

# Predicting ICU Admission of Confirmed COVID19 Cases

---



INDRAPRASTHA INSTITUTE *of*  
INFORMATION TECHNOLOGY  
DELHI

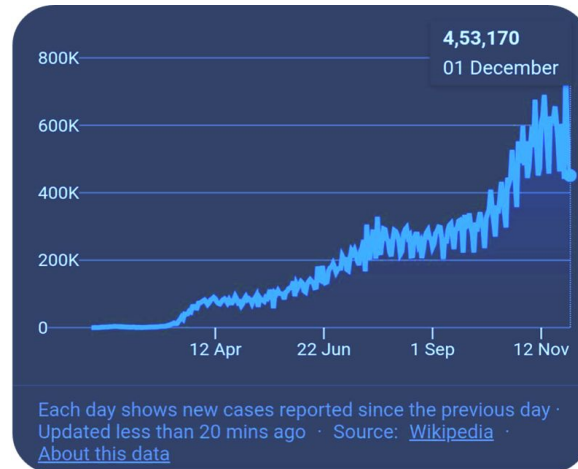
**Group- 44**



# Motivation



- ❖ COVID-19, The biggest challenge that world is currently facing.
- ❖ A large population is affected by the coronavirus.
- ❖ Healthcare Systems are not ready to handle this load.



# Motivation



- ❖ Limited Number of Advanced Medical Resources like ICUs.
- ❖ Uneven requirement and load on ICUs across hospitals.
- ❖ Utilization of clinical data can be improved to a great extent.
- ❖ Casualties due to time wastage in patient transfers.



What can help us to overcome these challenges ?

- ❖ A strategy that can assess and analyse the clinical data of confirmed COVID19 patients.
- ❖ A model that can accurately predict the need of ICU on the basis of clinical data.
- ❖ Predicting as soon as the patient is admitted to hospital. The earlier the better.
- ❖ Planning flow of operations on the basis of prediction.



## Research Paper 1

Link - <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7356638/>

### Objective

Approximately 20–30% of patients with COVID-19 require hospitalization, and 5–12% may require critical care in an intensive care unit (ICU). A rapid surge in cases of severe COVID-19 will lead to a corresponding surge in demand for ICU care.

They developed a machine learning-based risk prioritization tool that predicts ICU transfer within 24 h, seeking to facilitate efficient use of care providers' efforts and help hospitals plan their flow of operations.

## Methods

Study Cohort and Features, Sampling Strategy, Labeling, Training, Testing, and Cross-Validation, Feature Selection, Model Testing.

## Results

The median time to ICU transfer from the time of admission was 2.45 days. The study group included a higher proportion of women, and about two-thirds of the group was between 18 and 65 years old. The median duration of hospital stay was 4.2 days and ranged between 1 to 43.6 days. About one-quarter of the patients in the group had more than one comorbidity, including COPD, diabetes, hypertension, obesity, or cancer.


## Research Paper 2

Link - <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0236618>

### Objective

This study aimed to develop risk scores based on clinical characteristics at present to predict intensive care unit (ICU) admission and mortality in COVID-19 patients. Five significant variables lactate dehydrogenase, procalcitonin, pulse oxygen saturation, smoking history, and lymphocyte count were used for predicting ICU admission and seven significant variables predicting mortality were heart failure, procalcitonin, lactate dehydrogenase, chronic obstructive pulmonary disease, pulse oxygen saturation, heart rate, and age.

This risk score system may prove useful for frontline physicians in clinical decision-making under time-sensitive and resource-constrained environment.



## Methods

Study population and data collection, Statistical analysis and modeling, Patient selection, Characteristics of the ICU admission group, Characteristics of mortality group

## Results

The sensitivity and specificity of the risk scores were 10.5% and 99.2% for predicting ICU admission, and 7.1% and 100% to predict mortality. The high specificity and low sensitivity were due to the imbalance of sample sizes where there were more patients in general admission group compared to ICU admission or death group.



# Dataset Description

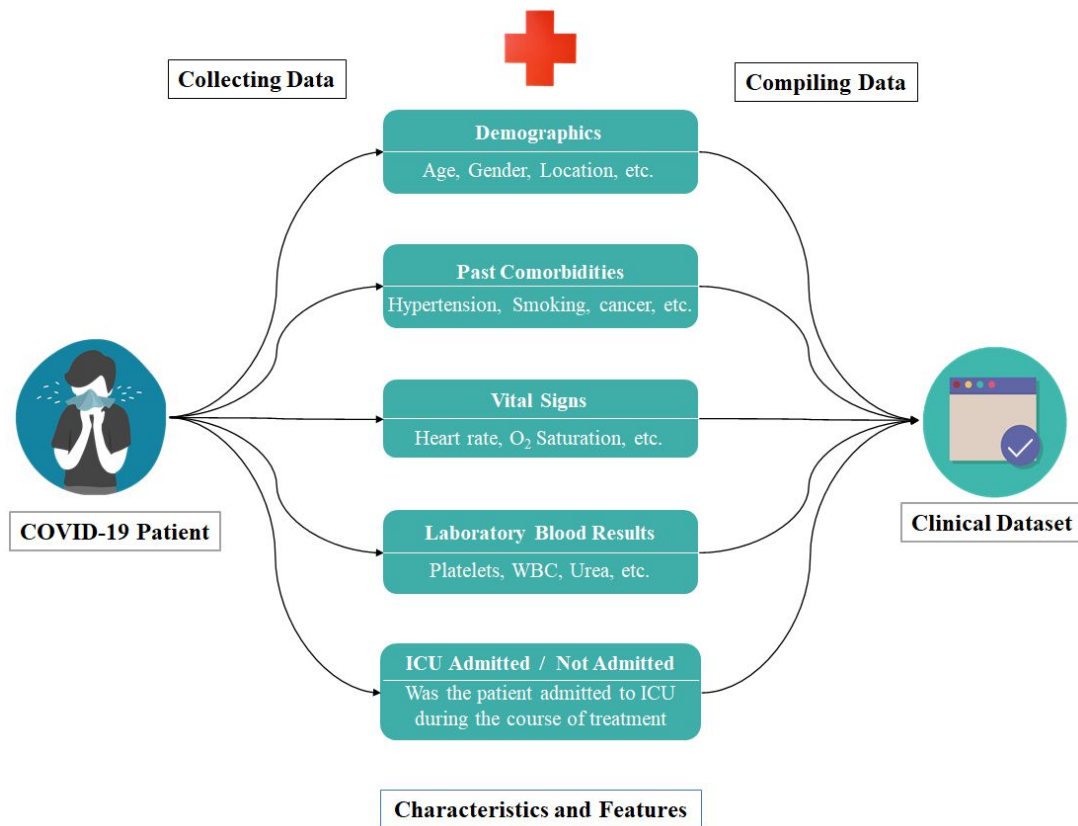


**The dataset contains anonymized data from Hospital Sírio-Libanês, São Paulo and Brasília.**

<https://www.kaggle.com/S%C3%ADrio-Libanes/covid19/download>

- ❖ Data contains clinical details of 384 patients in time-window format. There are 5 windows for each patient i.e. 0-2, 2-4, 4-6, 6-12, 12+ hours according to the admission in hospital.
- ❖ Data is cleaned and scaled by column according to Min Max Scaler to fit between -1 and 1.
- ❖ There are 54 features in the dataset. Although, they are further expanded by calculating mean, median, max, min, diff and relative diff. So, getting a total of
- ❖ **There are 4 categories of attributes:**
  - Patient demographic information (03)
    - Age, Gender.
  - Patient previous grouped diseases (09)
    - Hypertension, Immunosuppressed, 6 other confidential groupings.
  - Blood results (36)
    - Albumin, Calcium, Glucose, Hemoglobin, Sodium, Urea, Platelets, etc.
  - Vital signs (06)
    - Blood Pressure, Heart rate, Respiratory rate, Temperature, Oxygen saturation.

# Dataset Description



## Preprocessing

1. Scaling was not required as the data is already scaled by column according to Min Max Scaler to fit between -1 and 1.
2. As there is a concept of windows, some modification was required. We could not use the windows in which patient is admitted to ICU as we need to predict for future not for present. So, we removed those rows.
3. We can use all the windows before the ICU admission, but in the project we are using only the first window i.e. 0-2 because the earlier we predict the more clinically relevant our prediction is. So, we removed all the windows except 0-2.
4. There was some missing data. We filled it using the mean of values of all windows of the patient.
5. Performed Binary Hotcoding for some of the columns like window, age percentile, etc.

## Reading, Modification and Preprocessing data

- ❖ Data was in .csv format so we read it using csv library in python.
- ❖ One hot encoding to convert not float columns. (Age Percentile and Window)
- ❖ Marking label of 0-2 window as 1 if the patient was admitted to ICU in any of the future windows.
- ❖ Removing all the records of the windows in which patients were actually admitted to the ICU (windows with ICU label 1 before the previous step).
- ❖ Filling the NaN values of window 0-2 with the help of mean of values in all the windows of that patient.
- ❖ Removing all the rows still having NaN values.
- ❖ At the beginning we started with data of 384 patients but after the data cleaning, modification and preprocessing the final size reduced to 293.

## Analyzing Data

- ❖ Data was analysed on these categories of attributes: Demographics, Previous Grouped Diseases, Blood Test Results, Vital Signs.
- ❖ We used matplotlib for plotting and visualization
- ❖ For the purpose of comparison three different views of dataset were formed: Total, ICU Admitted, Non ICU patients.
- ❖ Plots like histogram, pie charts, dotplots, etc. were used.



## Feature Engineering & Training

- ❖ Based on analysis, we found features that were impacting the target label most.
- ❖ Created a correlation matrix to find the correlation among the features and remove the redundant features. We got 93 features after this step.
- ❖ Out of the reduced features, we selected the top 43 features based on the literature.



## Training & Testing

- ❖ We used sklearn module to train and test various models.
- ❖ Applied grid search technique on each model to find the optimal hyperparameters for it.
- ❖ Plotted learning curves for the visualization that includes train-test error vs parameter, accuracy vs parameter, etc.
- ❖ Compared the models on the basis of accuracy, precision, recall and AUC-ROC curve.



## Demographics

World has witnessed that mortality rate is very high in case of older generation. Covid19 cases become more critical when the patient is from the older.

We analysed the demographics of the data and found some really interesting facts.



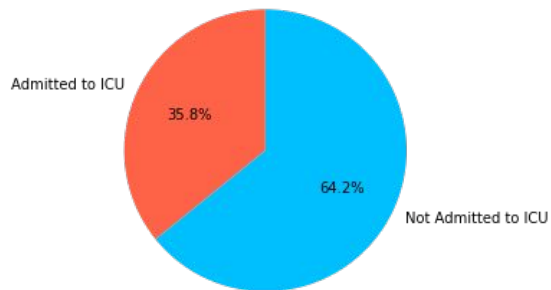


# Analysis

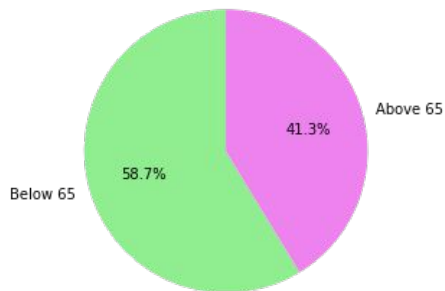


## Demographics

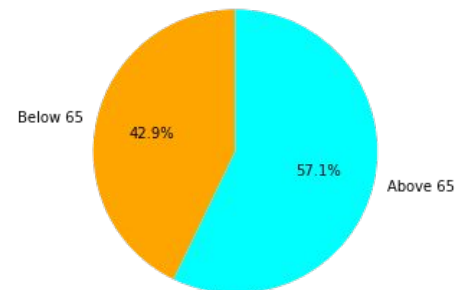
ICU Distribution of data



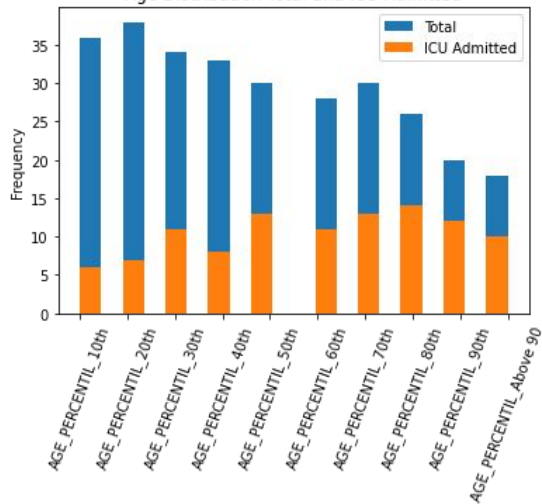
Age Distribution of data



Age Distribution of ICU Admitted patients



Age Distribution Total and ICU Admitted



## Past Comorbidities

Literature has stated that the past health conditions among patients like diseases, hypertension, being immunosuppressed have a correlation with the severity of the covid19. Also, some studies suggest that the habits like smoking, drinking, etc. also impact the criticality of the covid19.

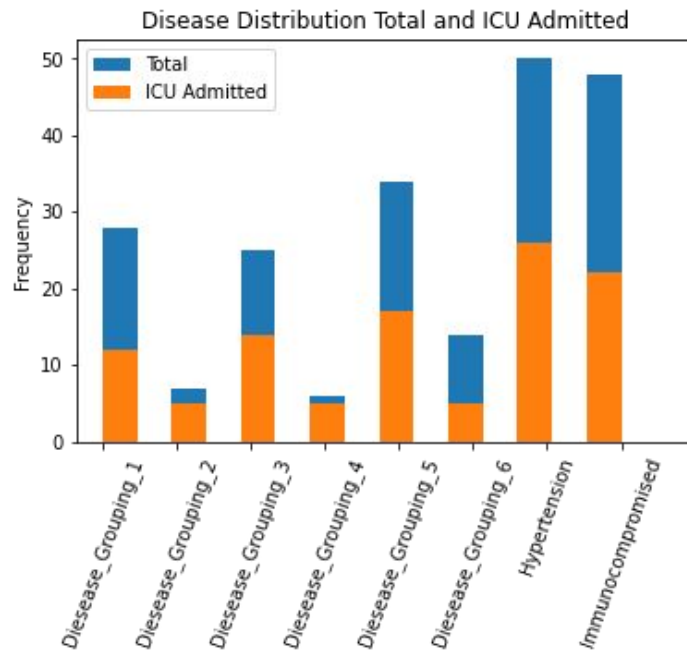
We analysed the past comorbidities among the patients and compared the statistics of Non ICU patients with ICU admitted patients.



# Analysis



## Past Comorbidities



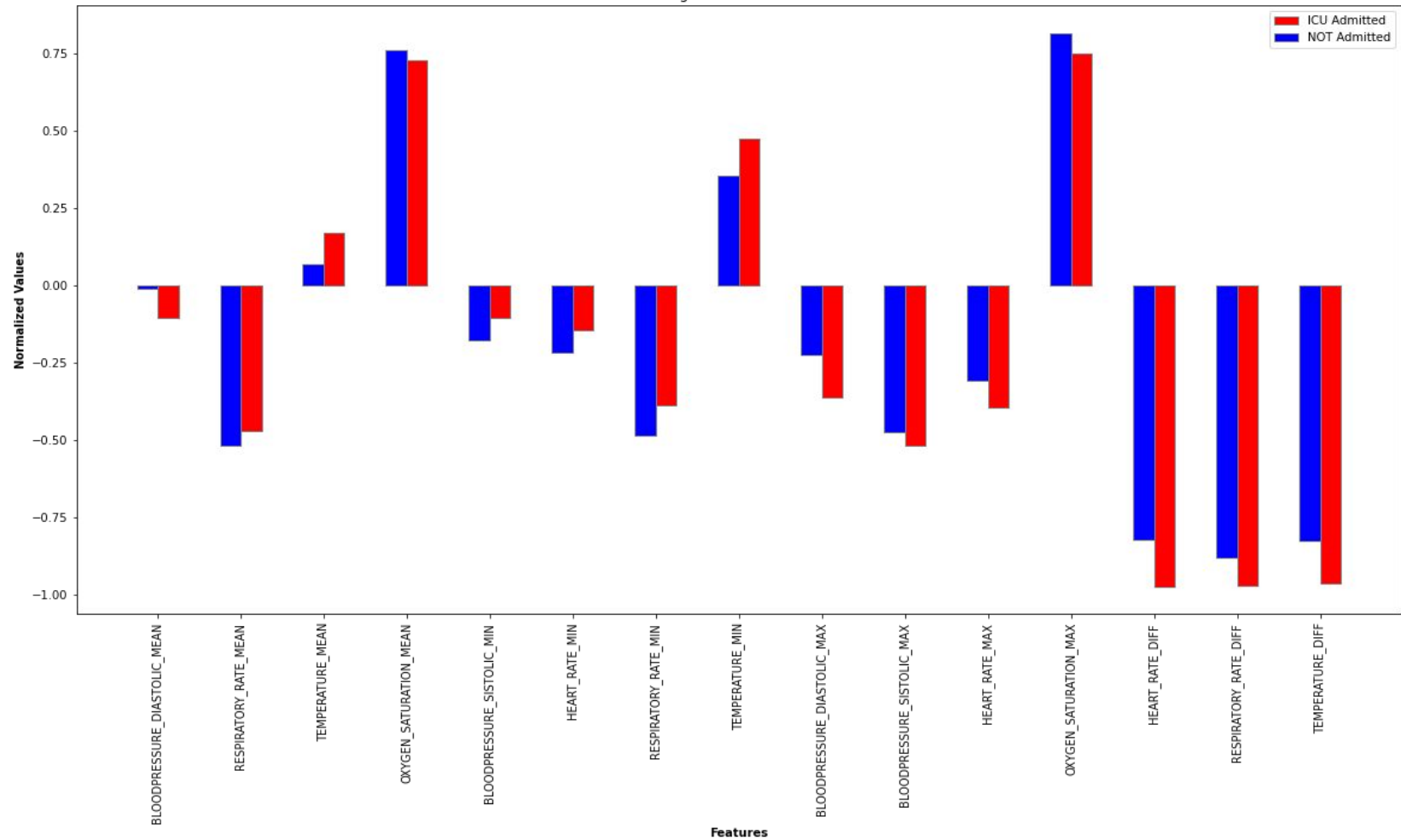
## Vital Signs

These are the vital clinical characteristics of a patient that can be monitored live. These signs are generally used to know about the present health condition of the patient.

We analysed the vital sign attributes in both ICU admitted & Not admitted patients and found that the blood pressure, oxygen saturation and temperature are significant indicators.



Vital Signs of Covid19 Patients



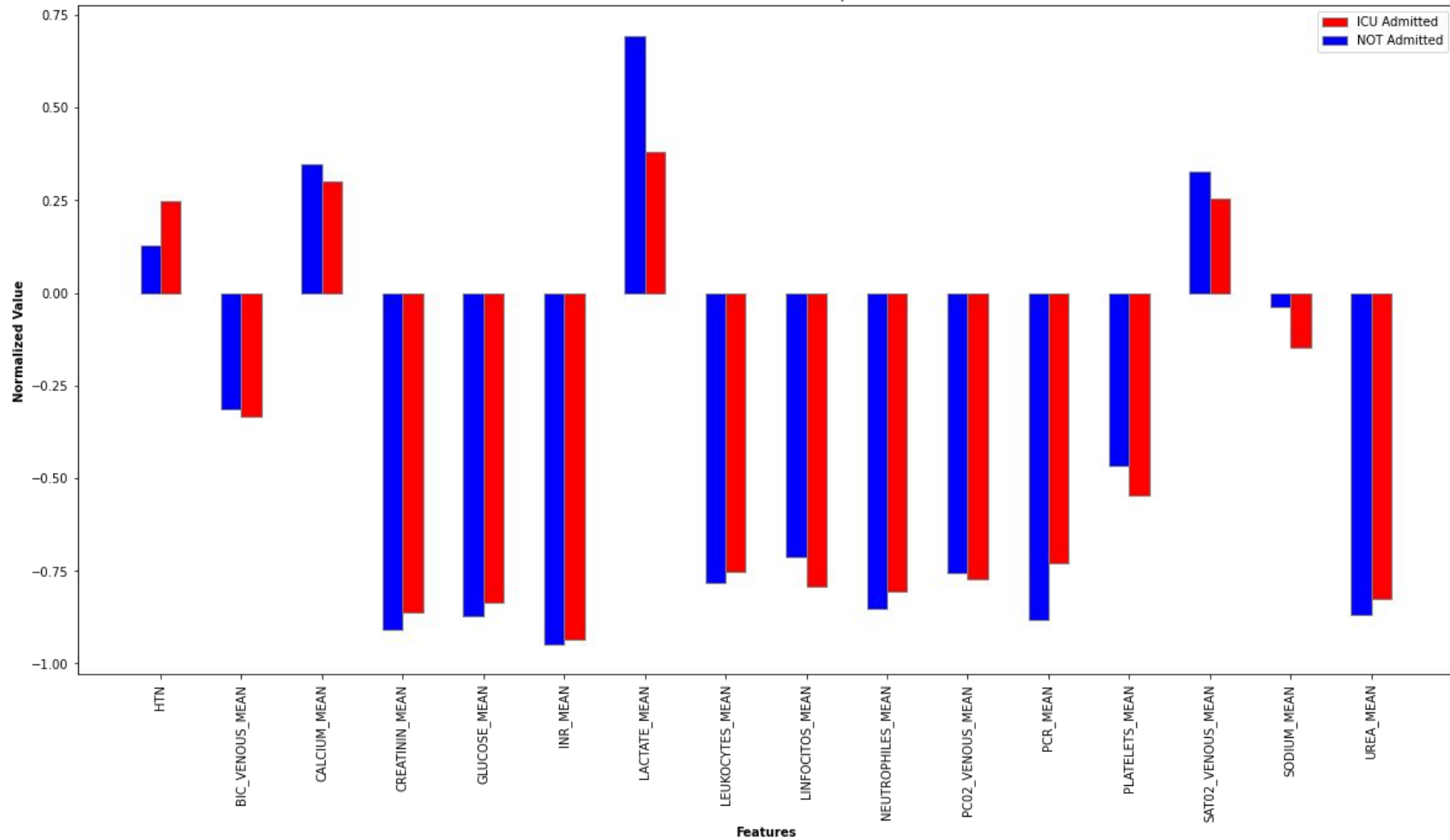
## Laboratory Results

Our blood is composed of various kind of molecules that is utilised by different part of our bodies.

When there is something wrong in the body like presence of the virus. Our body behaves in a different way to counter it, this results in change of requirement, production and consumptions of these molecules. So, it is expected to see difference in the levels in ICU and Non-ICU patients.

We analysed the laboratory results in both ICU admitted and Not Admitted patients and found that sodium, lactate, platelets, Hemoglobin are some of the major indicators where we saw a significant difference. We couldn't include all the attributes for the visualization as the number is large.

Lab Test Results of Covid19 patients



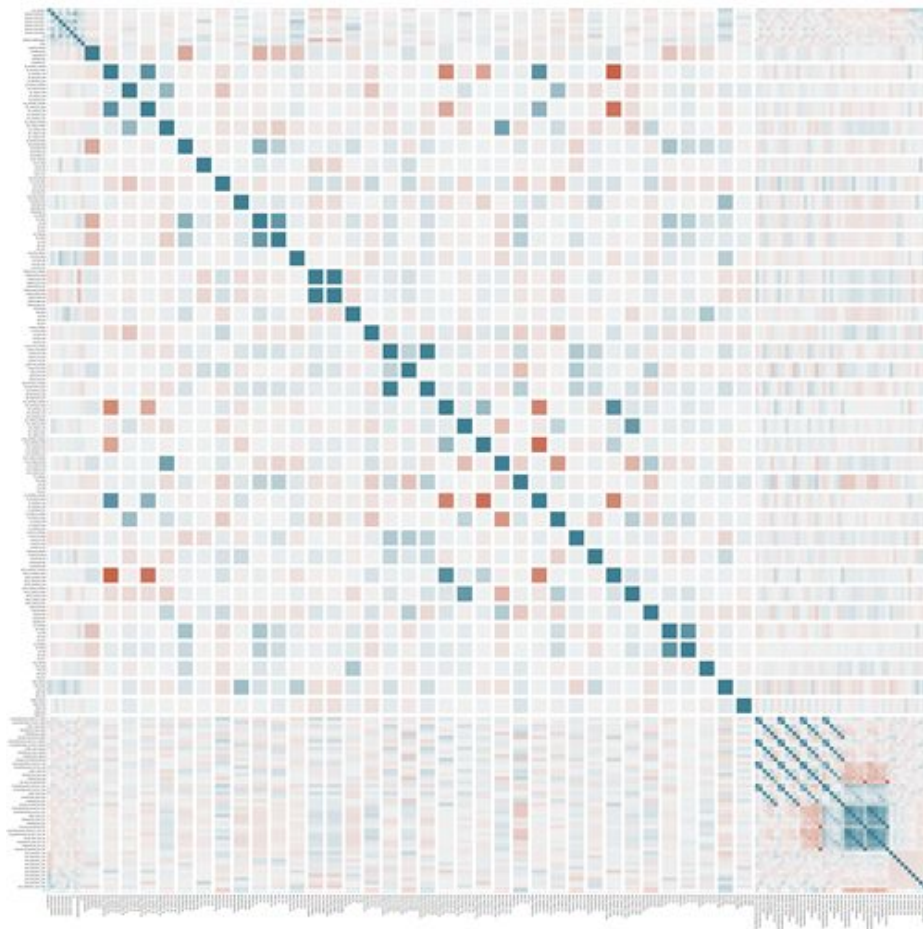
## Correlation Matrix

We analyzed the correlation in features and removed the highly correlated features. This helped us in reducing the dimensions of the data and made further analysis simpler. We visualized the correlation matrix in the form of heatmaps.

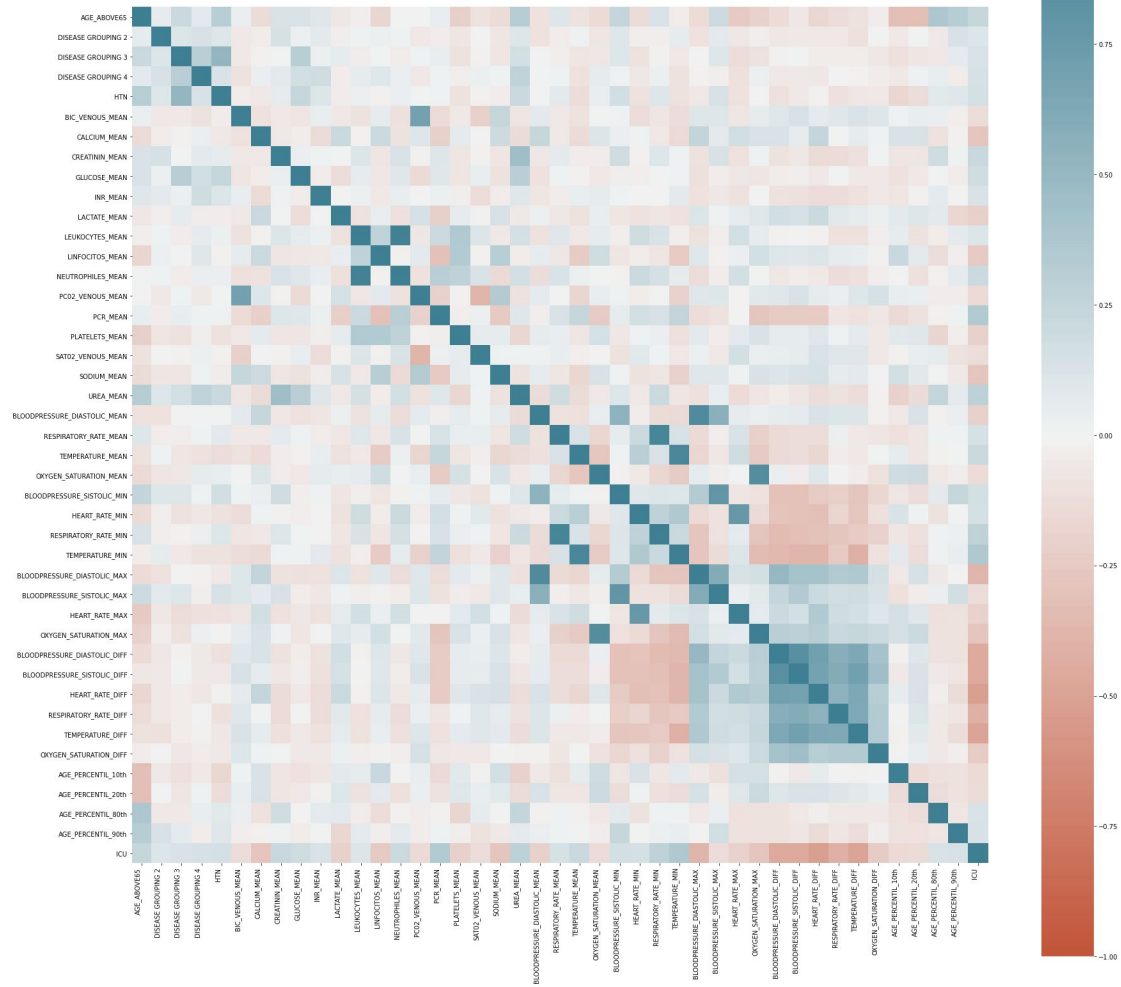




**Before**



After



## Final Selected Features

After all the analysis and feature engineering we selected these features.

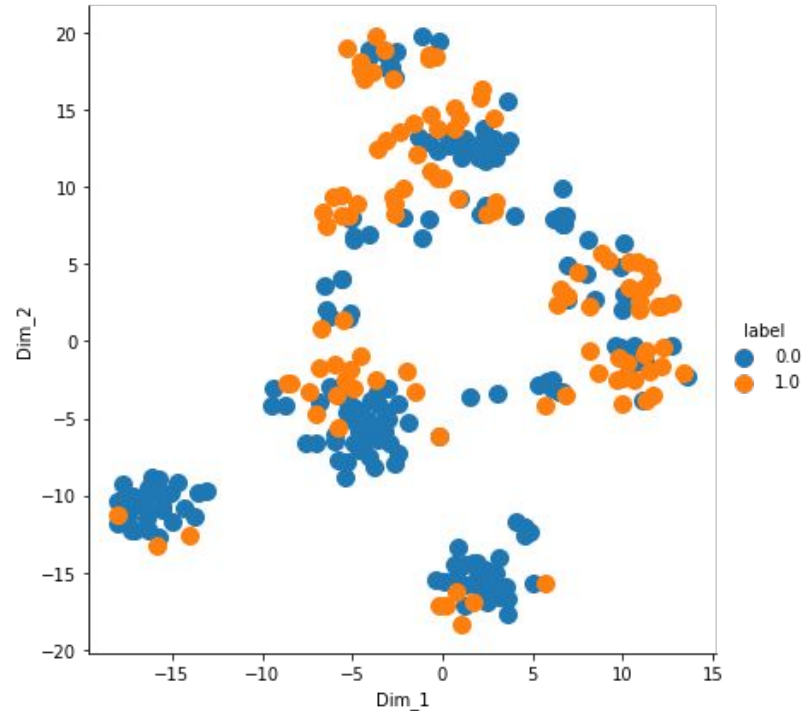
```
['AGE_ABOVE65', 'DISEASE GROUPING 2', 'DISEASE GROUPING 3', 'DISEASE GROUPING 4', 'HTN', 'BIC_VENOUS_MEAN', 'CALCIUM_MEAN',  
'CREATININ_MEAN', 'GLUCOSE_MEAN', 'INR_MEAN', 'LACTATE_MEAN', 'LEUKOCYTES_MEAN', 'LYMPHOCYTES_MEAN', 'NEUTROPHILES_MEAN',  
'PCO2_VENOUS_MEAN', 'PCR_MEAN', 'PLATELETS_MEAN', 'SATO2_VENOUS_MEAN', 'SODIUM_MEAN', 'UREA_MEAN', 'BLOODPRESSURE_DIASTOLIC_MEAN',  
'RESPIRATORY_RATE_MEAN', 'TEMPERATURE_MEAN', 'OXYGEN_SATURATION_MEAN', 'BLOODPRESSURE_SISTOLIC_MIN', 'HEART_RATE_MIN',  
'RESPIRATORY_RATE_MIN', 'TEMPERATURE_MIN', 'BLOODPRESSURE_DIASTOLIC_MAX', 'BLOODPRESSURE_SISTOLIC_MAX', 'HEART_RATE_MAX',  
'OXYGEN_SATURATION_MAX', 'BLOODPRESSURE_DIASTOLIC_DIFF', 'BLOODPRESSURE_SISTOLIC_DIFF', 'HEART_RATE_DIFF', 'RESPIRATORY_RATE_DIFF',  
'TEMPERATURE_DIFF', 'OXYGEN_SATURATION_DIFF', 'AGE_PERCENTIL_10th', 'AGE_PERCENTIL_20th', 'AGE_PERCENTIL_80th', 'AGE_PERCENTIL_90th']
```

# Analysis



## t-SNE Visualization

We used t-SNE to visualize the final data in low dimensional space.

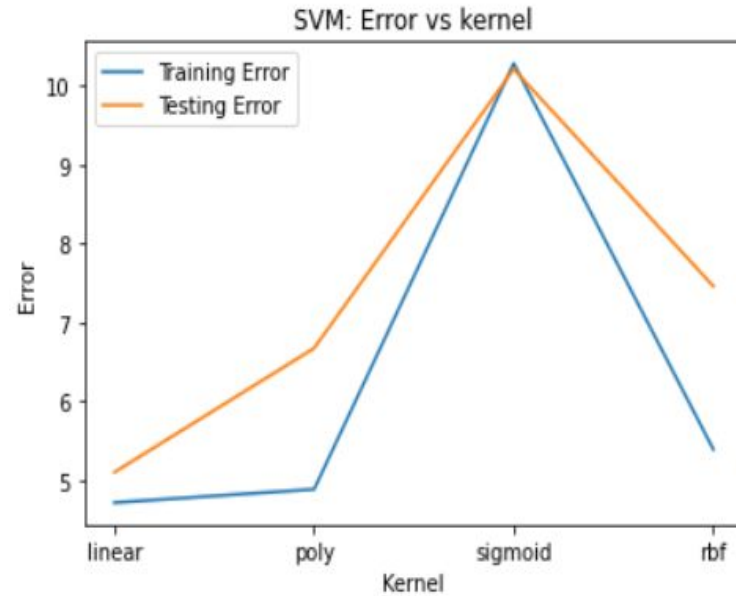


# Results



## Support Vector Machine

Optimal Parameters	Accuracy	Sensitivity	Specificity	ROC/AUC Score
kernel='linear'	85.22	66.67	93.44	0.80

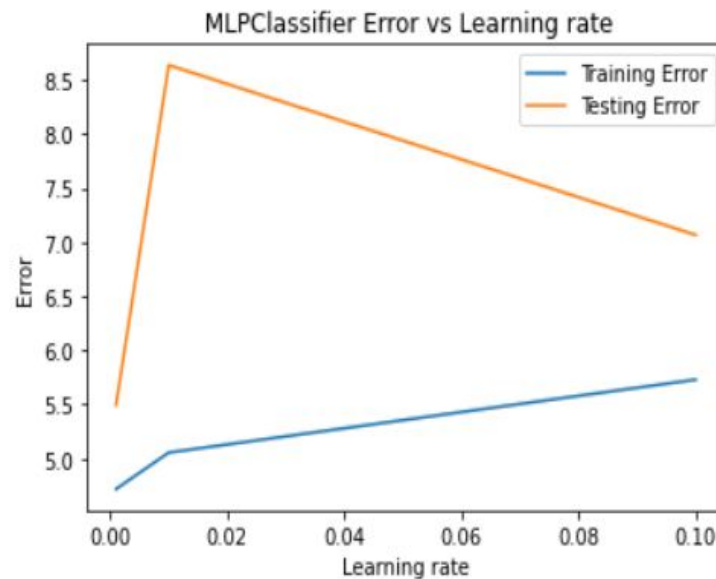


# Results



## Multi-Layer Perceptron

Optimal Parameters	Accuracy	Sensitivity	Specificity	ROC/AUC Score
Activation functions = ['identity', 'Logistic', 'Tanh', 'relu'] max_iter = 10000 batch_size=64 Learning rate=0.1	[85.22, 80.68, 84.09, 82.95]	[66.67, 62.96, 62.96, 66.67]	[93.44, 88.52, 93.44, 90.16]	[80.05, 75.74, 78.20, 78.41]

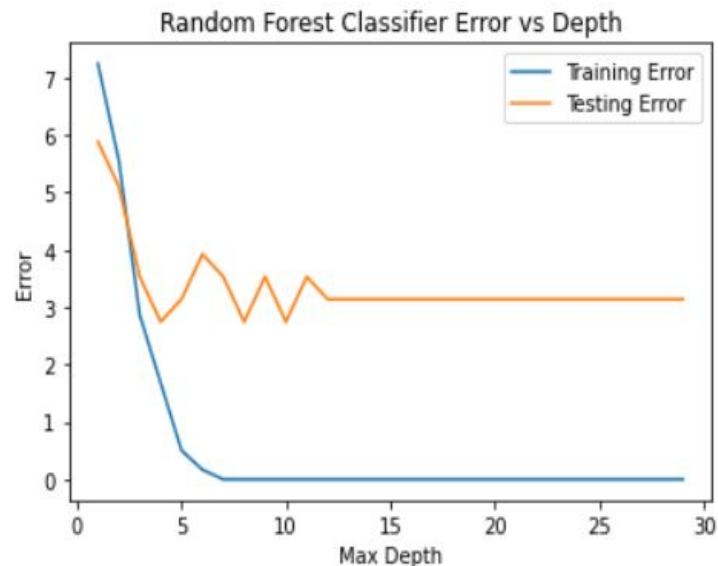


# Results



## Random Forest

Optimal Parameters	Accuracy	Sensitivity	Specificity	ROC/AUC Score
Criterion = 'gini' Random_state = 23 max_depth=6 bootstrap=True	93.18	88.89	95.08	0.919



# Results



## K-Neighbors

Optimal Parameters	Accuracy	Sensitivity	Specificity	ROC/AUC Score
N_neighbours = 25 p=1	82.95	51.85	96.72	0.742



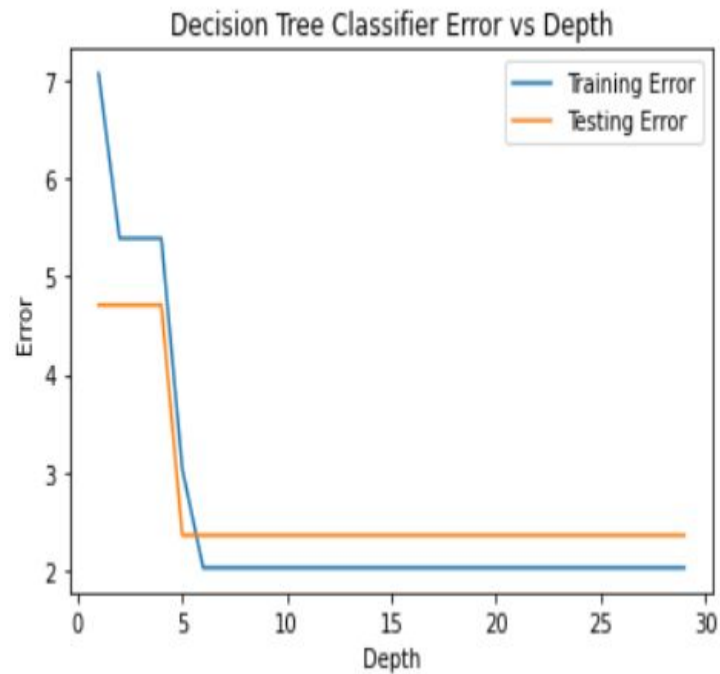


# Results



## Decision Tree (Best Performance)

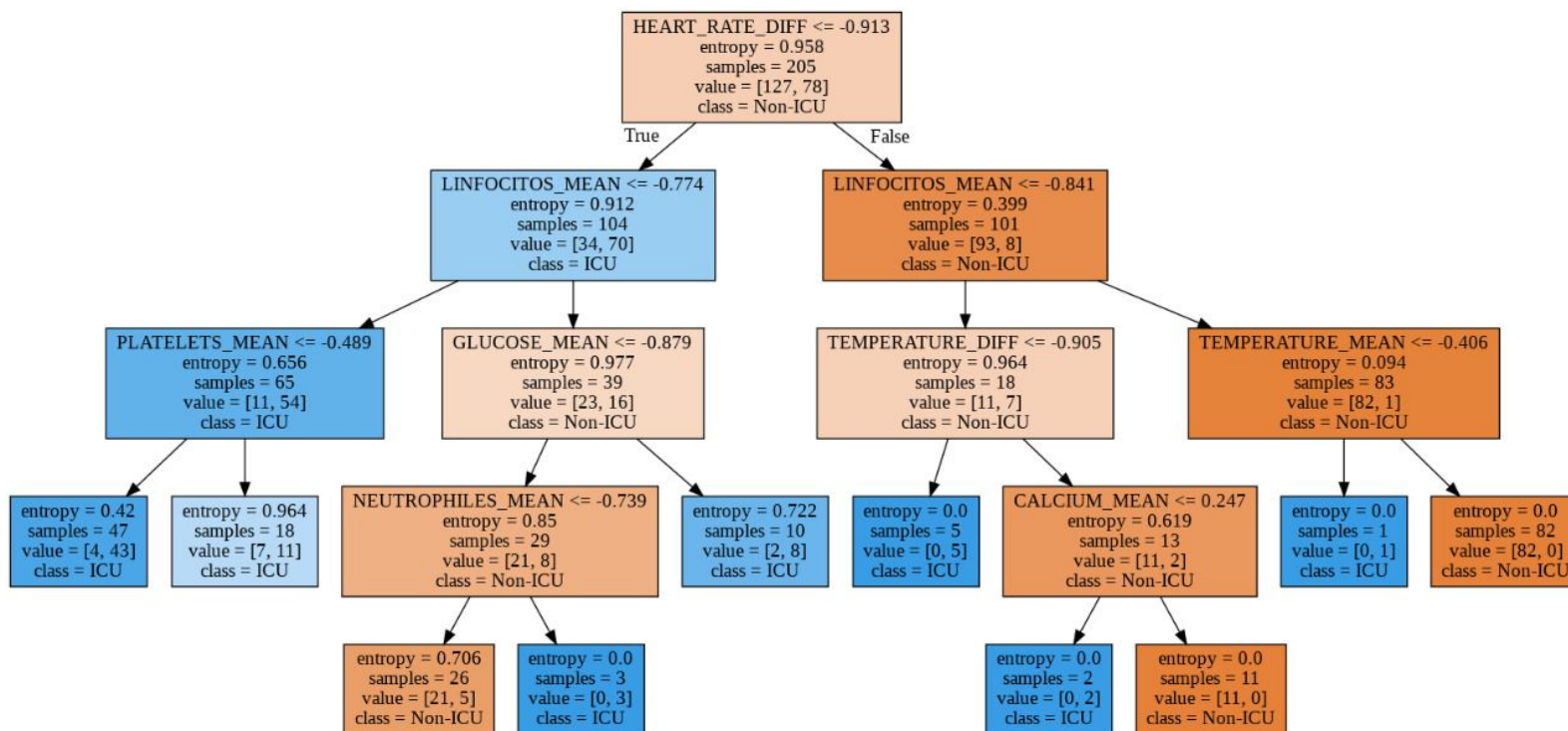
Optimal Parameters	Accuracy	Sensitivity	Specificity	ROC/AUC Score
Criterion = 'entropy' max_depth=4 max_leaf_nodes=10	94.31	92.59	95.08	0.938



# Results



## Decision Tree (Best Performance)



<b>Classifier Algorithm</b>	<b>Optimal parameters</b>	<b>Accuracy</b>	<b>Sensitivity</b>	<b>Specificity</b>	<b>ROC AUC Score</b>
<b>Logistic Regression</b>	random_state=1 cv=1 max_iter=5000	84.09	66.67	91.80	0.79
<b>Gaussian Naive Bias</b>	Default	82.95	66.66	90.16	0.784
<b>SGD</b>	random_state=285	88.63	74.07	95.08	0.845
<b>SVM</b>	kernel='linear'	85.22	66.67	93.44	0.80
<b>Decision Tree</b>	criterion='entropy' max_depth=3 max_leaf_nodes=10	94.31	92.59	95.08	0.938
<b>Random Forest</b>	criterion='gini' random_state=23 max_depth=6 bootstrap=True	93.18	88.89	95.08	0.919
<b>K-Nearest Neighbour</b>	N_neighbours = 25 p=1	82.95	51.85	96.72	0.742
<b>Multi-Layer Perceptron</b>	Activation functions=['identity', logistic', 'tanh', 'relu'] max_iter = 10000 batch_size=64 Learning rate=0.1	[85.22, 80.68, 84.09, 82.95]	[66.67, 62.96, 62.96, 66.67]	[93.44, 88.52, 93.44, 90.16]	[80.05, 75.74, 78.20, 78.41]

# Team Member's Contribution

---



## ❖ **Hemant Dhankar**

- Preprocessing, Feature Engineering, Model selection, Hyperparameter Tuning, Result Analysis, Report Writing.

## ❖ **Nitesh Jaiswal**

- Literature Review, Visualization, Training Models, Hyperparameter Tuning, Result Analysis, Report Writing.

## ❖ **Bhaskar Singh**

- Preprocessing, Visualization, Training Models, Hyperparameter Tuning, Result Analysis, Report Writing.

These were the assigned responsibilities. However, all the members equally contributed to all the work done.

Thank You