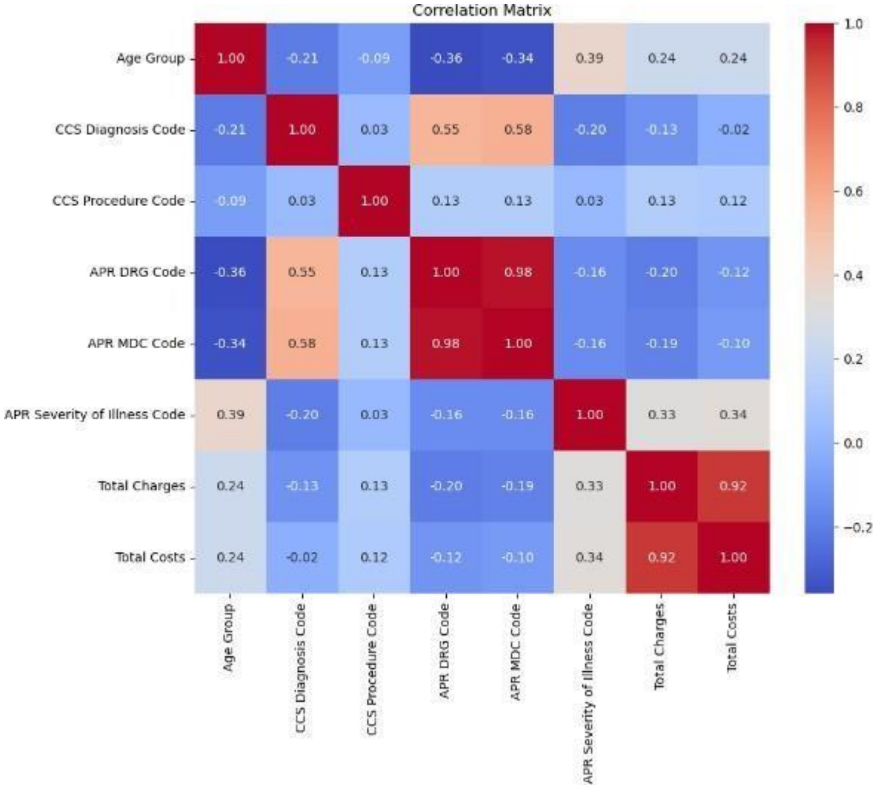
iv. Pair Plot between all the features:

From the below Pair Plot graph, we can observe all the relations between every feature and decide which features can be used and which can be dropped accordingly.



v. Correlation Matrix:

Following is the Correlation Matrix for all the Numerical Features and identifying the uncorrelated or unrelated features and dropping those features with a low correlation coefficient are be identified and subsequently dropped from the dataset to increase the model performance.



MACHINE LEARNING ALGORITHMS:

I have applied 3 ML Algorithms, and they are:

i. **Linear Regression:**

Brief Description: It is a method used for modeling the relationship between a dependent variable and one or more independent variables.

Mathematical Representation:

The basic equation for linear regression is $y = mx + c$.

Key Features: It's easy to interpret.

Advantages: Simplicity, computationally efficient.

Disadvantages: It is sensitive to outliers.

ii. **Gradient Boost:**

Brief Description: This estimator builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage a regression tree is fit on the negative gradient of the given loss function.

Mathematical Representation: $L(y, F\_k(x)) = \sum_{i=1}^{n} L(y\_i, F\_k - 1(x\_i) + h\_k(x\_i))$

Key Features: It's highly flexible and can handle various types of data.

Advantages: It's robust to outliers and performs well in practice, Produces highly accurate predictions.

Disadvantages: It's computationally expensive.

iii. **Random Forest Regressor:**

Brief Description: Random forests (RF) construct many individual decision trees at training. Predictions from all trees are pooled to make the final prediction.

Mathematical Representation: $y(x) = (1/T) \sum_{t=1}^{T} f\_t(x)$

Key Features: It's highly accurate, robust to overfitting.

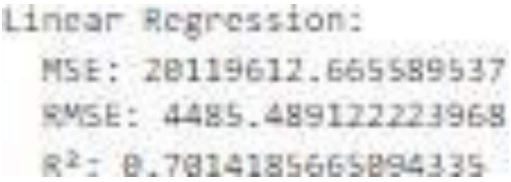
Advantages: Provides high accuracy.

Disadvantages: It is slow to evaluate for large datasets.

2. Provide your loss value and accuracy for all 3 methods.

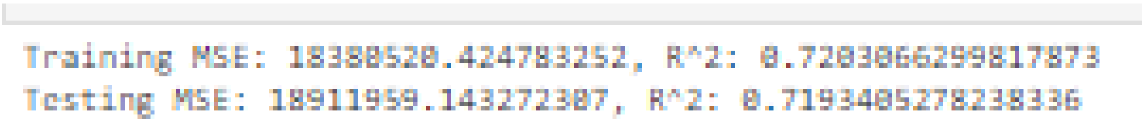
Below are the loss value and accuracy of all the 3 methods: (MSE and R^2 values are mentioned)

1. **Linear Regression:** R^2 value is 0.70



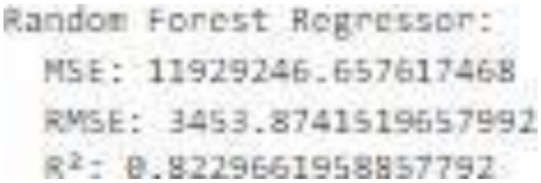
```
Linear Regression:
   MSE: 20119612.665589537
   RMSE: 4485.489122223968
   R²: 0.7014185665094335
```

2. **Gradient Boosting:** R^2 value of training is 0.72 and R^2 value of testing is 0.71



```
Training MSE: 18380520.424783252, R^2: 0.7203066299817873
Testing MSE: 18911959.143272307, R^2: 0.7193405278238336
```

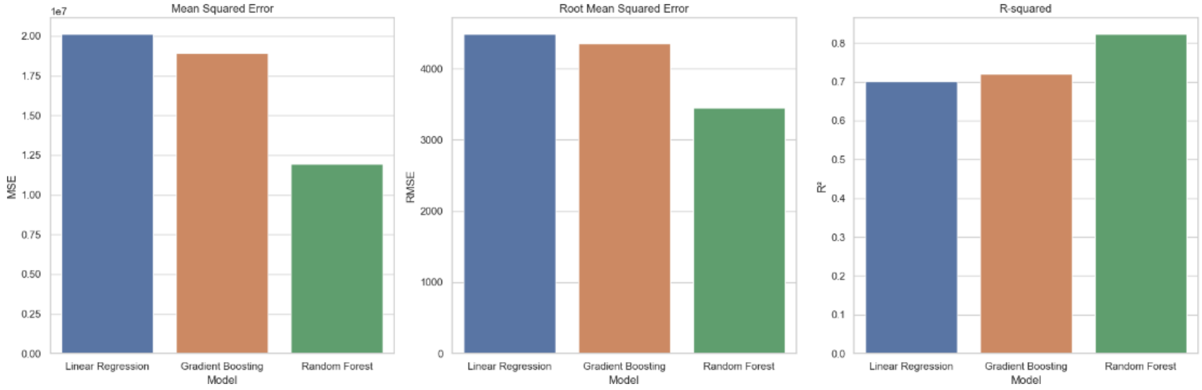
3. **Random Forest Regressor:** R^2 value is 0.82



```
Random Forest Regressor:
  MSE: 11929246.657617468
  RMSE: 3453.8741519657992
  R²: 0.8229661958857792
```

COMPARING MODEL PREDICTIONS:

Below is the Plot comparing the predictions for all the methods used is below:



From the above graphs, we can observe that Random Forest Regressor is predicting well when compared with the other two algorithms.

NEURAL NETWORK:

Defining the NN:

1. Input Layer:

Receives input data

2. First Hidden Layer:

Type: Fully connected layer.

Output from this layer: 64 size

Activation: ReLU.

3. Second Hidden Layer:

Type: Fully connected (dense) layer.

Output from this layer: 32 size

Activation: ReLU.

4. Output Layer:

Type: Fully connected (dense) layer.

Output from this layer: 1 size

Activation Functions:

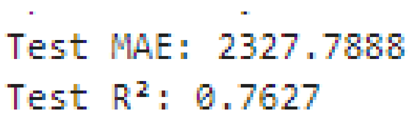
ReLU (Rectified Linear Unit): Used in the hidden layers to introduce non-linearity.

Training:

Loss Function: Mean Squared Error (MSE)

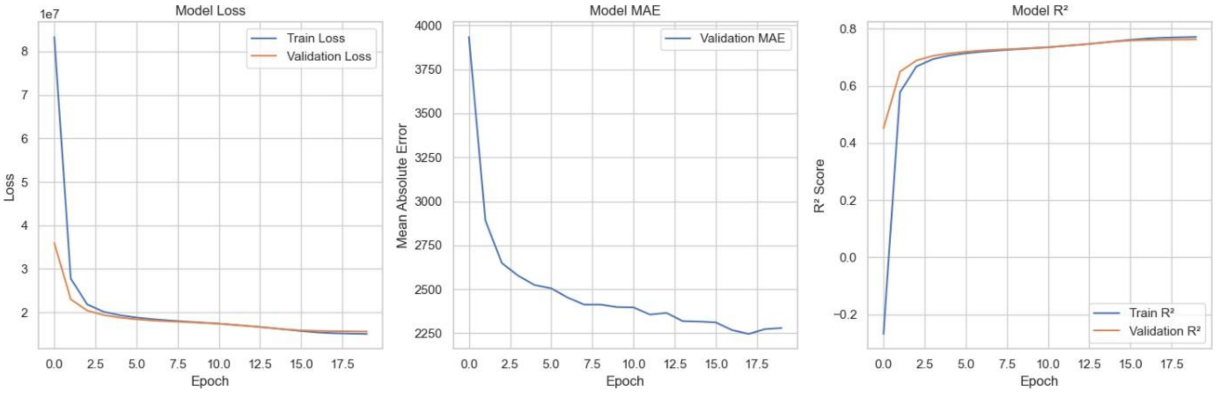
Optimization: Adam optimizer.

Results:



```
Test MAE: 2327.7888
Test R²: 0.7627
```

Plot:



By observing the plots, we can observe that the Mean Absolute Error and Model is decreasing during each epoch. I have considered 20 epochs. By observing this plot, we can say that the neural network is performing well.

Conclusion: If we compare the NN with the ML models, NN performs better than most of the ML models like Linear Regression and Gradient Boosting.

REFERENCES:

Dataset: https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De Identified/u4ud-w55t/about_data

[2] Gradient Boosting: https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/

[3] Random Forest Regressor: https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3

[4] https://scikit-learn.org/stable/supervised_learning.html