Health Care Data Set Analysis

By:

Mahmoud Mohamed Ahmed

Youssif Mohamed Mostafa

Hamza Ahmed Abdo

Supervised by:

Dr / Samya Heshmat

ENG / Abd Alrahman Yahya

Data analysis

At the beginning, we presented the data to thoroughly read and understand the content. After that, we worked on improving the data to achieve the best results.

In the first stage, we examined the data to check if it contained any missing information so we could work on improving it. Fortunately, we found that it did not contain any Null values.

After that, we identified the type of features in the data to format and convert what is useful in order to achieve better results.

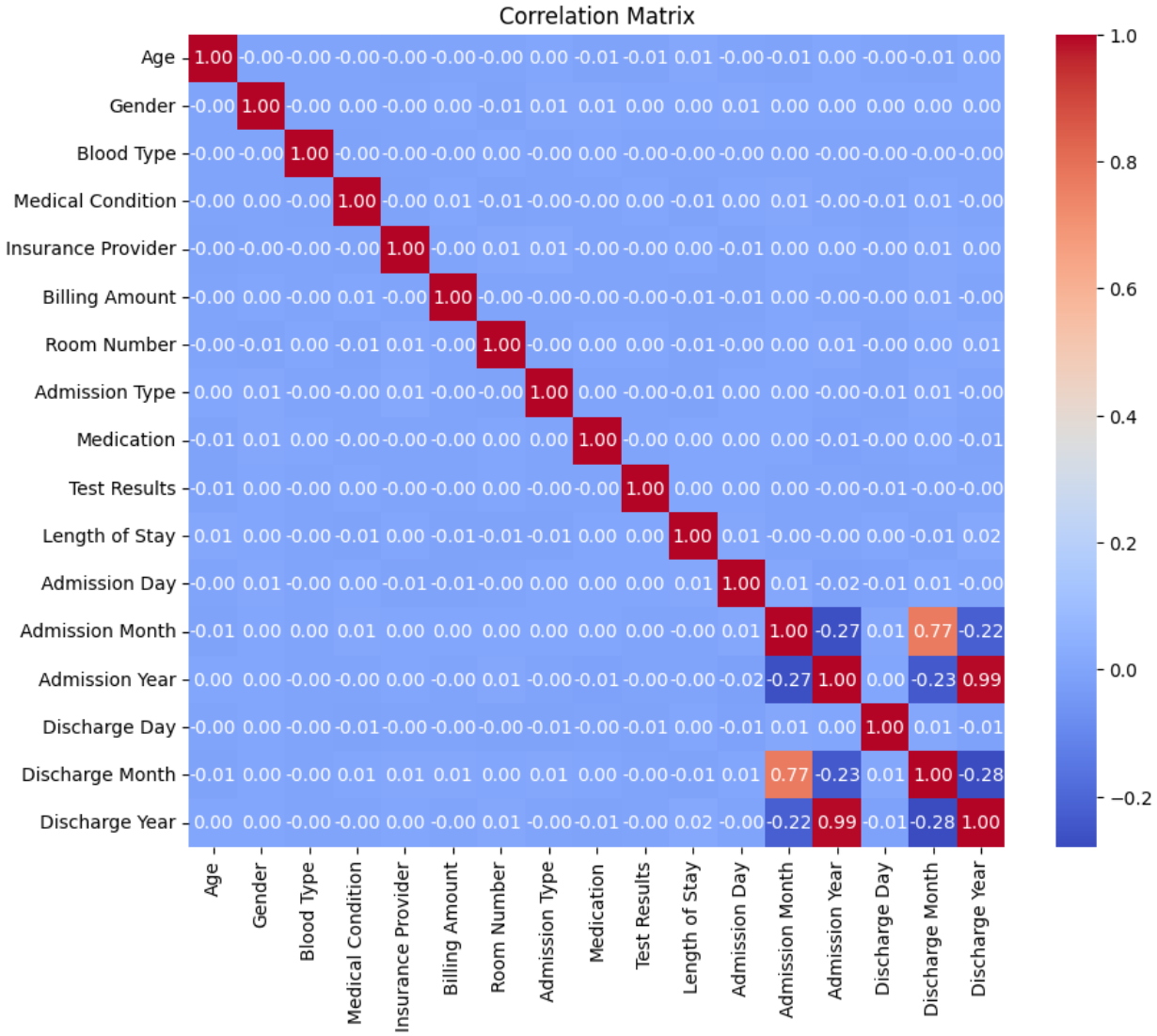
- After the examination, we found that it was best to use the Encoder method.

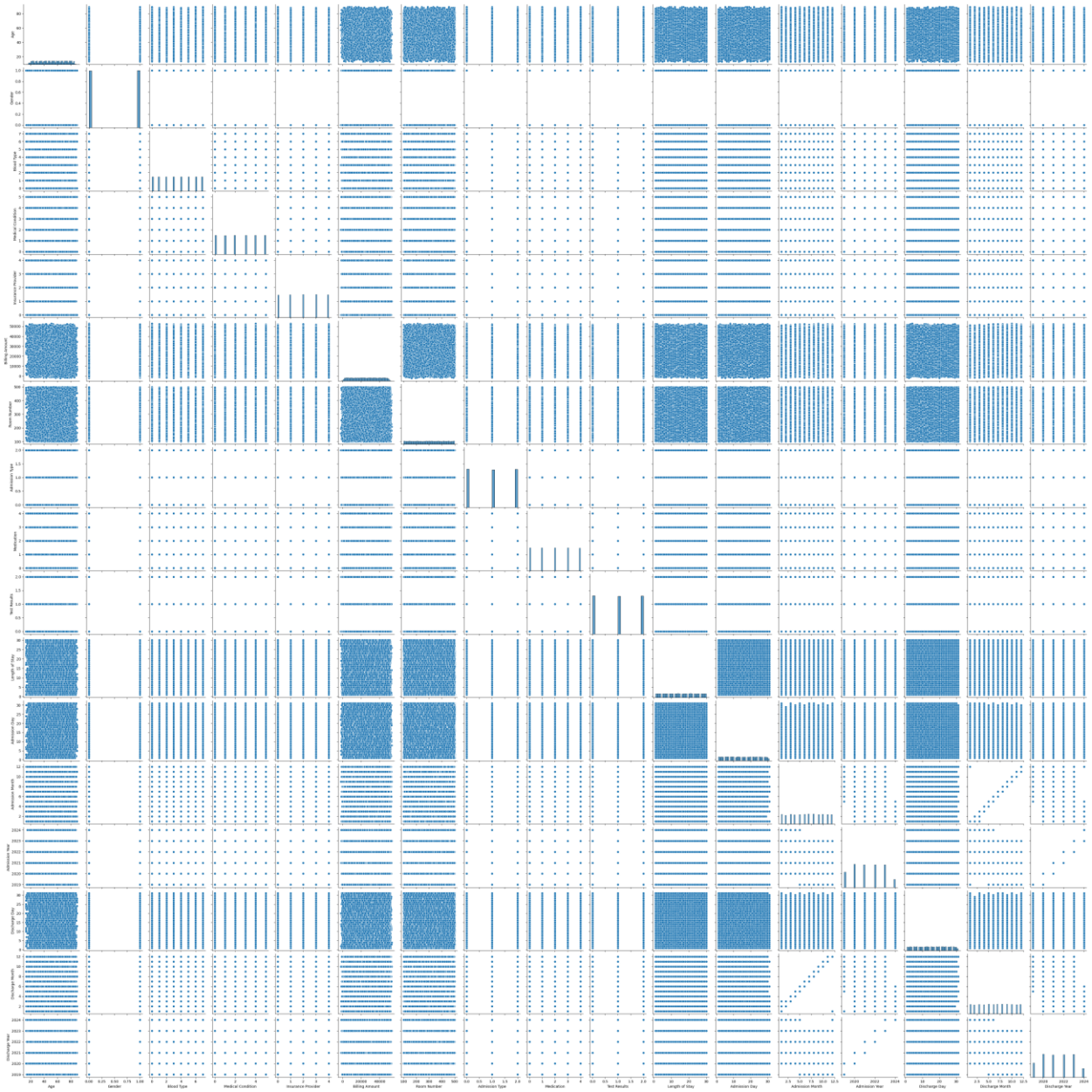
-We converted these elements to the integer type:r

Gender
Blood Type
Medical Condition
Insurance Provider
Admission Type
Medication
Test Results

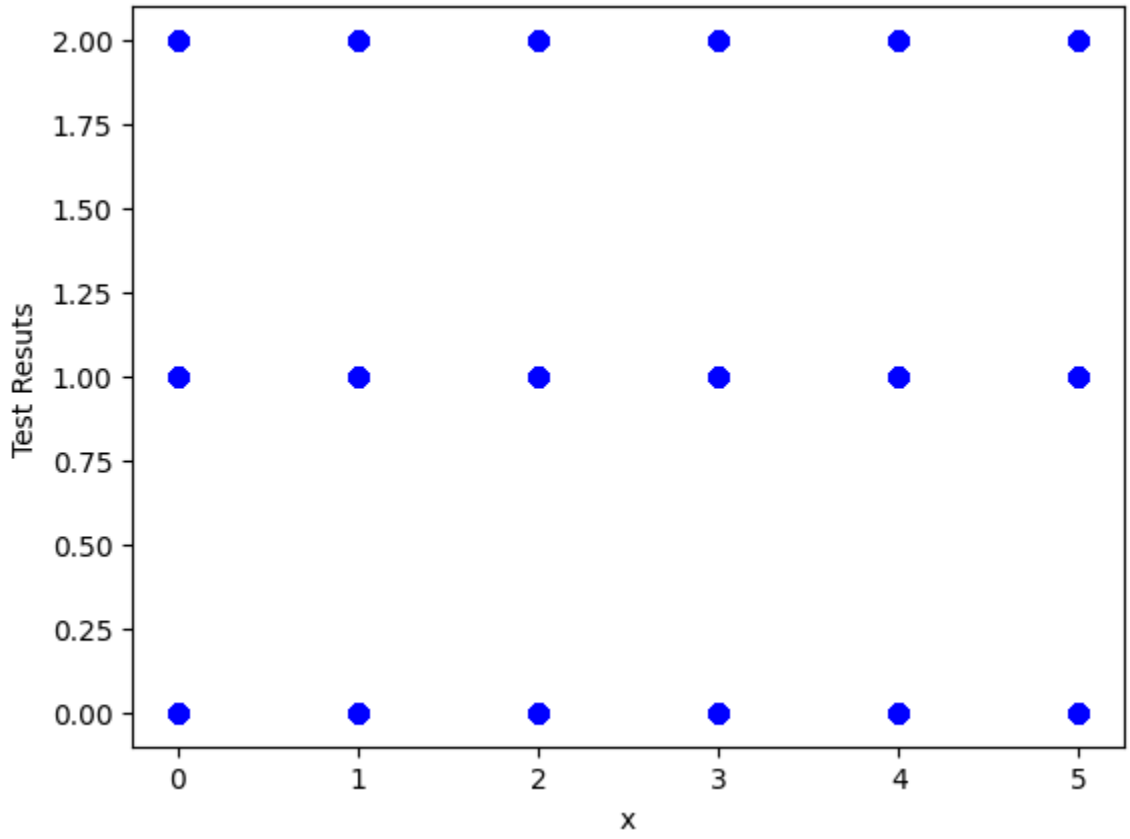
- We converted the dates to day, month, and year, and calculated the number of days the patient stayed.

- We used the Correlation Matrix function to determine the relationship between the features:




Age
Gender
Blood Type
Medical Condition
Insurance Provider
Billing Amount
Room Number
Admission Type
Medication
Test Results
Length of Stay
Admission Day
Admission Month
Admission Year
Discharge Day
Discharge Month
Discharge Year

Scatter:



- After modifying the data to achieve the best results, the data is split into Train and Test sets.

* Using Function : train_test_split

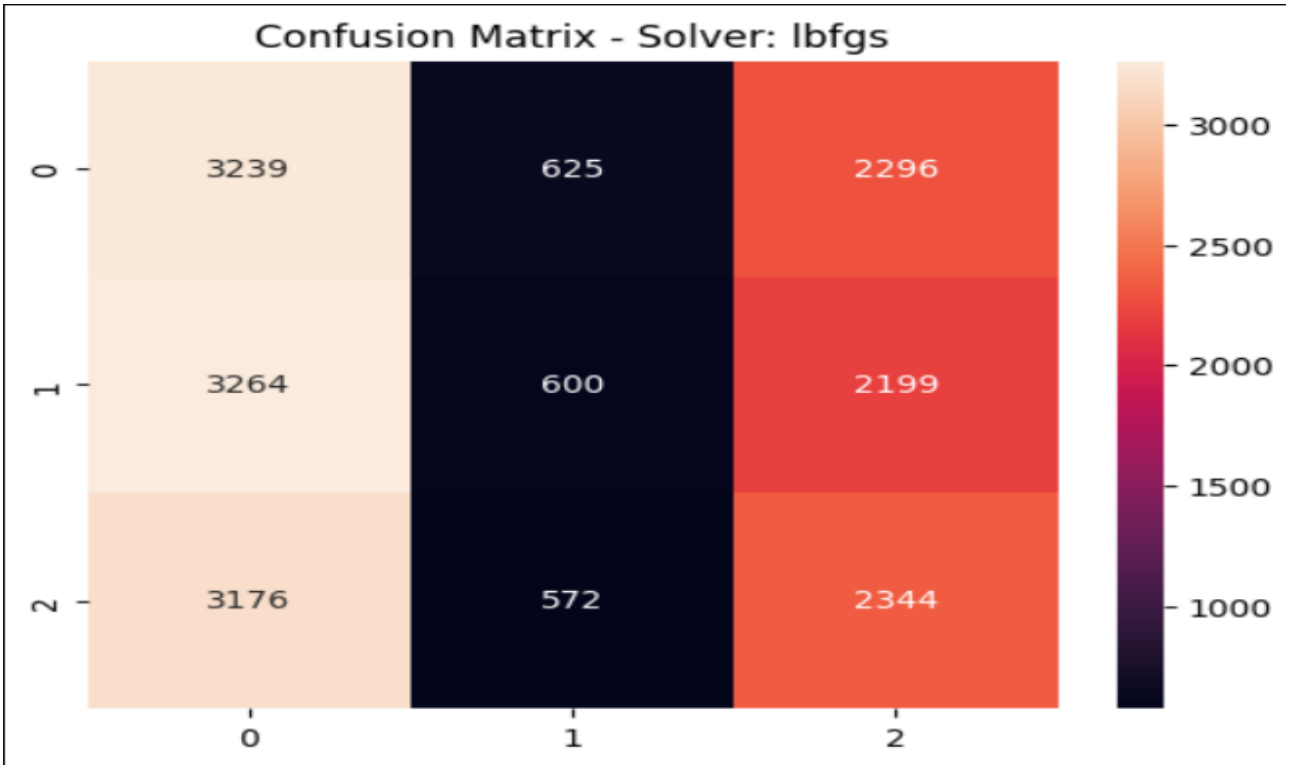
-The goal of this analysis is to explore and compare different machine learning models to identify the most effective approach for predicting test results. We experimented with a range of models, including:

1. Logistic Regression
2. Decision Trees
3. Random Forests
4. Neural Networks
5. Support Vector Machines (SVM)
6. Gradient Boosting Machines

Model name: Logistic Regression

Test#1:- Input features : 'Medical Condition' ,'Length of Stay' ,'Admission Type' , 'Age', 'Admission Day' ,'Admission Month', 'Admission Year'

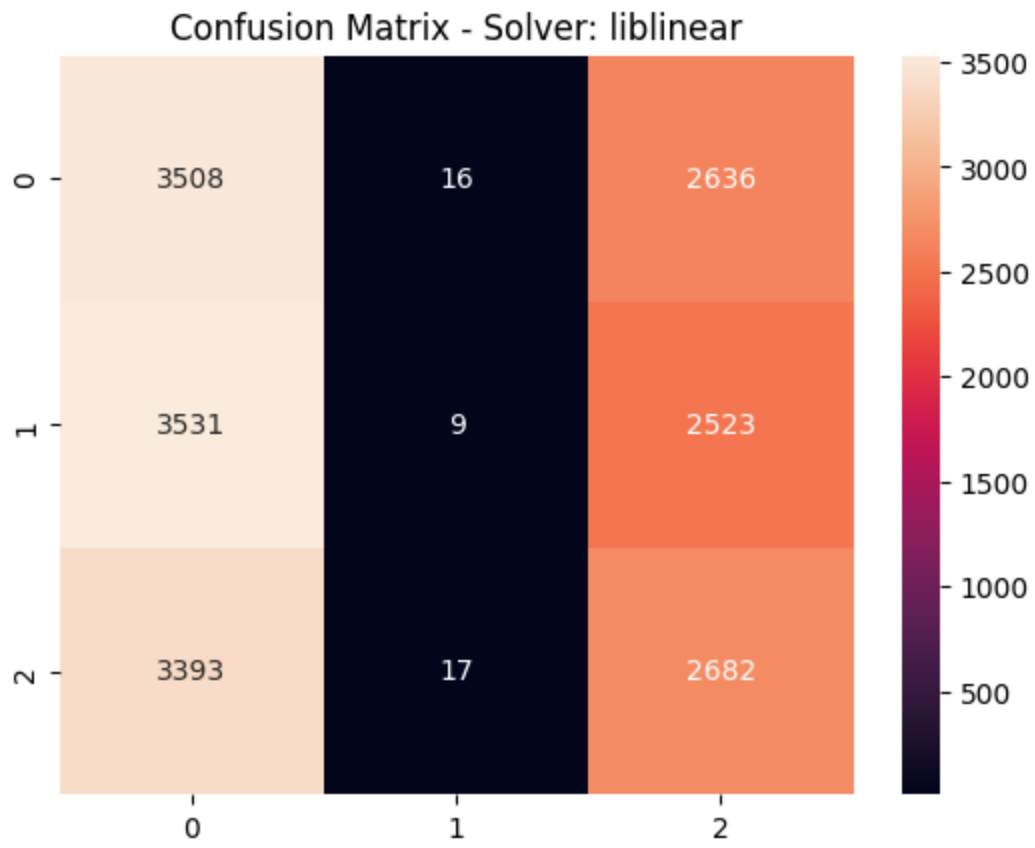
Hyperparameter Tuning and Performance Metrics :



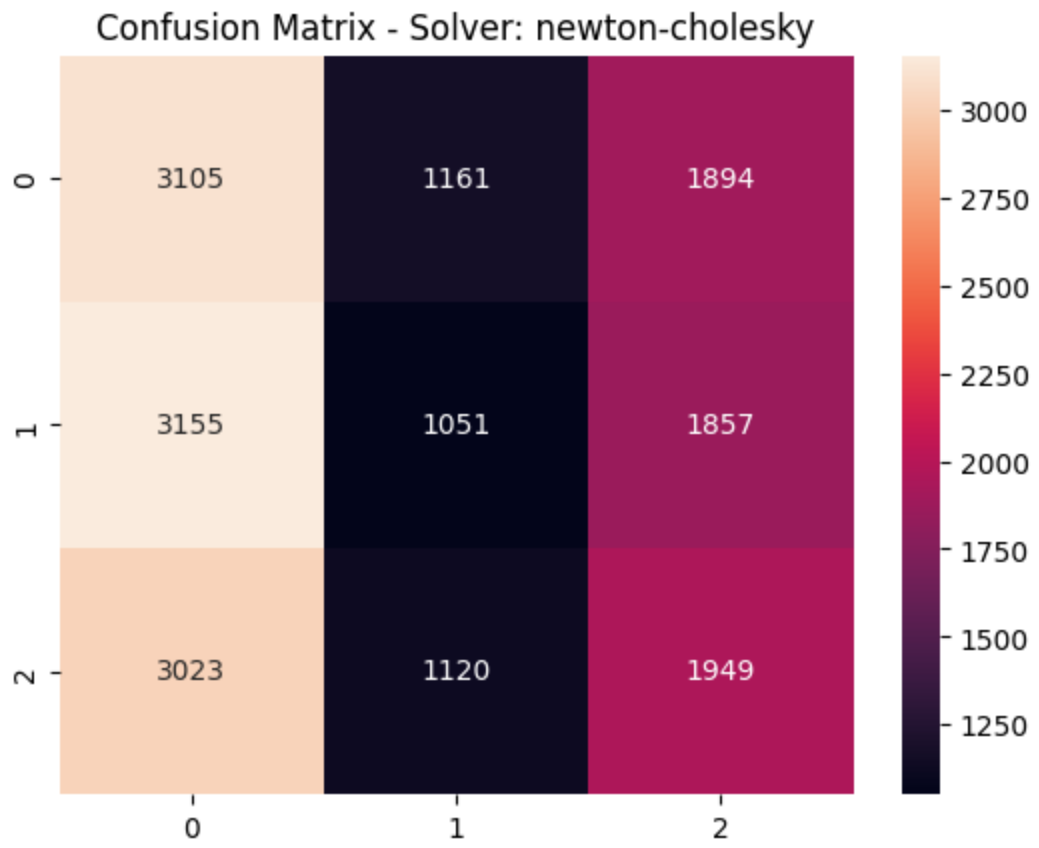
	Precision	Recall	F1-score
0	0.33	0.53	0.41
1	0.33	0.10	0.15
2	0.33	0.38	0.36
accuracy			0.34

Test#2:- Input features : 'Medical Condition', 'Length of Stay' ,'Admission Type',' Admission Day',' Admission Month'

Hyperparameter Tuning and Performance Metrics :



	Precision	Recall	F1-score
0	0.34	0.57	0.42
1	0.21	0.00	0.00
2	0.34	0.44	0.38
Accuracy			0.33

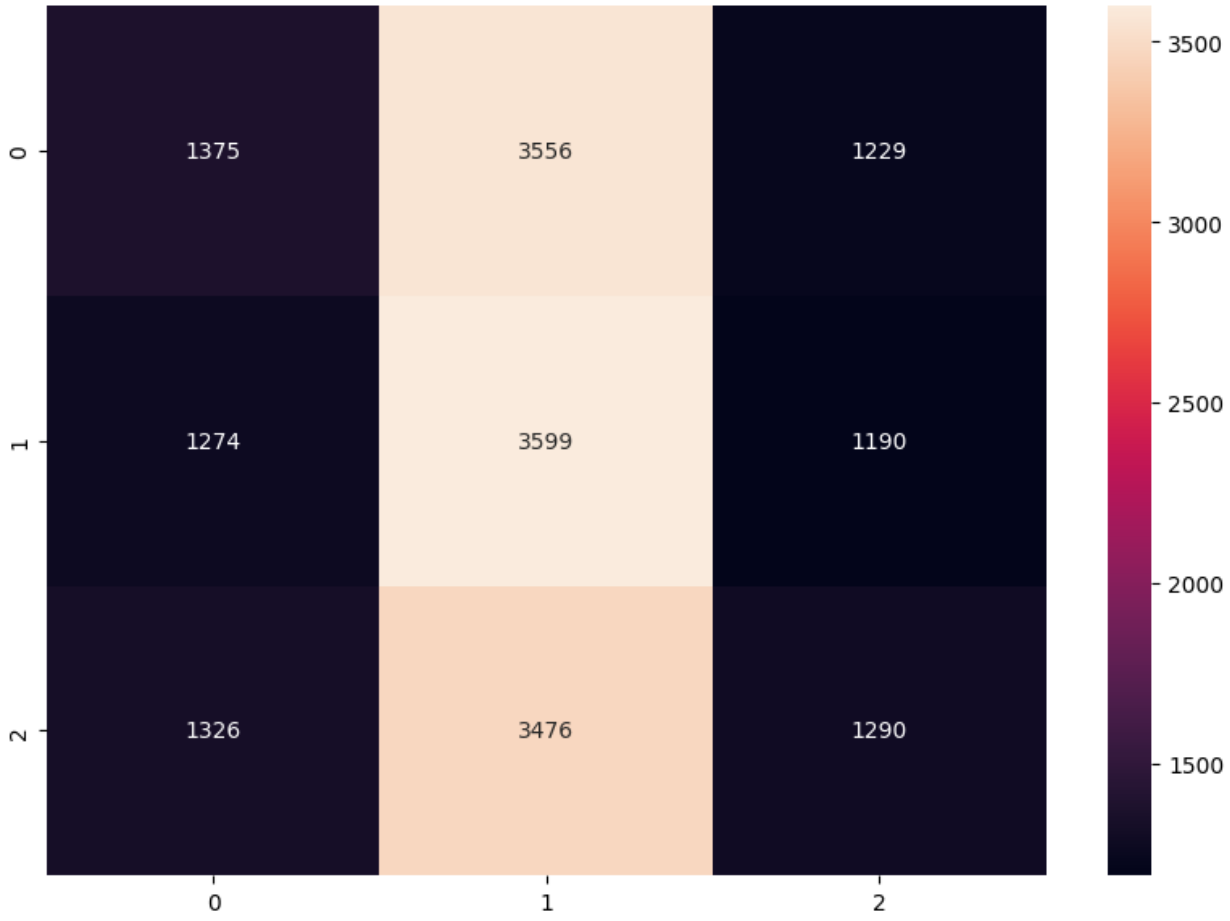


	Precision	Recall	F1-score
0	0.34	0.57	0.40
1	0.22	0.00	0.00
2	0.34	0.44	0.35
Accuracy			0.34

Model name: Decision Tree

Test#1:- Gini

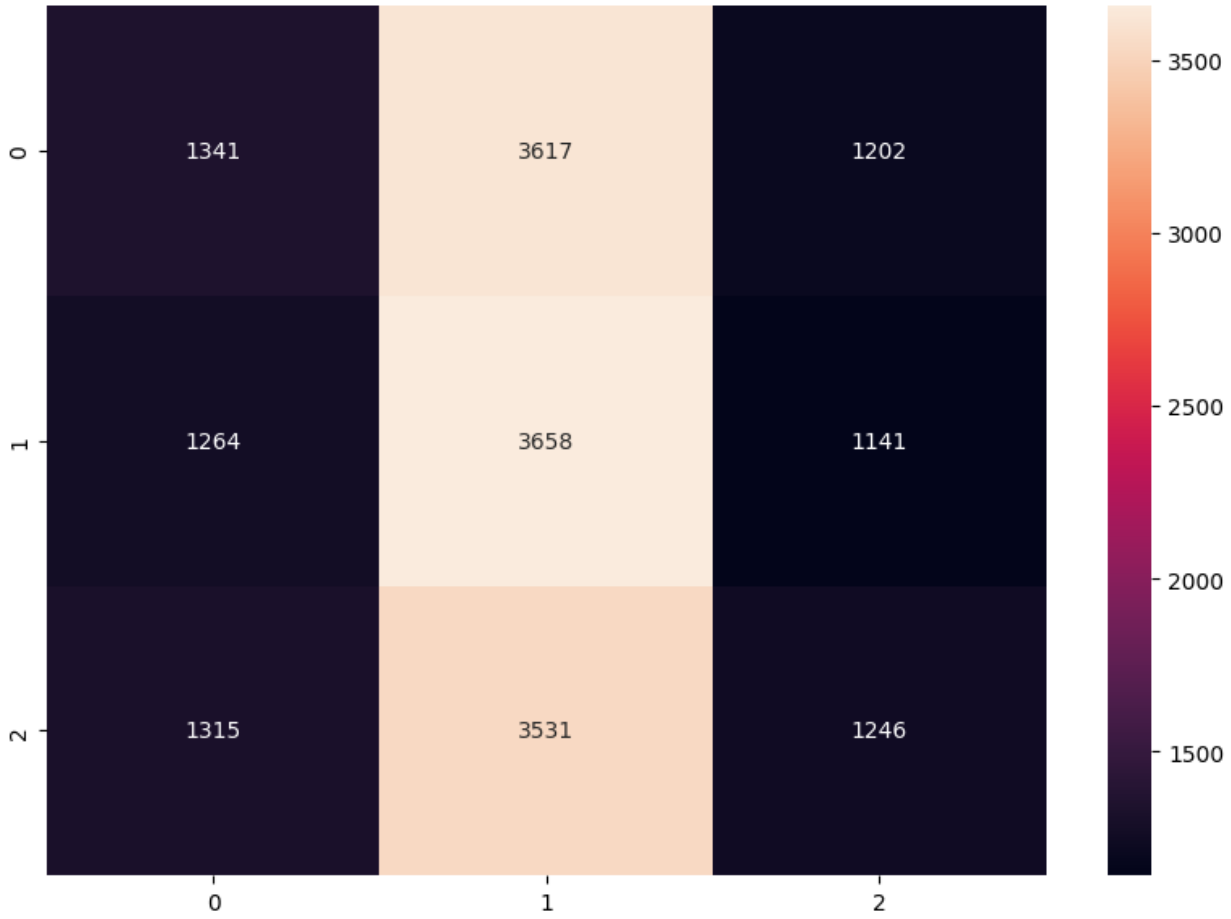
Hyperparameter Tuning and Performance Metrics :



	Precision	Recall	F1-score	Support
0	0.35	0.22	0.27	6160
1	0.34	0.59	0.43	6063
2	0.35	0.21	0.26	6092
accuracy				0.34

Test#2:- Entropy

Hyperparameter Tuning and Performance Metrics :

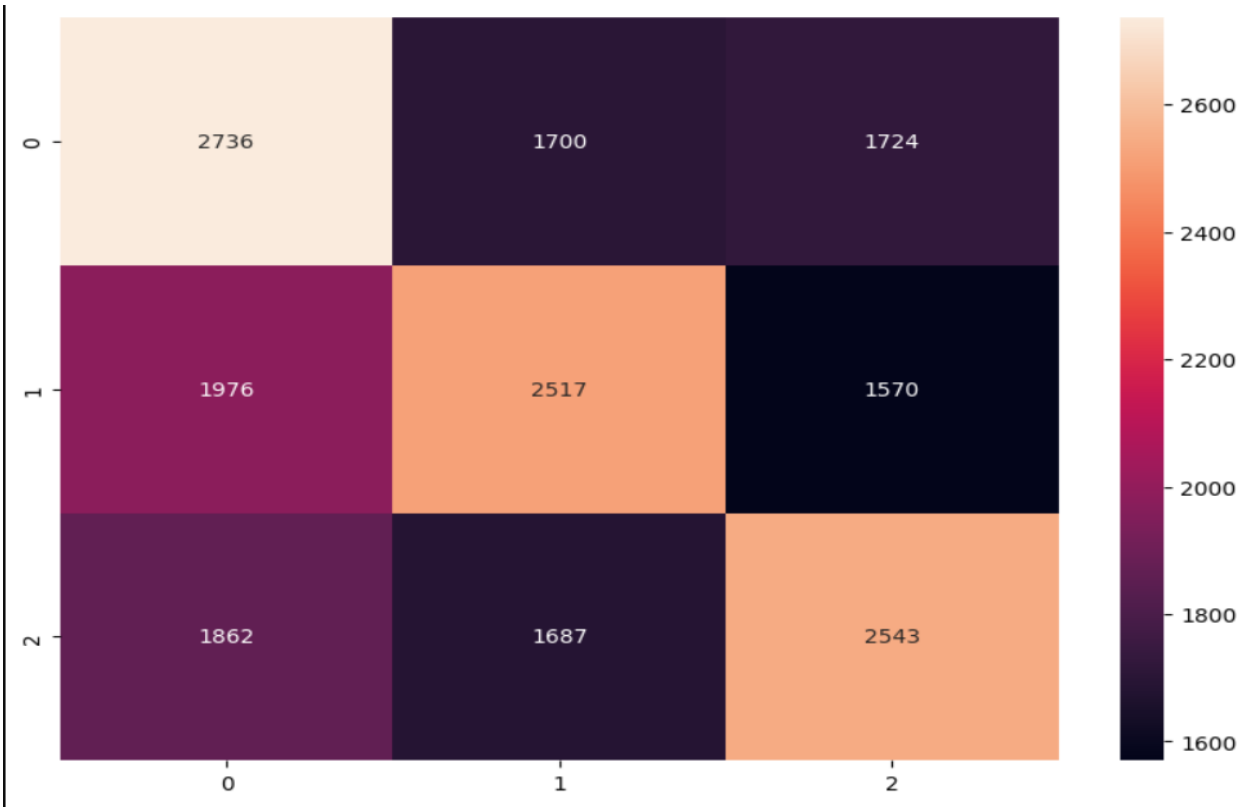


	Precision	Recall	F1-score	Support
0	0.34	0.22	0.27	6160
1	0.35	0.60	0.43	6063
2	0.34	0.20	0.26	6092
accuracy				0.34

Model name: Random Forest

Test#1:- Input features : 'Medical Condition' ,'Length of Stay' ,'Admission Type' , 'Age' ,'Admission Day' ,'Admission Month' ,'Admission Year'

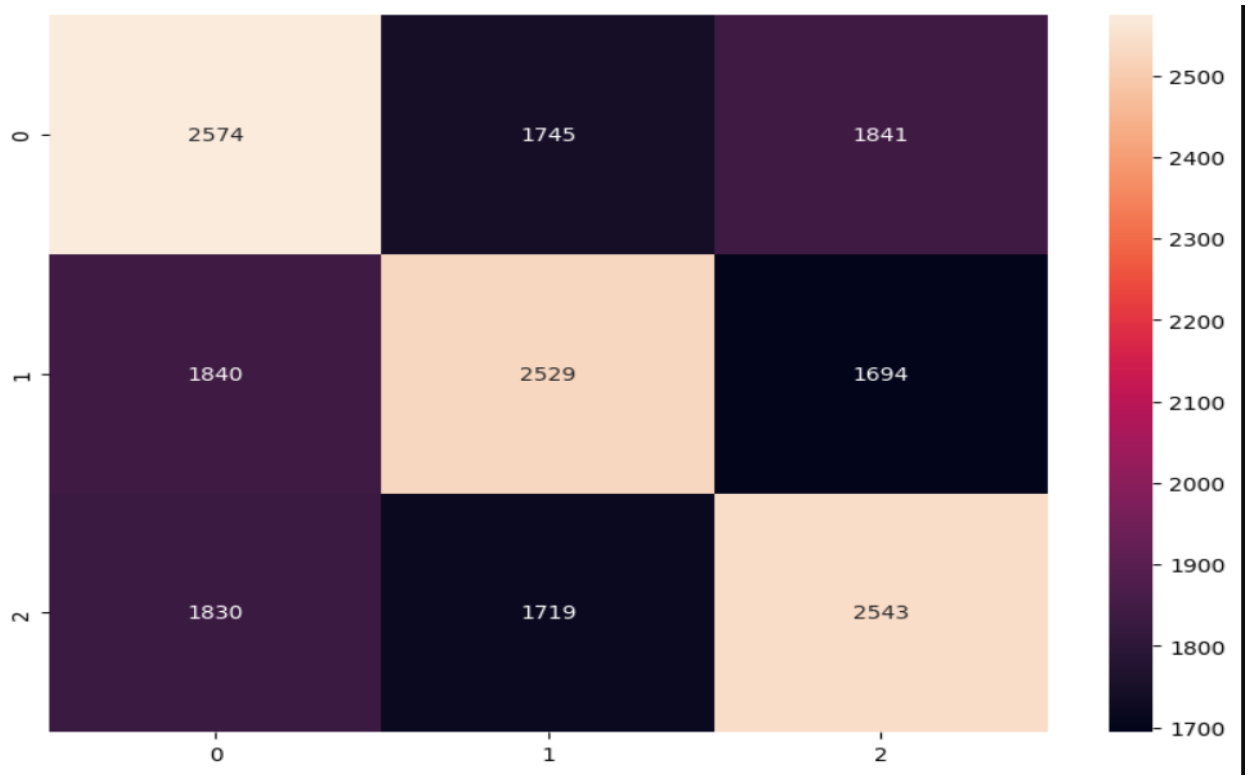
Hyperparameter Tuning and Performance Metrics :



	Precision	Recall	F1-score	Support
0	0.42	0.44	0.43	6160
1	0.43	0.42	0.42	6063
2	0.44	0.42	0.43	6092
accuracy				0.43

Test#2:- Input features : 'Medical Condition', 'Length of Stay' ,'Admission Type' ,'Admission Day' ,'Admission Month'

Hyperparameter Tuning and Performance Metrics :

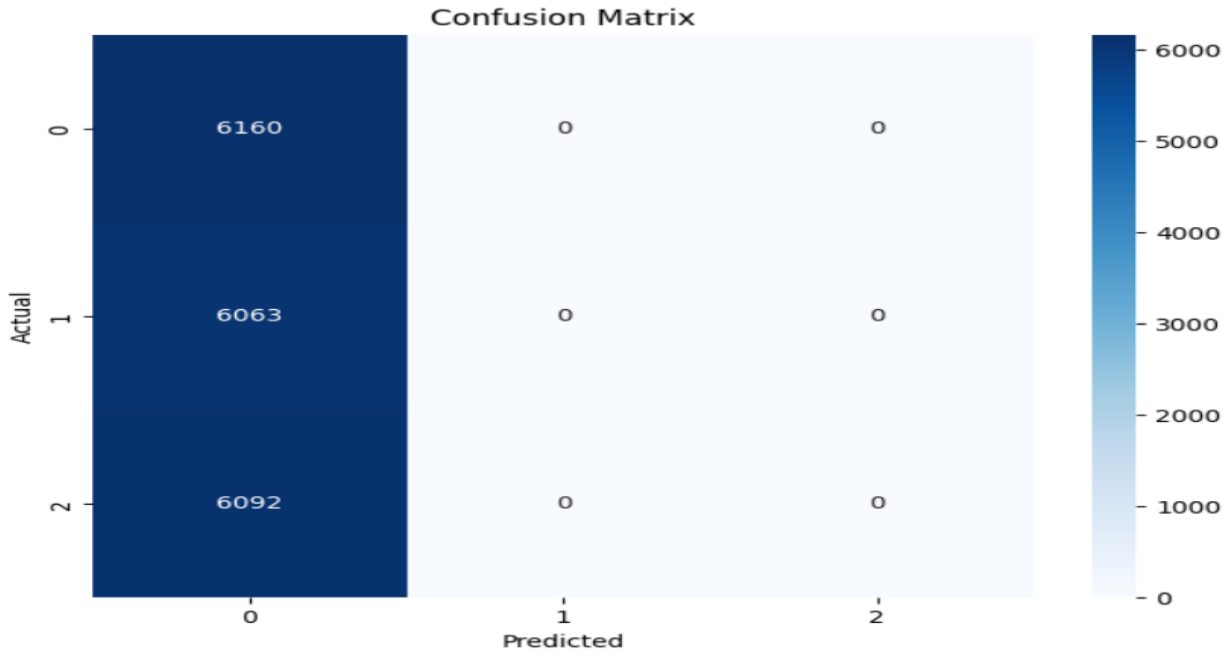


	Precision	Recall	F1-score	Support
0	0.41	0.42	0.42	6160
1	0.42	0.42	0.42	6063
2	0.42	0.42	0.42	6092
accuracy				0.42

Model name: Deep_Learning NN

Test#1:- Input features : 'Medical Condition' ,'Length of Stay', 'Admission Type' , 'Age', 'Admission Day', 'Admission Month', 'Admission Year'

Hyperparameter Tuning and Performance Metrics :

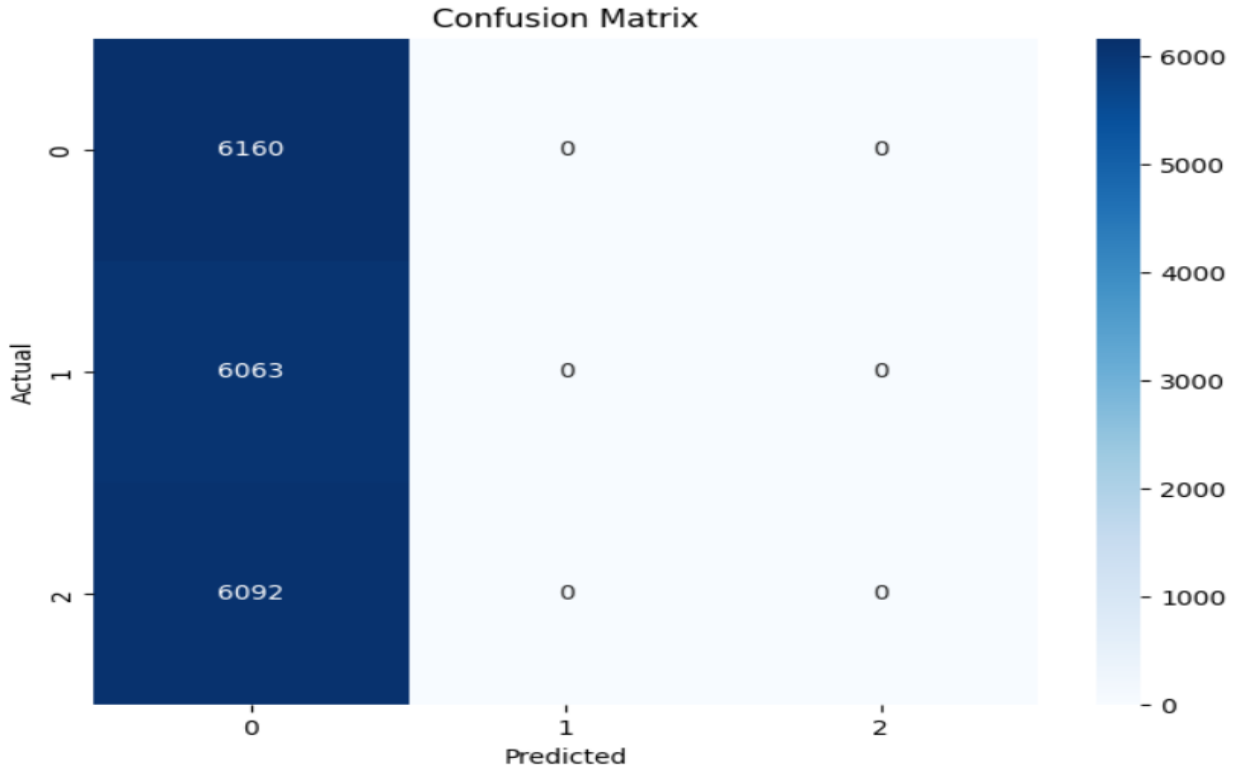


	Precision	Recall	F1-score	Support
0	0.34	1.00	0.50	6160
1	0.00	0.00	0.00	6063
2	0.00	0.00	0.00	6092

accuracy				0.34

Test#2:- Input features : 'Medical Condition', 'Length of Stay' ,'Admission Type' ,'Admission Day', 'Admission Month'

Hyperparameter Tuning and Performance Metrics :

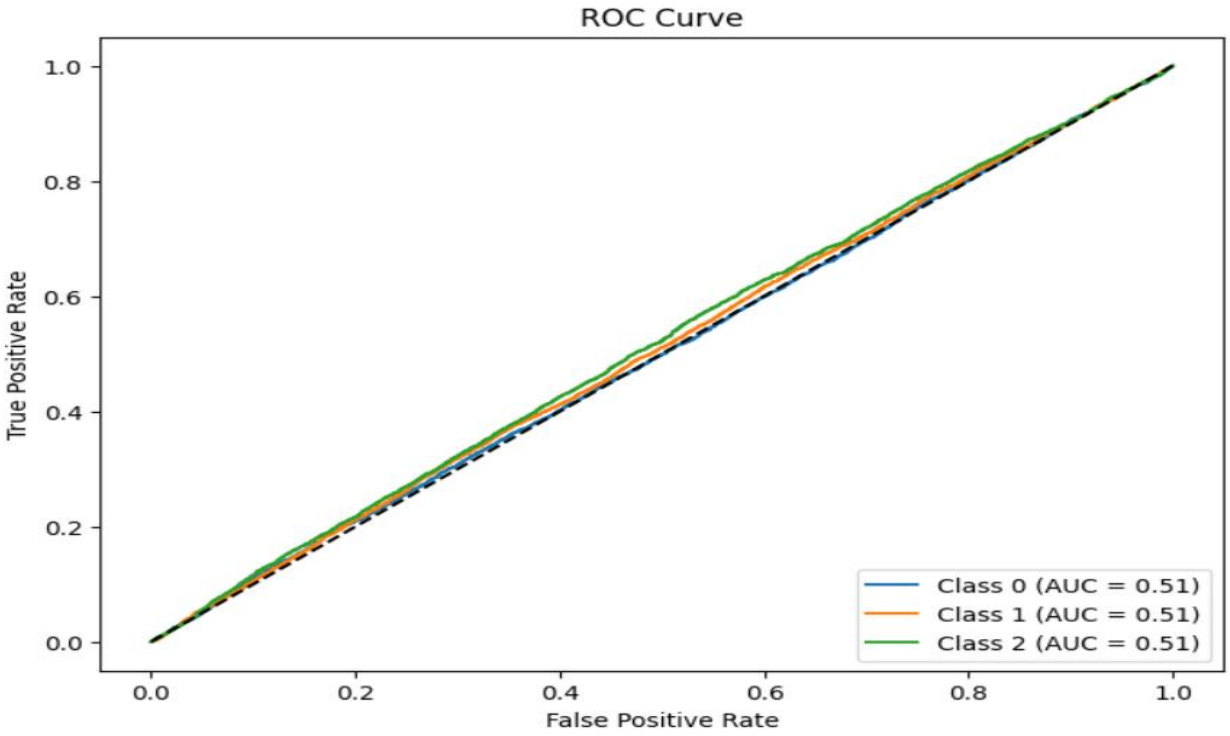


	Preision	Recall	F1-score	Support
0	0.34	1.00	0.50	6160
1	0.00	0.00	0.00	6063
2	0.00	0.00	0.00	6092
accuracy				0.34

Model name: Gradient_Boosting

Test#1:- Input features : 'Medical Condition' ,'Length of Stay', 'Admission Type', 'Age','Admission Day', 'Admission Month', 'Admission Year'

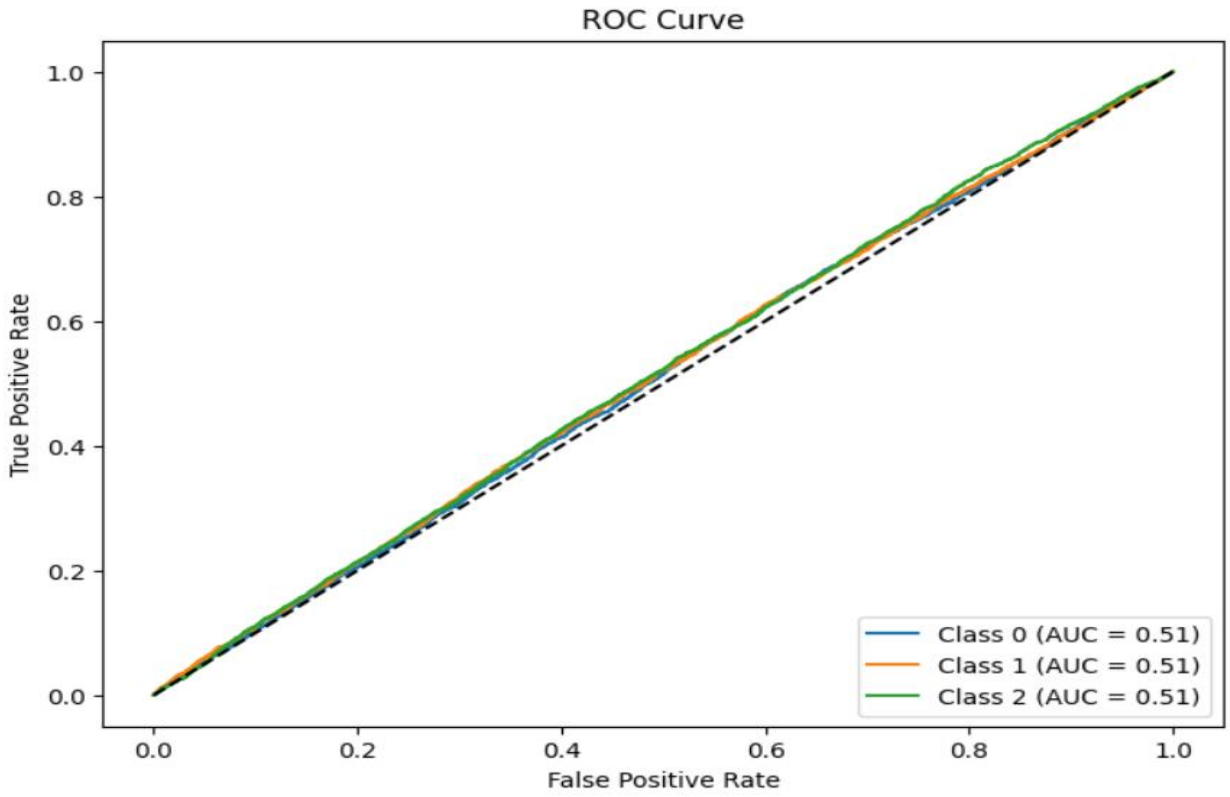
Hyperparameter Tuning and Performance Metrics :



	Precision	Recall	F1-score	Support
0	0.34	0.38	0.36	6160
1	0.34	0.29	0.31	6063
2	0.35	0.36	0.35	6092
accuracy				0.34

Test#2:- Input features : 'Medical Condition' ,'Length of Stay', 'Admission Type', 'Admission Day', 'Admission Month'

Hyperparameter Tuning and Performance Metrics :

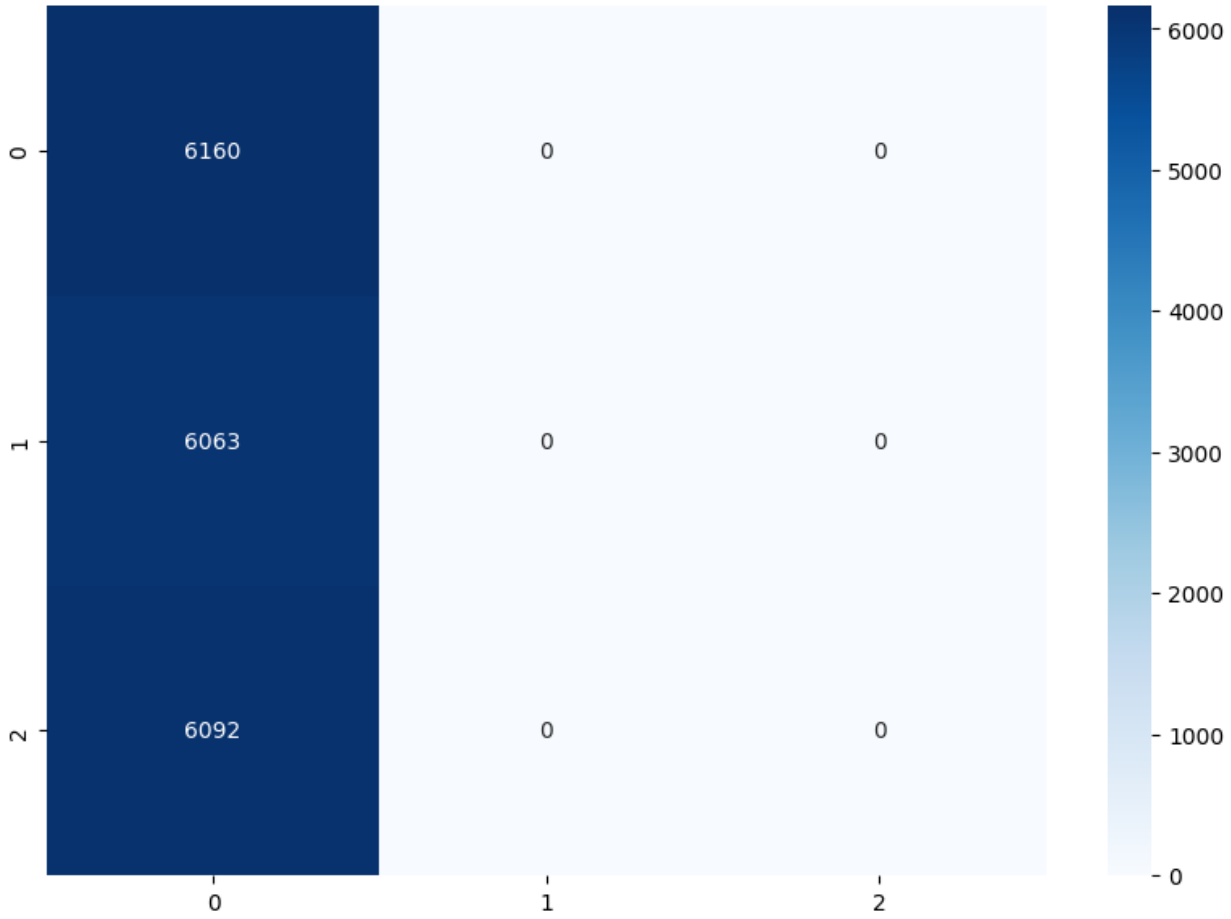


	Precision	Recall	F1-score	Support
0	0.35	0.38	0.36	6160
1	0.34	0.31	0.33	6063
2	0.35	0.35	0.35	6092
accuracy				0.35

Model name: SVM

Test#1:- Input features : 'Medical Condition' ,'Length of Stay', 'Admission Type', 'Age' ,'Admission Day', 'Admission Month', 'Admission Year'

Hyperparameter Tuning and Performance Metrics :



	Precision	Recall	F1-score	Support
0	0.34	1.00	0.50	6160
1	0.00	0.00	0.00	6063
2	0.00	0.00	0.00	6092
accuracy				0.34

-PCA:

I used PCA to improve the data and obtain better results, but we did not get better results.

If you need further assistance or want to explore other methods, feel free to ask.

In the end, the best results we obtained were with the Random Forest model, but the results are still unsatisfactory.