PATIENT SURVIVAL PREDICTION

CAPSTONE PROJECT – FINAL REPORT

In partial completion of

POST GRADUATE PROGRAM IN

DATA SCIENCE & ENGINEERING

Submitted By

Bhupendra Mewada

Madhavkumar

Tumu Adithya

Ajinkya Dandgavhal

Capstone Group 9 - PGPDSE-FT Online Aug21-B

Project Mentor

Ms. Vibha Santhanam

GREAT LEARNING

greatlearning

Learning for Life

www.mygreatlearning.com

CERTIFICATE OF COMPLETION

I hereby certify that the project titled ‘Patient Survival Prediction’ was undertaken and completed under my supervision by Bhupendra Mewada, Madhavkumar, Tumu Adithya and Ajinkya Dandgavhal, students of the PG Program in Data Science & Engineering (PDPDSE-FT Online Aug21-B) at Great Learning.

Date: May 8, 2022

Ms. Vibha Santhanam
Mentor

ACKNOWLEDGEMENTS

It is a great pleasure for us to present this Capstone Project Final Report of the 9-Month PG Program in Data Science & Engineering at Great Learning. This project provided us with the best opportunity to put our knowledge into practical use.

We take this opportunity to thank our project mentor Ms. Vibha Santhanam for guiding us throughout the project. She helped us narrow down on the choice of the Project as well as the scope and focus area of the project. She gave us valuable feedback at every stage to enhance the process and the outputs. We are very grateful to Great Learning for assigning us such knowledgeable mentor.

We would like to express our gratitude towards all faculties and mentors for equipping us with the necessary knowledge and understanding of concepts which helped us to successfully undertake and complete this project.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

Date: May 8, 2022

Group Members:
Bhupendra Mewada
Madhavkumar
Tumu Adithya
Ajinkya Dandgavhal

TABLE OF CONTENTS

Sr No.	Topic	Page No
1	Abstract	5
2	Approach	6
3	Topic Survey & Assessment	8
4	Data Understanding	9
5	Pre-Processing & Statistical Analysis	14
6	Data Exploration (EDA)	16
7	Basic Model Building Steps	22
8	Using SMOTE to Handle Imbalance	24
9	Using Under & Over Sampling Combination	26
10	Using Balanced Class Weights Approach	28
11	Final Model	30
12	Results & Discussion	32
13	Implications & Scope	33
14	Closing Reflections	34
15	References	35

ABSTRACT

Introduction

Survival prediction after hospitalization and first diagnosis is important for both the doctor administering the treatment and patients or their family members. First, as the survival ability of the patients largely depends on their medical condition, fatality of the disease and other health markers, accurately forecasting the prognosis would be extremely helpful for estimating the degree of severity and planning the possible medical treatment. On the other hand, patients and their families can get adequate time to make the necessary arrangements on the basis of accurate survival prediction. As a result, timely prevention and treatment can be made and worse treatment decisions, such as over-treatment or late palliative care, can be effectively avoided.

Business Problem

Through our project ‘Patient Survival Prediction’ we have attempted to develop a model which will predict whether a patient will survive or not based on certain health markers which are recorded when the patient is hospitalized. The survival ability of a patient largely depends upon his medical condition, fatality of the disease and other health markers. We have attempted to study the same using data analysis tools and machine learning techniques and prepare a model and use it to predict the patient’s survival.

Findings

We have observed that there are certain health parameters which largely affect a patient’s survival ability which include age, icu admit source, icu type, ethnicity, apache body system, whether the patient was ventilated or not, pre-icu loss days etc. Also, there are some parameters which do not significantly affect the survivability like bmi, gender etc.

Model

Our final model uses Logistic Regression Algorithm with balanced class weights and gives a recall score of 0.79 for both the classes in the target columns (survival or death) on the test data.

Implications

Using accurate predictions of the model, the hospital staff can readily understand which patients are more likely to survive and which are not. This can help them to prioritize as to which patients need more attention and make proper medical arrangements within the critical time available and plan proper medical interventions to reduce the fatality rate.

APPROACH

We have approached this problem in a step-by-step manner. We have studied the health markers which are recorded once a patient is hospitalized. We have attempted to understand how each health marker can potentially indicate the severity of the patient's condition. We have then studied how these different health markers interact with each other and their effect on the survival ability of the patient using EDA, Statistical Analysis. And further we have built a model using Classification Algorithms.

Here, we are following the CRISP-DM methodology that stands for Cross-Industry Standard Procedure for Data Mining. Following image captures the steps which we will be following.

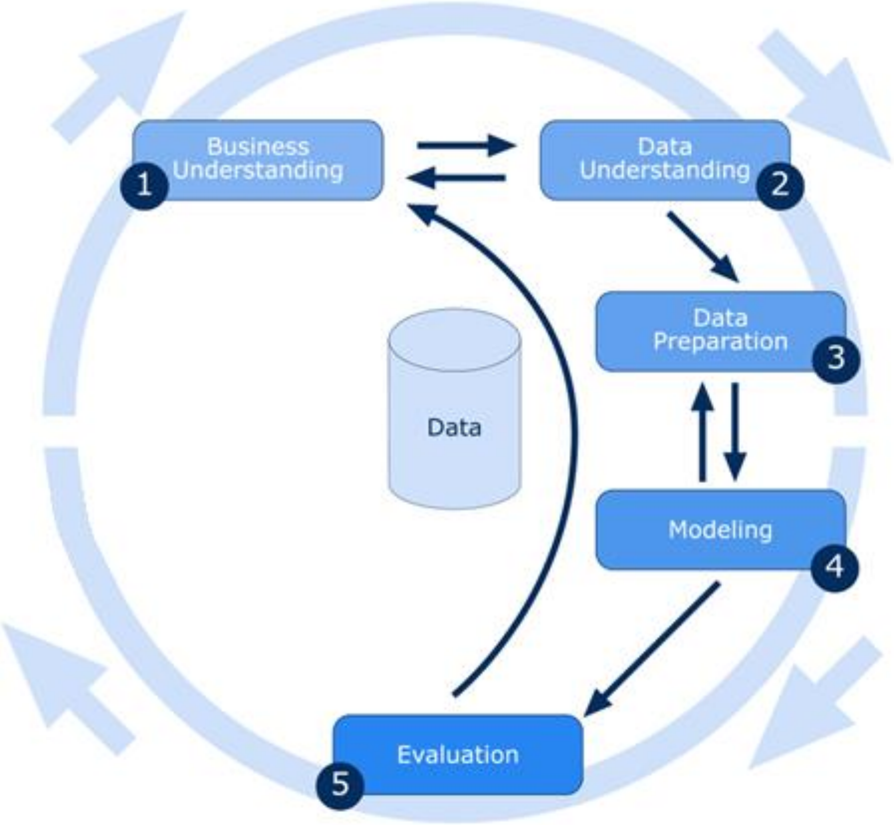


Fig : CRISP-DM (Cross-Industry Standard Procedure for Data Mining)

Tools and Techniques

- Summary Statistics for each variable
- Using graphs and box plots to visually represent the distribution of the variables
- Correlation Matrix and Correlation Heat Map
- Classification Algorithms – Logistic Regression, KNN, Decision Tree, Random Forest, AdaBoost, XGBoost
- Tools used: Python & MS-Excel
- Techniques: Box Plot, Histogram, Bar Chart, Correlation Matrix, Classification Algorithms
- We have used Jupyter Programming Environment for our analysis and data visualization

Analytical Approach

- Data extraction from secondary data source
- Data quality check
- Data cleaning and data preparation
- Study each of the variables by exploring the data
- Study the variables for its relevance for the study
- Identifying Y variable
- Performing Univariate and Bivariate analysis for all variables
- Division of data into train and test
- Model Development
- Final Model
- Model Validation & Model Validation on Test

TOPIC SURVEY & ASSESSMENT

Topic Survey in Brief

Various research articles available online were considered to understand the recent advances in patient survival prediction. Also, an attempt was made to understand the survival prediction systems in use currently. Currently, survival ability of the patient is predicted using APACHE-II scoring system (Acute Physiology and Chronic Health Evaluation II). APACHE-II is a severity-of-disease classification system, one of several ICU scoring systems. It is applied within 24 hrs of admission of a patient to an Intensive Care Unit (ICU). It is an integer score from 0 to 71 computed based on several measurements. Higher scores correspond to more severe disease and a higher risk of death. It was developed by Knaus et al. in 1981. The score is calculated from 12 physiological parameters like body temperature, blood pH, heart rate, arterial pressure, serum sodium, creatinine, WBC count, age etc. Also, prediction is made by specialist doctors based on their medical experience.

Critical Assessment of Topic Survey

One of the major gaps which was observed during the topic survey phase is that there are many studies regarding survival prediction for certain specific diseases like cancer, sepsis, brain tumor, but very few studies which include many diseases together. We have tried to bridge this gap by including as many diseases which can adversely affect a patient's survival. These diseases are namely liver cirrhosis, cardiovascular, respiratory failure, trauma, sepsis, metabolic, neurologic and gastrointestinal diseases. Also, using the model, time required to predict the survival status is very less as compared to that of APACHE-II Scoring System (24 hrs).

DATA UNDERSTANDING

We have considered the ‘Patient Survival Prediction’ dataset available on Kaggle. Following are the details of our dataset :

- No. of records : 91,713
- No. of columns : 84
- No. of numeric columns : 63
- No. of categorical columns : 22
- Dataset Link : https://www.kaggle.com/mitishaagarwal/patient

Data Dictionary :

1. encounter_id : Unique identifier associated with a patient unit stay
2. patient_id : Unique identifier associated with a patient
3. hospital_id : Unique identifier associated with a hospital
4. age : The age of the patient on unit admission
5. bmi : The body mass index of the person on unit admission
6. elective_surgery : Whether the patient was admitted to the hospital for an elective surgical operation
7. ethnicity : The common national or cultural tradition which the person belongs to
8. gender : Sex of the patient
9. height : The height of the person on unit admission
10. icu_admit_source : The location of the patient prior to being admitted to the unit
11. icu_id : A unique identifier for the unit to which the patient was admitted
12. icu_stay_type : whether the patient was admitted or transferred
13. icu_type : A classification which indicates the type of care the unit is capable of providing
14. pre_icu_los_days : The length of stay of the patient between hospital admission and unit admission
15. weight : The weight (body mass) of the person on unit admission
16. apache_2_diagnosis : The APACHE II diagnosis for the ICU admission
17. apache_3j_diagnosis : The APACHE III-J sub-diagnosis code which best describes the reason for the ICU admission
18. apache_post_operative : The APACHE operative status; 1 for post-operative, 0 for non-operative
19. arf_apache : Whether the patient had acute renal failure during the first 24 hours of their unit stay, efined as a 24 hour urine output <410ml, creatinine >=133 micromol/L and no chronic dialysis

20. gcs_eyes_apache : The eye opening component of the Glasgow Coma Scale measured during the first 24 hours which results in the highest APACHE III score
21. gcs_motor_apache : The motor component of the Glasgow Coma Scale measured during the first 24 hours which results in the highest APACHE III score
22. gcs_unable_apache : Whether the Glasgow Coma Scale was unable to be assessed due to patient sedation
23. gcs_verbal_apache : The verbal component of the Glasgow Coma Scale measured during the first 24 hours which results in the highest APACHE III score
24. heart_rate_apache : The heart rate measured during the first 24 hours which results in the highest APACHE III score
25. intubated_apache : Whether the patient was intubated at the time of the highest scoring arterial blood gas used in the oxygenation score
26. map_apache : The mean arterial pressure measured during the first 24 hours which results in the highest APACHE III score
27. resprate_apache : The respiratory rate measured during the first 24 hours which results in the highest APACHE III score
28. temp_apache : The temperature measured during the first 24 hours which results in the highest APACHE III score
29. ventilated_apache : Whether the patient was invasively ventilated at the time of the highest scoring arterial blood gas using the oxygenation scoring algorithm, including any mode of positive pressure ventilation delivered through a circuit attached to an endo-tracheal tube or tracheostomy
30. d1_diasbp_max : The patient's highest diastolic blood pressure during the first 24 hours of their unit stay, either non-invasively or invasively measured
31. d1_diasbp_min : The patient's lowest diastolic blood pressure during the first 24 hours of their unit stay, either non-invasively or invasively measured
32. d1_diasbp_noninvasive_max : The patient's highest diastolic blood pressure during the first 24 hours of their unit stay, non-invasively measured
33. d1_diasbp_noninvasive_min : The patient's lowest diastolic blood pressure during the first 24 hours of their unit stay, non-invasively measured
34. d1_heartrate_max : The patient's highest heart rate during the first 24 hours of their unit stay
35. d1_heartrate_min : The patient's lowest heart rate during the first 24 hours of their unit stay
36. d1_mbp_max : The patient's highest mean blood pressure during the first 24 hours of their unit stay, either non-invasively or invasively measured
37. d1_mbp_min : The patient's lowest mean blood pressure during the first 24 hours of their unit stay, either non-invasively or invasively measured
38. d1_mbp_noninvasive_max : The patient's highest mean blood pressure during the first 24 hours of their unit stay, non-invasively measured

39. d1_mbp_noninvasive_min : The patient's lowest mean blood pressure during the first 24 hours of their unit stay, non-invasively measured
40. d1_resprate_max : The patient's highest respiratory rate during the first 24 hours of their unit stay
41. d1_resprate_min : The patient's lowest respiratory rate during the first 24 hours of their unit stay
42. d1_spo2_max : The patient's highest peripheral oxygen saturation during the first 24 hours of their unit stay
43. d1_spo2_min : The patient's lowest peripheral oxygen saturation during the first 24 hours of their unit stay
44. d1_sysbp_max : The patient's highest systolic blood pressure during the first 24 hours of their unit stay, either non-invasively or invasively measured
45. d1_sysbp_min : The patient's lowest systolic blood pressure during the first 24 hours of their unit stay, either non-invasively or invasively measured
46. d1_sysbp_noninvasive_max : The patient's highest systolic blood pressure during the first 24 hours of their unit stay, invasively measured
47. d1_sysbp_noninvasive_min : The patient's lowest systolic blood pressure during the first 24 hours of their unit stay, invasively measured
48. d1_temp_max : The patient's highest core temperature during the first 24 hours of their unit stay, invasively measured
49. d1_temp_min : The patient's lowest core temperature during the first 24 hours of their unit stay
50. h1_diasbp_max : The patient's highest diastolic blood pressure during the first hour of their unit stay, either non-invasively or invasively measured
51. h1_diasbp_min : The patient's lowest diastolic blood pressure during the first hour of their unit stay, either non-invasively or invasively measured
52. h1_diasbp_noninvasive_max : The patient's highest diastolic blood pressure during the first hour of their unit stay, invasively measured
53. h1_diasbp_noninvasive_min : The patient's lowest diastolic blood pressure during the first hour of their unit stay, invasively measured
54. h1_heartrate_max : The patient's highest heart rate during the first hour of their unit stay
55. h1_heartrate_min : The patient's lowest heart rate during the first hour of their unit stay
56. h1_mbp_max : The patient's highest mean blood pressure during the first hour of their unit stay, either non-invasively or invasively measured
57. h1_mbp_min : The patient's lowest mean blood pressure during the first hour of their unit stay, either non-invasively or invasively measured
58. h1_mbp_noninvasive_max : The patient's highest mean blood pressure during the first hour of their unit stay, non-invasively measured
59. h1_mbp_noninvasive_min : The patient's lowest mean blood pressure during the first hour of their unit stay, non-invasively measured

60. h1_resprate_max : The patient's highest respiratory rate during the first hour of their unit stay
61. h1_resprate_min : The patient's lowest respiratory rate during the first hour of their unit stay
62. h1_spo2_max : The patient's highest peripheral oxygen saturation during the first hour of their unit stay
63. h1_spo2_min : The patient's lowest peripheral oxygen saturation during the first hour of their unit stay
64. h1_sysbp_max : The patient's highest systolic blood pressure during the first hour of their unit stay, either non-invasively or invasively measured
65. h1_sysbp_min : The patient's lowest systolic blood pressure during the first hour of their unit stay, either non-invasively or invasively measured
66. h1_sysbp_noninvasive_max : The patient's highest systolic blood pressure during the first hour of their unit stay, non-invasively measured
67. h1_sysbp_noninvasive_min : The patient's lowest systolic blood pressure during the first hour of their unit stay, non-invasively measured
68. d1_glucose_max : The highest glucose concentration of the patient in their serum or plasma during the first 24 hours of their unit stay
69. d1_glucose_min : The lowest glucose concentration of the patient in their serum or plasma during the first 24 hours of their unit stay70. d1_potassium_max : The highest potassium concentration for the patient in their serum or plasma during the first 24 hours of their unit stay
70. d1_potassium_min : The lowest potassium concentration for the patient in their serum or plasma during the first 24 hours of their unit stay
71. apache_4a_hospital_death_prob : The APACHE IVa probabilistic prediction of in-hospital mortality for the patient which utilizes the APACHE III score and other covariates, including diagnosis.
72. apache_4a_icu_death_prob : The APACHE IVa probabilistic prediction of in ICU mortality for the patient which utilizes the APACHE III score and other covariates, including diagnosis
73. aids : Whether the patient has a definitive diagnosis of acquired immune deficiency syndrome (AIDS) (not HIV positive alone)
74. cirrhosis : Whether the patient has a history of heavy alcohol use with portal hypertension and varices, other causes of cirrhosis with evidence of portal hypertension and varices, or biopsy proven cirrhosis. This comorbidity does not apply to patients with a functioning liver transplant.
75. diabetes_mellitus : Whether the patient has been diagnosed with diabetes, either juvenile or adult onset, which requires medication.
76. hepatic_failure : Whether the patient has cirrhosis and additional complications including jaundice and ascites, upper GI bleeding, hepatic encephalopathy, or coma.

77. immunosuppression : Whether the patient has their immune system suppressed within six months prior to ICU admission for any of the following reasons; radiation therapy, chemotherapy, use of non-cytotoxic immunosuppressive drugs, high dose steroids (at least 0.3 mg/kg/day of methylprednisolone or equivalent for at least 6 months).
78. leukemia : Whether the patient has been diagnosed with acute or chronic myelogenous leukemia, acute or chronic lymphocytic leukemia, or multiple myeloma.
79. lymphoma : Whether the patient has been diagnosed with non-Hodgkin lymphoma.
80. solid_tumor_with_metastasis : Whether the patient has been diagnosed with any solid tumor carcinoma (including malignant melanoma) which has evidence of metastasis.
81. apache_3j_bodysystem : Admission diagnosis group for APACHE III
82. apache_2_bodysystem : Admission diagnosis group for APACHE II
83. Unnamed : NA
84. hospital_death : Whether the patient died during this hospitalization

PRE-PROCESSING & STATISTICAL ANALYSIS

Data Pre-processing :

- Dataset shape before null value treatment : (91713 , 84)
- Dataset shape after null value treatment : (72454 , 80)
- Dataset shape after treating multi-collinearity : (72454, 58)
- Redundant Columns : 5

Multi-Collinearity :

We have dropped columns having correlation coefficient greater than 0.8. In total we have dropped 22 columns. Dataset shape after treating multi-collinearity : (72454, 58)

Outlier Analysis :

Outliers were present in the dataset for many columns. Instead of removing the outliers completely which might have resulted in loss of important information, we have used the capping technique to cap the outliers. The outliers above (Q3 + 1.5*IQR) were capped at (Q3 + 1.5*IQR) and those below (Q1 – 1.5*IQR) were capped at (Q1 – 1.5*IQR).

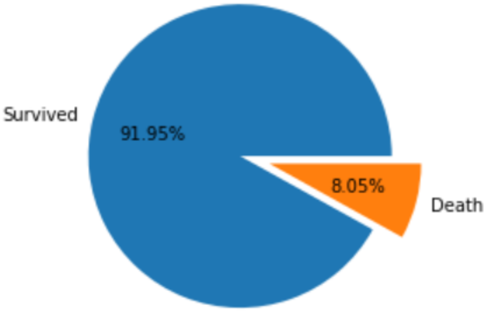
Handling imbalance in the Dataset :

Degree of imbalance Proportion of Minority Class:

- Mild : 20 - 40 % of dataset
- Moderate : 1 - 20 % of dataset
- Extreme : < 1 % of dataset

With respect to the above metric, the imbalance in our dataset falls in the category of moderate imbalance. Imbalance in the classes might lead to the algorithm being biased towards the majority class. We have experimented with various balancing techniques including SMOTE, SMOTE + Tomek T-Links, Under and Over Sampling and chosen the one which yields the best results.

Initial Distribution of the target column :



Feature Engineering :

Scaling the Data – We have scaled the numeric columns in the dataset using StandardSclaer() function from sklearn library.

Dummy Encoding – Also, we have encoded the categorical columns using One Hot Encoding.

Statistical Analysis :

The conclusions of the analysis are as below :

Chi-Sq Test for Independence of Attributes :

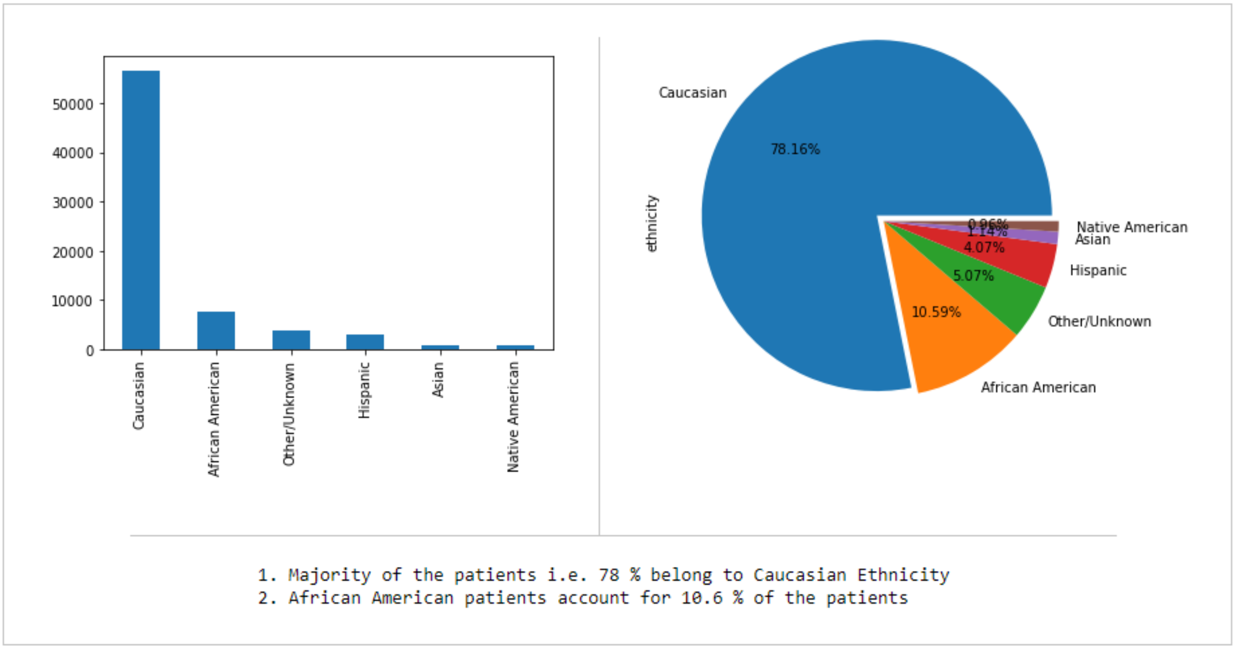
- The attributes – gender and hospital death are independent. Both males and females are at an equal risk of surviving or not surviving a particular medical condition.
- Ethnicity and hospital death are dependent.
- ICU admit source and hospital death are dependent.
- ICU type and hospital death are dependent.
- Apache_3j_bodysystem and hospital death are dependent.
- Ventilated_or_not and hospital death are dependent.

Z-Test for Equality of Means :

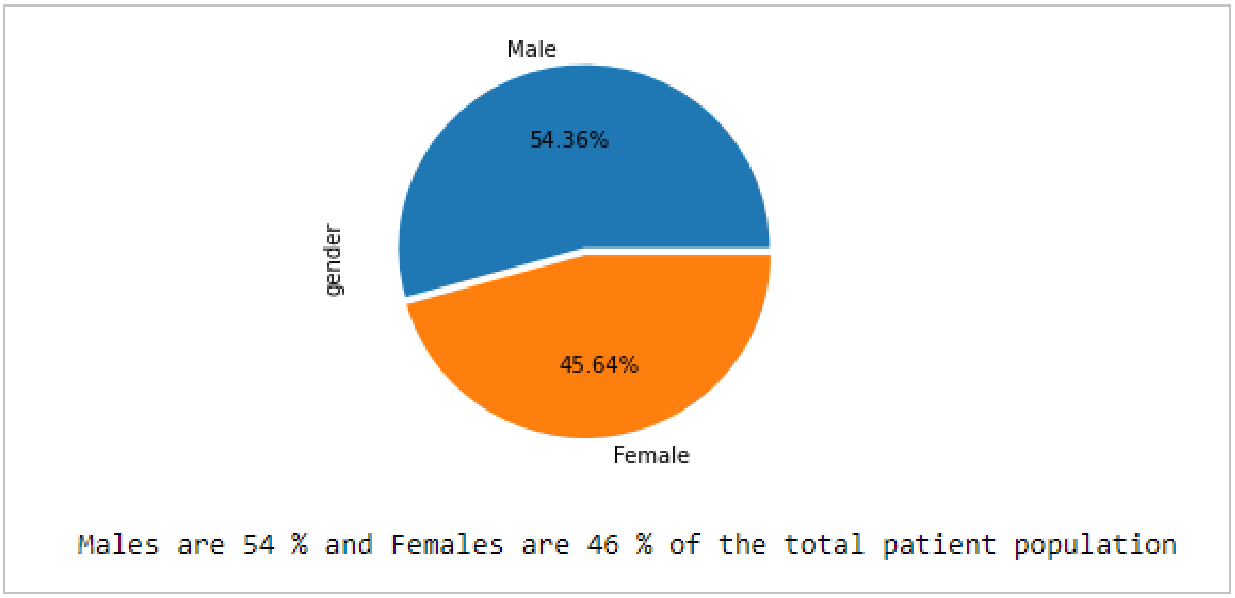
- The cases resulting in death are having a higher value of pre-icu loss-days.

DATA EXPLORATION (EDA)

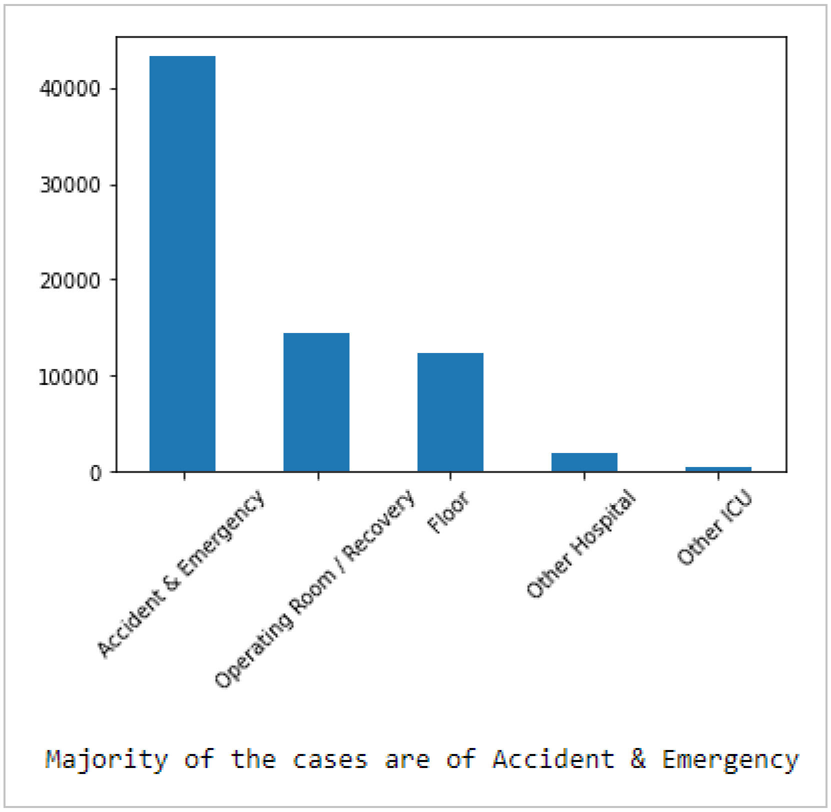
The findings from the dataset are summarized below :



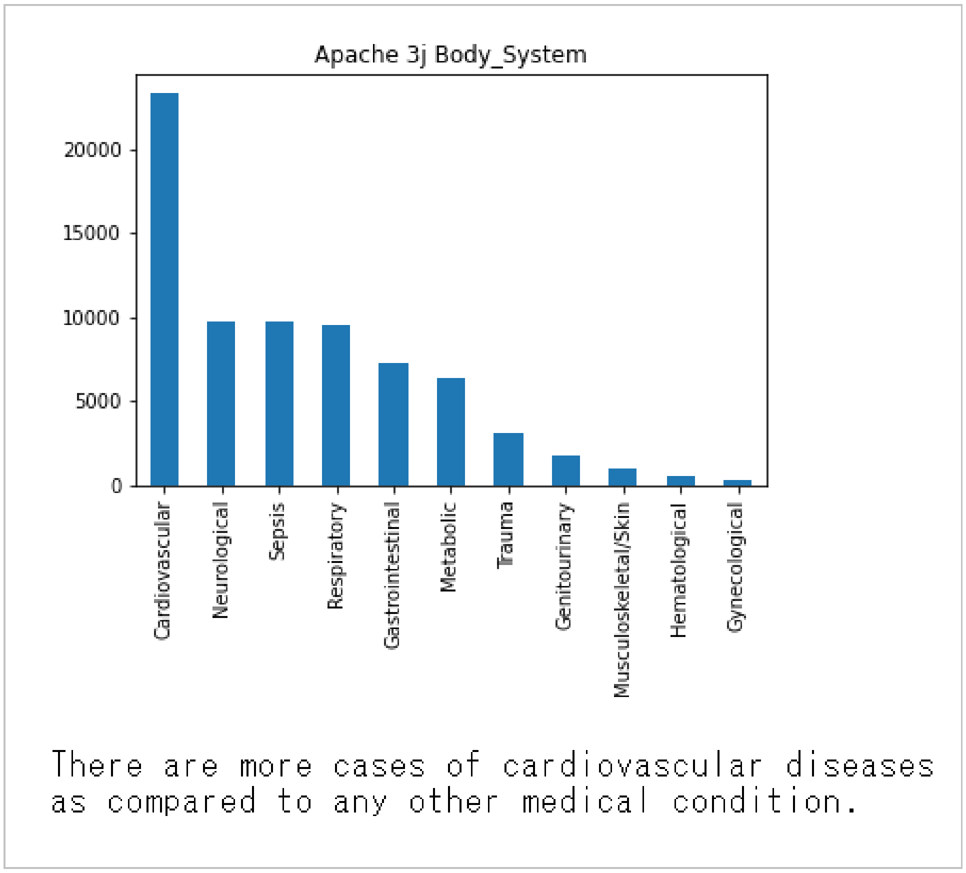
```
1. Majority of the patients i.e. 78 % belong to Caucasian Ethnicity
2. African American patients account for 10.6 % of the patients
```



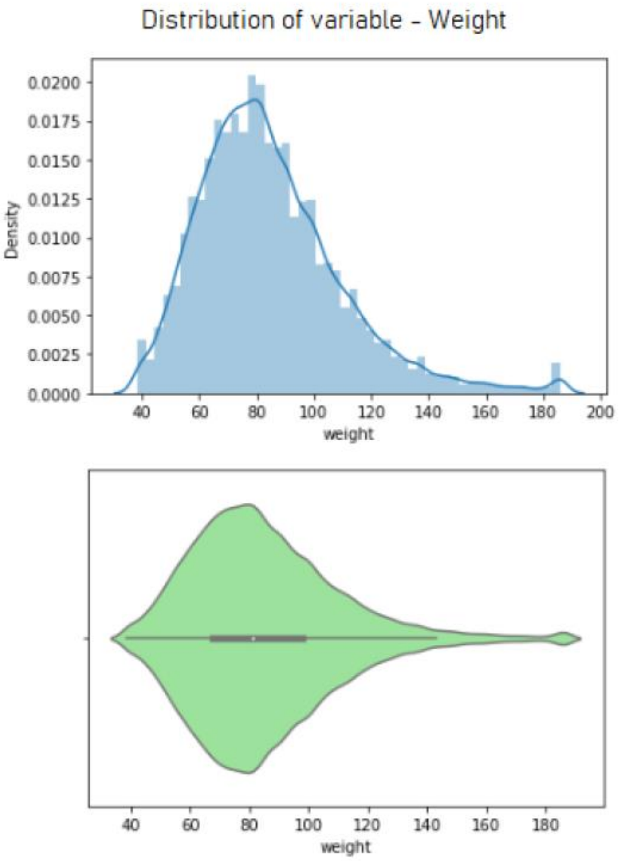
```
Males are 54 % and Females are 46 % of the total patient population
```



Majority of the cases are of Accident & Emergency

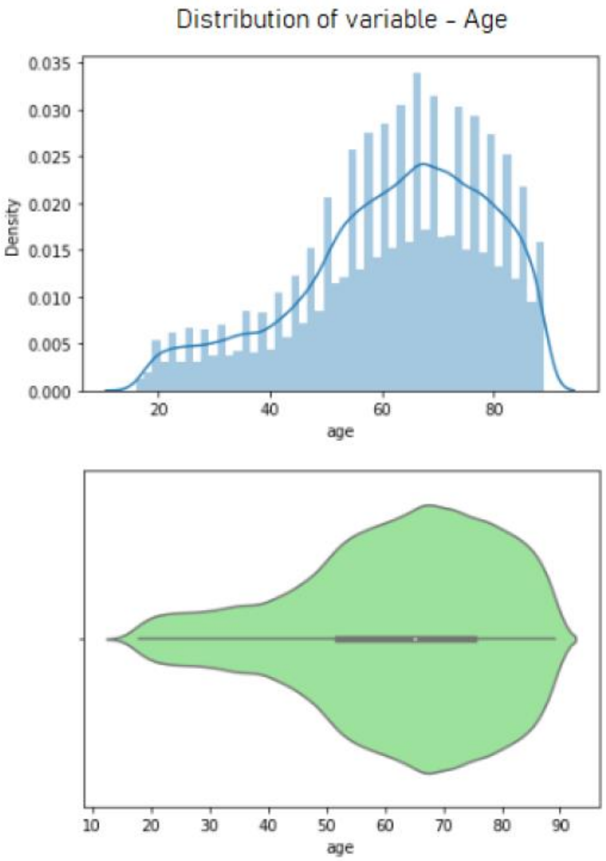


There are more cases of cardiovascular diseases as compared to any other medical condition.



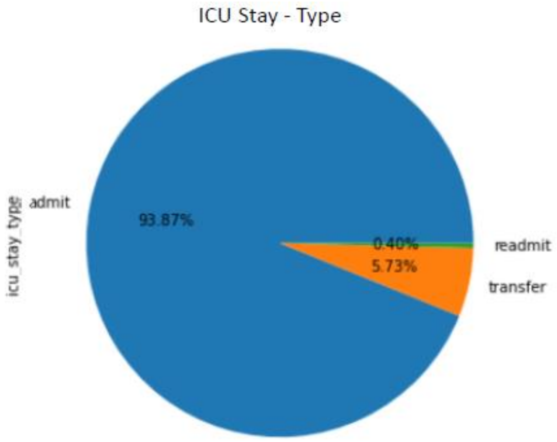
```
Skewness of Weight :  1.066220554588028
Mode of variable Weight :  84.55268653214341
Median of variable Weight :  81.0
```

The distribution of variable weight is positively skewed.
Majority of the patients belong to the weight category of 60 - 100 kg.

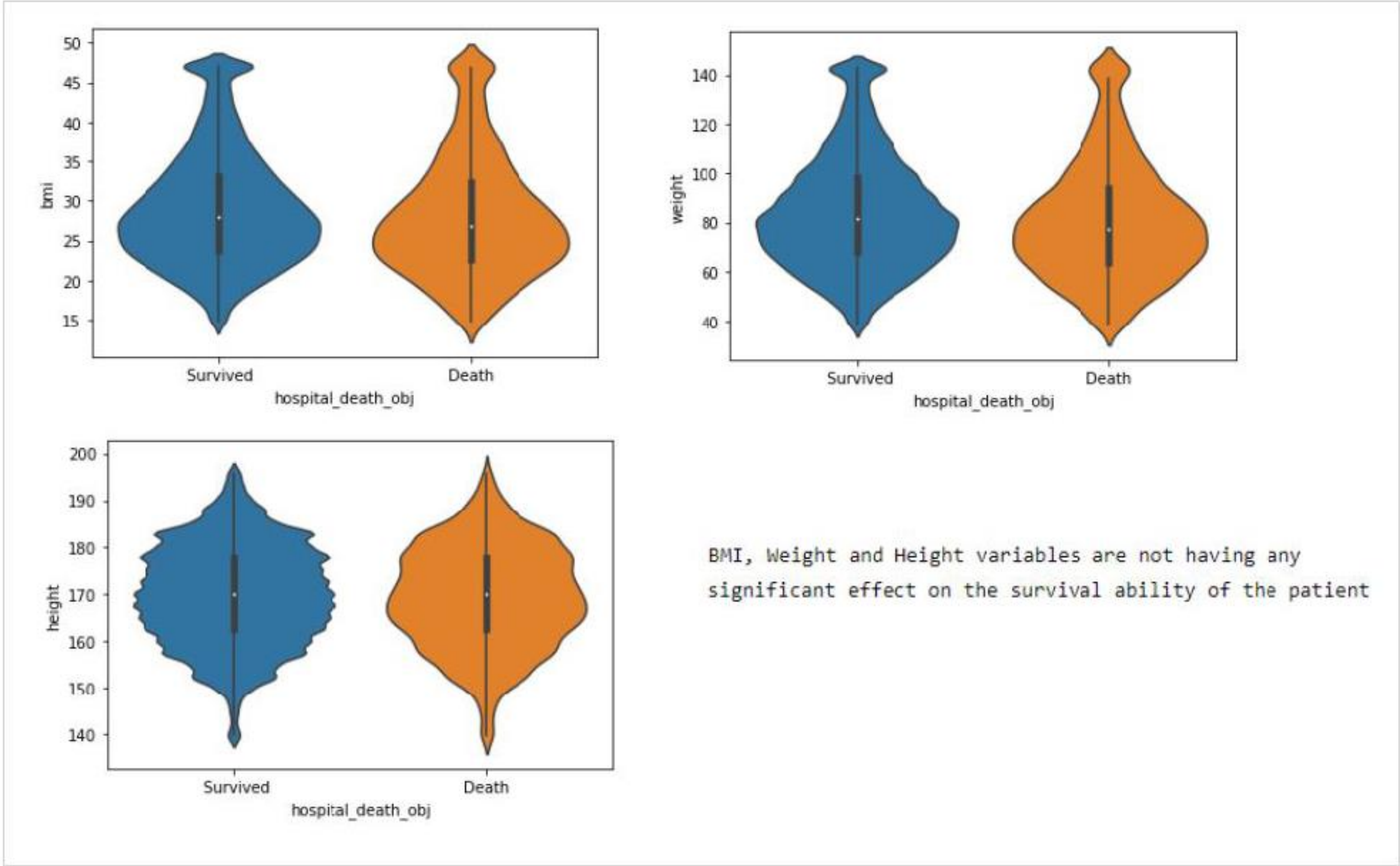


```
Skewness of Age :  -0.6270827791347964
Mode of variable Age :  67.0 yrs
Median of variable Age :  65.0 yrs
```

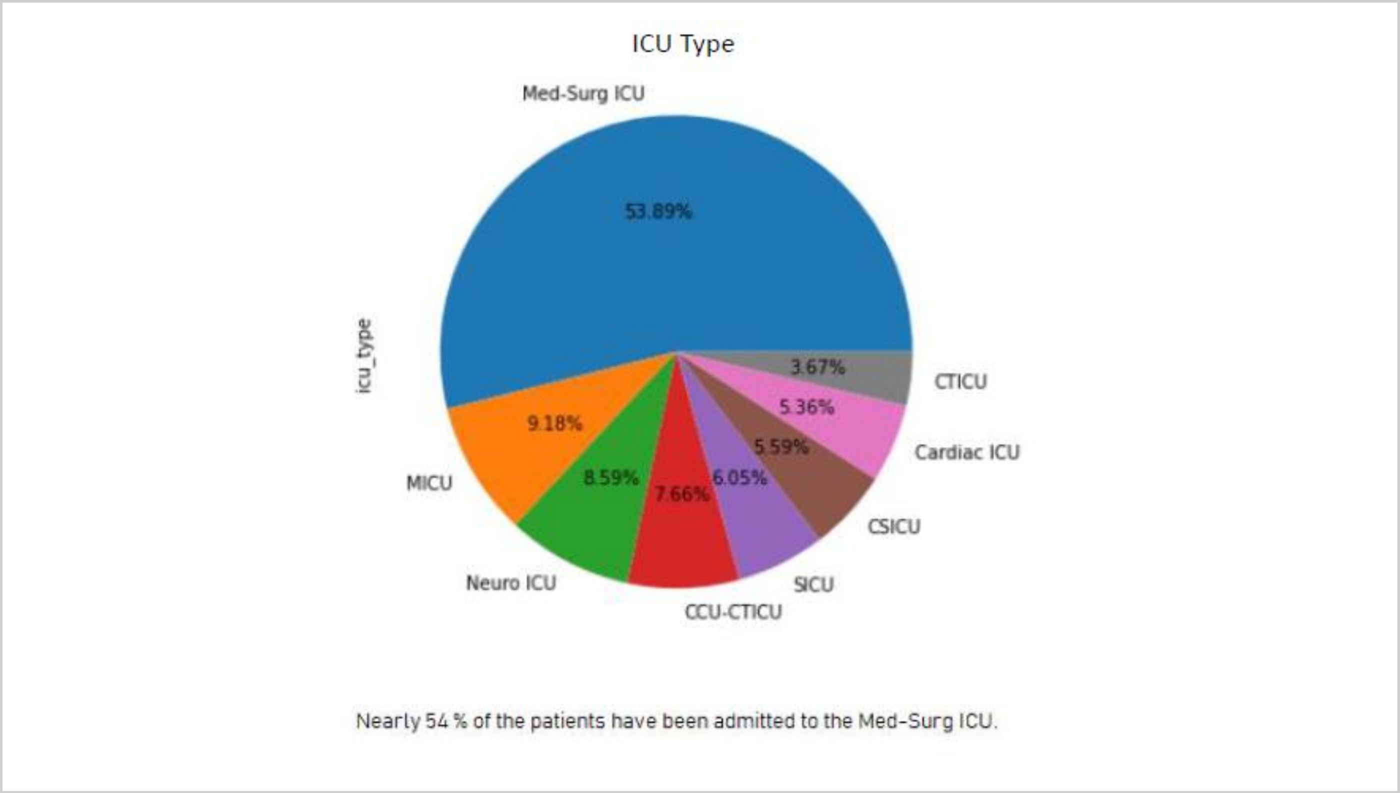
The distribution of variable age is negatively skewed.
Majority of the patients belong to the age group of 55 - 85 yrs.



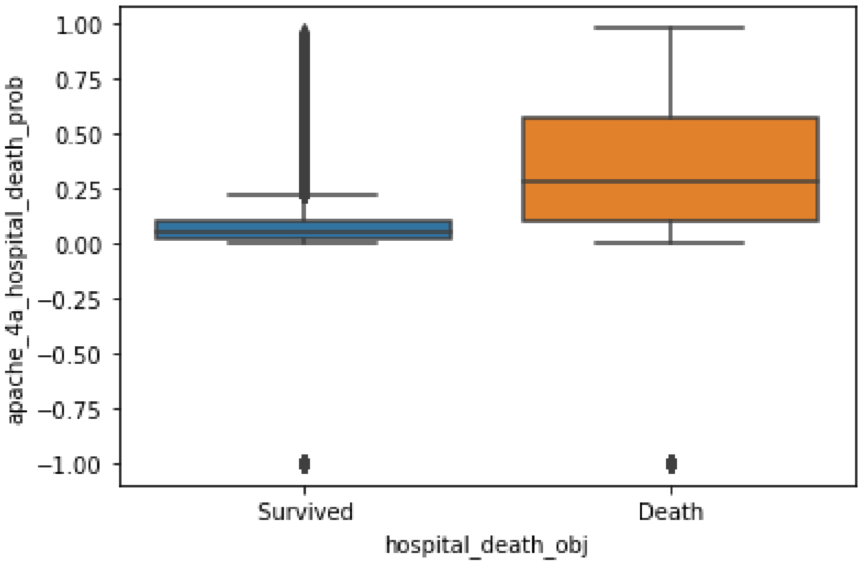
Nearly 94 % of the cases are of admit as compared to 6 % of transfer and only 0.4 % of readmit.



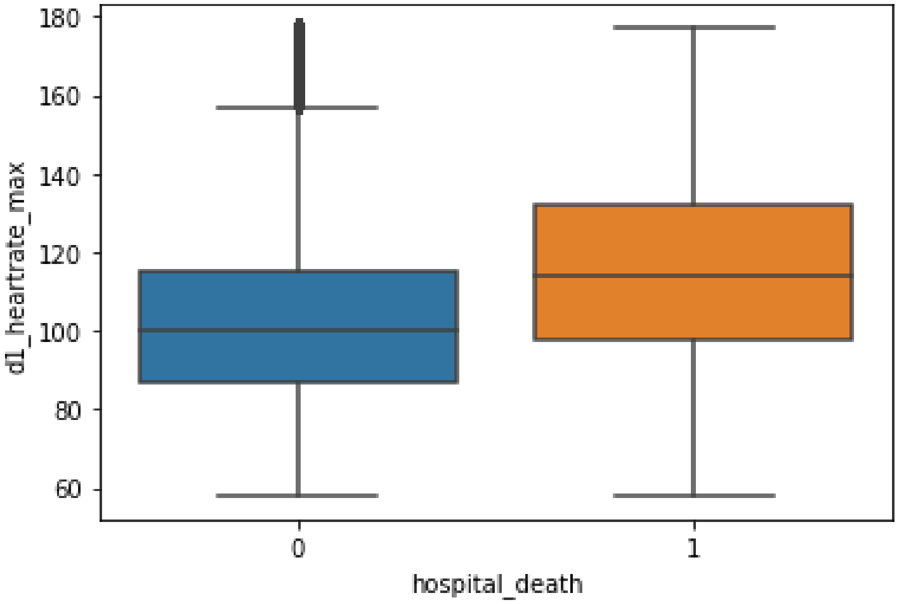
BMI, Weight and Height variables are not having any significant effect on the survival ability of the patient



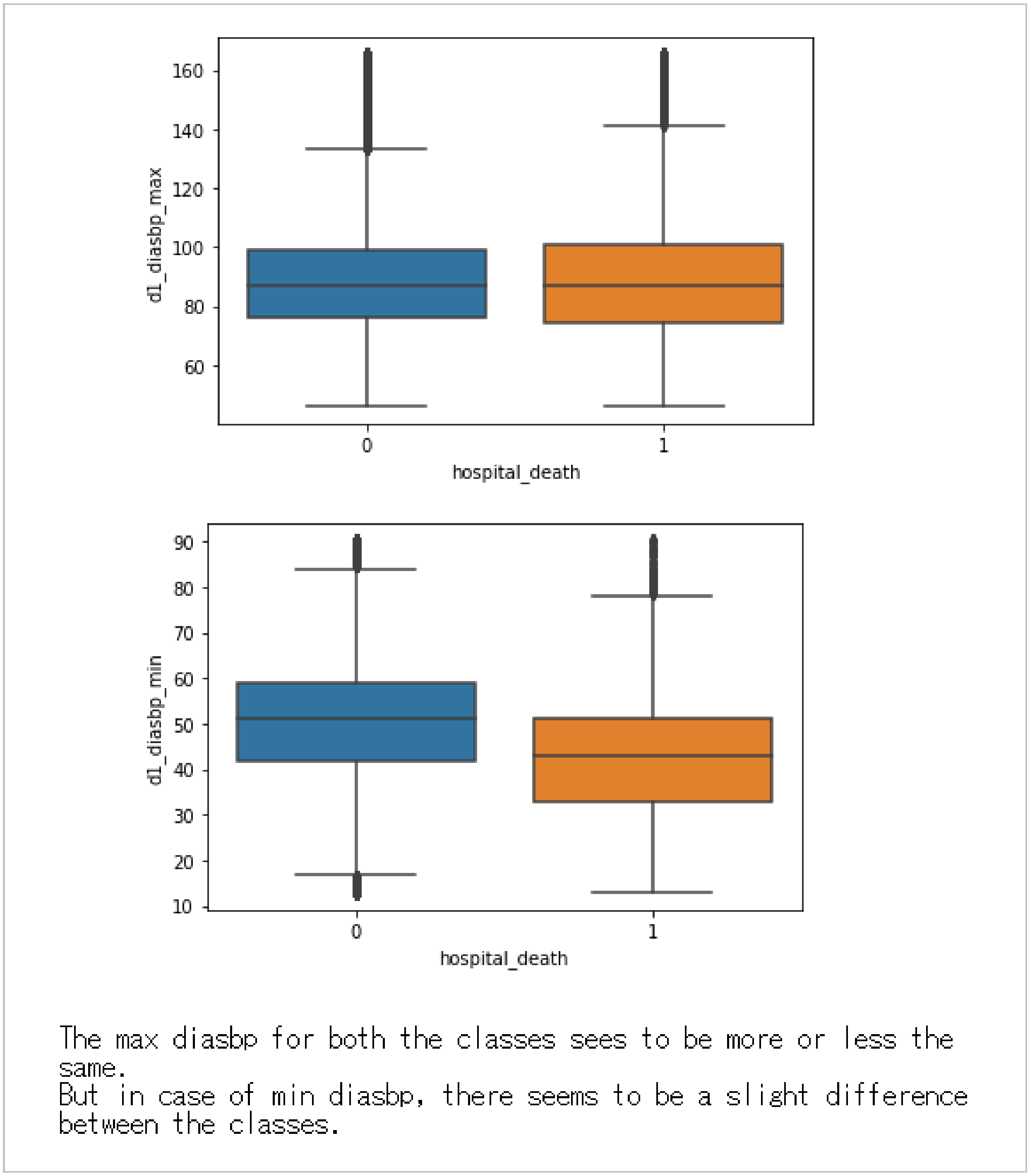
Nearly 54 % of the patients have been admitted to the Med-Surg ICU.



Cases which have resulted in death are having a higher value of apache_4a_hospital_death_prob.



There seems to be a difference between the max heartrate for the two classes.



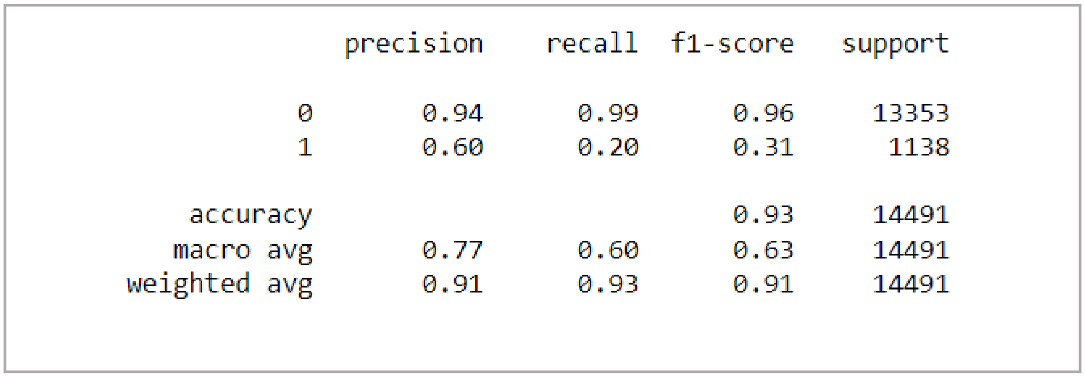
The max diasbp for both the classes sees to be more or less the same.
But in case of min diasbp, there seems to be a slight difference between the classes.

BASIC MODEL BUILDING STEPS

As our problem is a classification problem, we have experimented with different classification algorithms for building the base model including Logistic Regression, KNN, Decision Tree, Random Forest, Boosting Techniques etc.

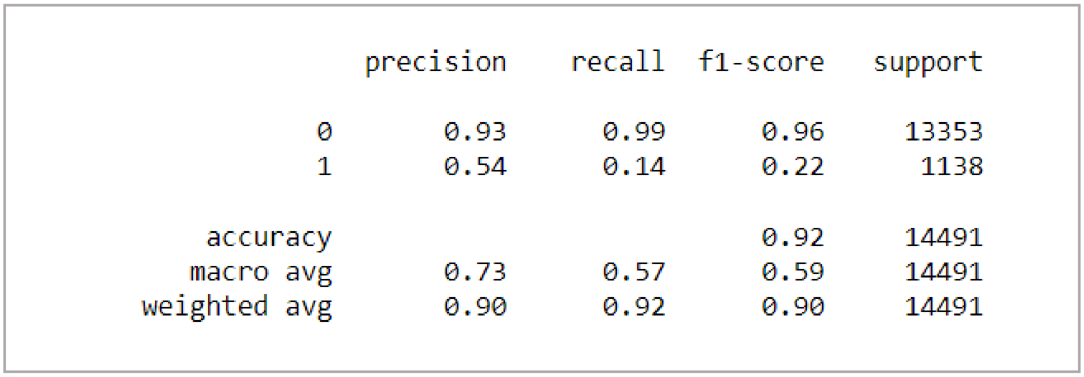
We have split the dataset into train and test set in the ratio 80:20. The test classification results of some of the base models considering different algorithms are summarized below :

- **Logistic Regression :**



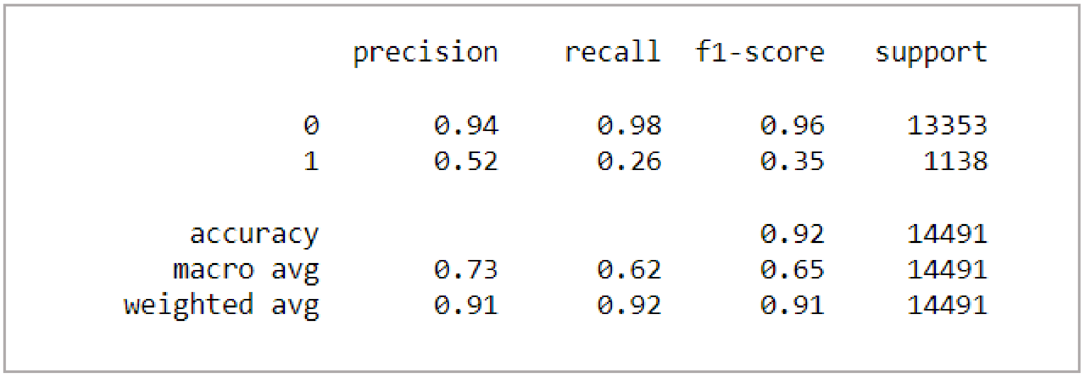
```
              precision    recall  f1-score   support

           0       0.94      0.99      0.96     13353
           1       0.60      0.20      0.31      1138

    accuracy                           0.93     14491
   macro avg       0.77      0.60      0.63     14491
weighted avg       0.91      0.93      0.91     14491
```

- **K- Nearest Neighbors (KNN) :**



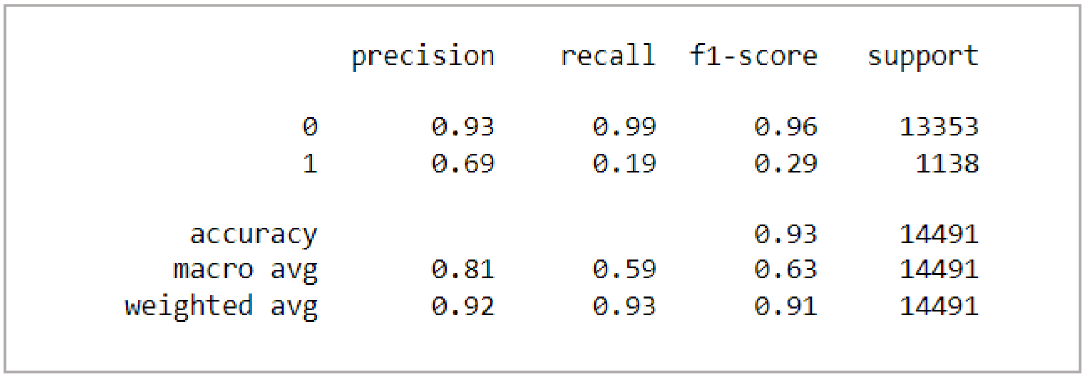
```
              precision    recall  f1-score   support

           0       0.93      0.99      0.96     13353
           1       0.54      0.14      0.22      1138

    accuracy                           0.92     14491
   macro avg       0.73      0.57      0.59     14491
weighted avg       0.90      0.92      0.90     14491
```

- **Decision Tree :**



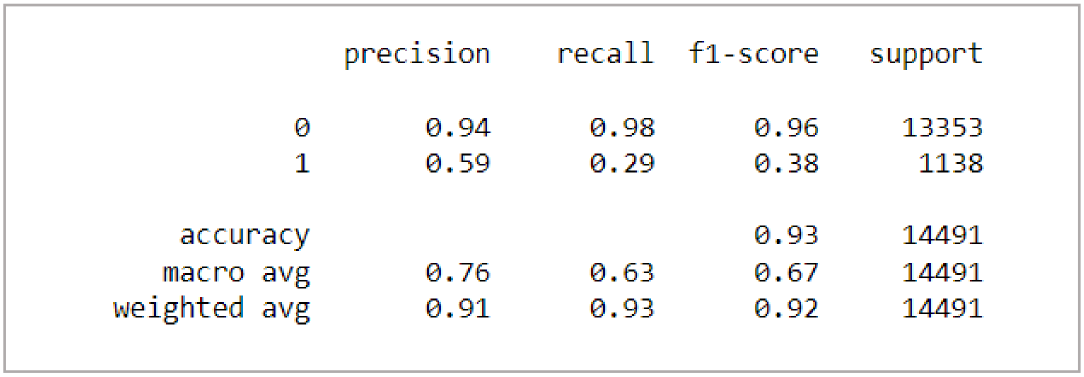
```
              precision    recall  f1-score   support

           0       0.94      0.98      0.96     13353
           1       0.52      0.26      0.35      1138

    accuracy                           0.92     14491
   macro avg       0.73      0.62      0.65     14491
weighted avg       0.91      0.92      0.91     14491
```

- **Random Forest :**



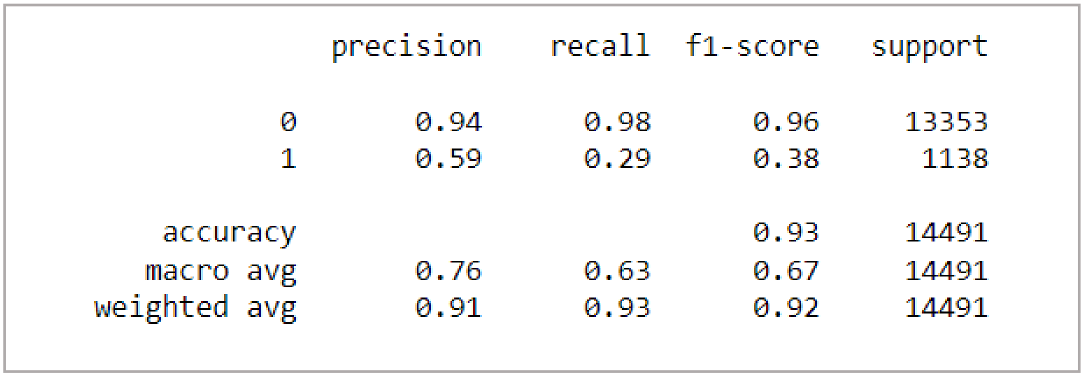
```
              precision    recall  f1-score   support

           0       0.93      0.99      0.96     13353
           1       0.69      0.19      0.29      1138

    accuracy                           0.93     14491
   macro avg       0.81      0.59      0.63     14491
weighted avg       0.92      0.93      0.91     14491
```

- **AdaBoost :**



```
              precision    recall  f1-score   support

           0       0.94      0.98      0.96     13353
           1       0.59      0.29      0.38      1138

    accuracy                           0.93     14491
   macro avg       0.76      0.63      0.67     14491
weighted avg       0.91      0.93      0.92     14491
```

- **XGBoost :**



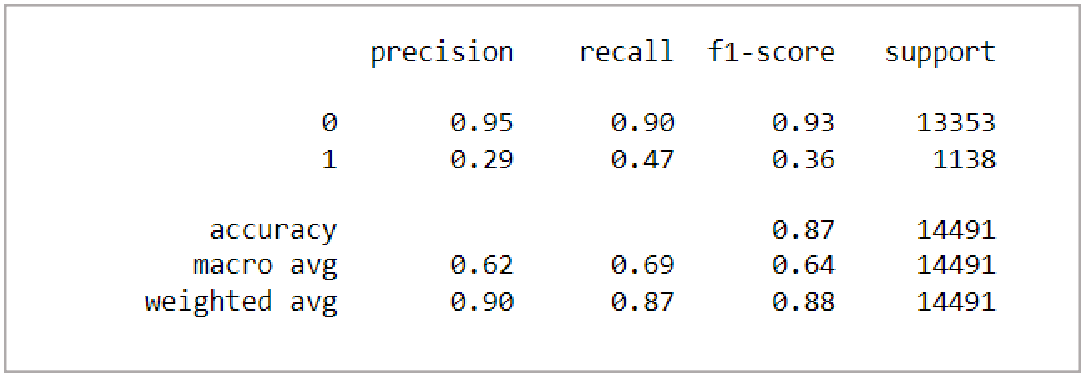
```
              precision    recall  f1-score   support

           0       0.94      0.98      0.96     13353
           1       0.59      0.29      0.38      1138

    accuracy                           0.93     14491
   macro avg       0.76      0.63      0.67     14491
weighted avg       0.91      0.93      0.92     14491
```

Observing the classification reports for test data for the different algorithms, we are not getting good results. The values of different metrics like precision, recall, f1-score for the minority class (1) i.e. hospital death are very low. We have also experimented with Stacking technique in which we have used AdaBoost and XGBoost along with Decision Tree as the final estimator. But we have not got any significant improvement in results.

USING SMOTE TO HANDLE IMBALANCE

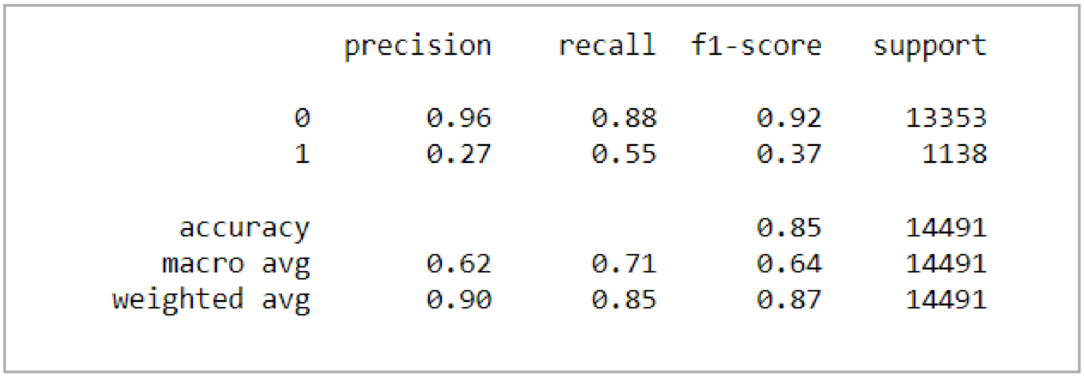
To overcome the poor values of metrics in the case of the minority class, we decided to use SMOTE (Synthetic Minority Oversampling Technique). Using SMOTE we have balanced the classes in the ratio 50:50 where the initial proportion was 92:8. The results after applying SMOTE are as follows :

- **Logistic Regression - SMOTE :**



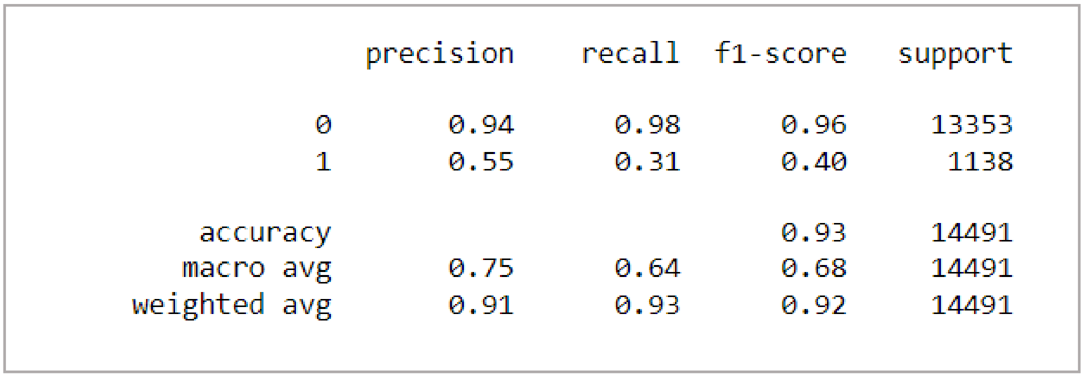
```
              precision    recall  f1-score   support

           0       0.95      0.90      0.93     13353
           1       0.29      0.47      0.36      1138

    accuracy                           0.87     14491
   macro avg       0.62      0.69      0.64     14491
weighted avg       0.90      0.87      0.88     14491
```

- **AdaBoost - SMOTE :**



```
              precision    recall  f1-score   support

           0       0.96      0.88      0.92     13353
           1       0.27      0.55      0.37      1138

    accuracy                           0.85     14491
   macro avg       0.62      0.71      0.64     14491
weighted avg       0.90      0.85      0.87     14491
```

- **XGBoost - SMOTE :**

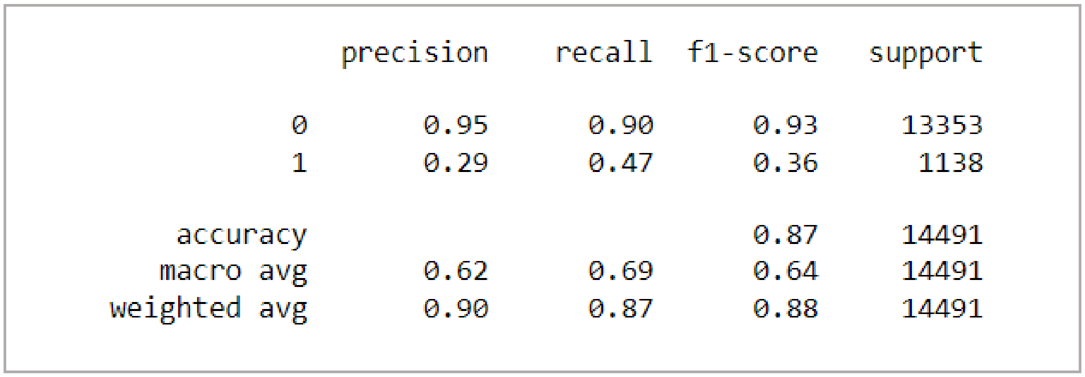


```
              precision    recall  f1-score   support

           0       0.94      0.98      0.96     13353
           1       0.55      0.31      0.40      1138

    accuracy                           0.93     14491
   macro avg       0.75      0.64      0.68     14491
weighted avg       0.91      0.93      0.92     14491
```

Looking at the classification report after applying SMOTE to balance the classes, we can observe that the recall has improved to a certain extent but still it is not good enough.

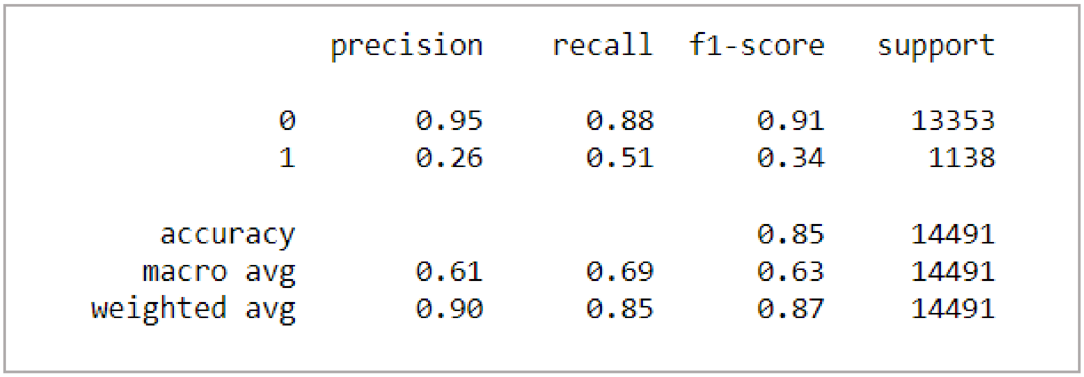
So, instead of only up-sampling the minority class using SMOTE, we decided to combine SMOTE up-sampling with Tomek T-Links down-sampling technique. We have used the SMOTETomek function from the imblearn library. The results using this approach are as follows :

- **Logistic Regression – SMOTE + Tomek T-Links :**



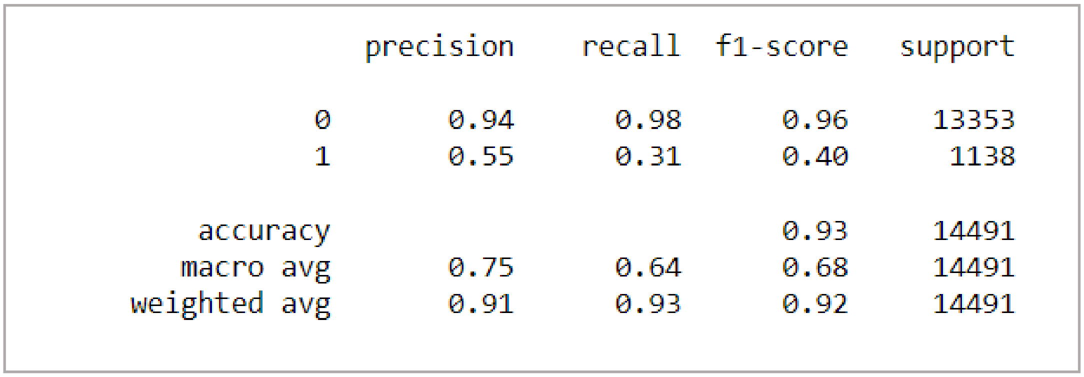
```
              precision    recall  f1-score   support

           0       0.95      0.90      0.93     13353
           1       0.29      0.47      0.36      1138

    accuracy                           0.87     14491
   macro avg       0.62      0.69      0.64     14491
weighted avg       0.90      0.87      0.88     14491
```

- **AdaBoost – SMOTE + Tomek T-Links :**



```
              precision    recall  f1-score   support

           0       0.95      0.88      0.91     13353
           1       0.26      0.51      0.34      1138

    accuracy                           0.85     14491
   macro avg       0.61      0.69      0.63     14491
weighted avg       0.90      0.85      0.87     14491
```

- **XGBoost - SMOTE + Tomek T-Links:**



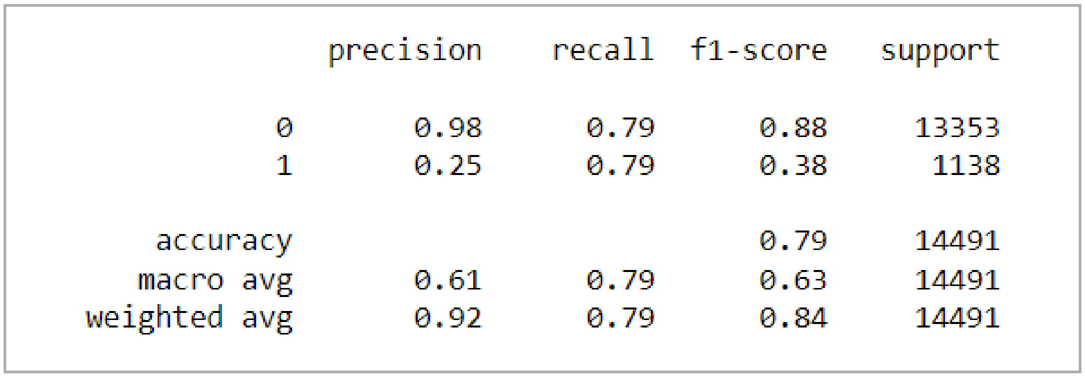
```
              precision    recall  f1-score   support

           0       0.94      0.98      0.96     13353
           1       0.55      0.31      0.40      1138

    accuracy                           0.93     14491
   macro avg       0.75      0.64      0.68     14491
weighted avg       0.91      0.93      0.92     14491
```

With this approach as well, we are not getting any significant improvements in results.

USING UNDER & OVER SAMPLING COMBINATION

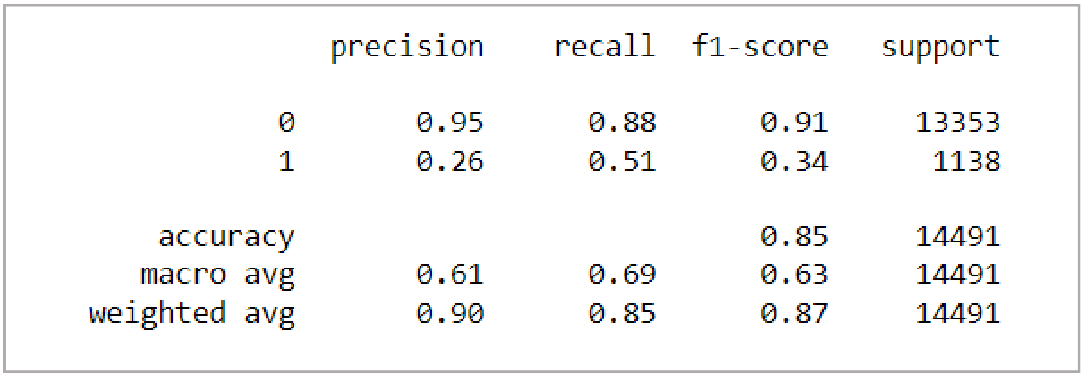
Using SMOTE and SMOTE + Tomek T-Links, we have not got any good results. We decided to use a combination of under sampling and over sampling. In this approach we have used RandomOverSampler and RandomUnderSampler function in combination from the imblearn library. Instead of taking the 50:50 balancing approach like in SMOTE, here we first decided to down-sample the majority class to 25 % and then over-sample the minority class to equal the majority class. The results are as follows :

- **Logistic Regression – 25 % Down-Sampling + Up-Sampling :**



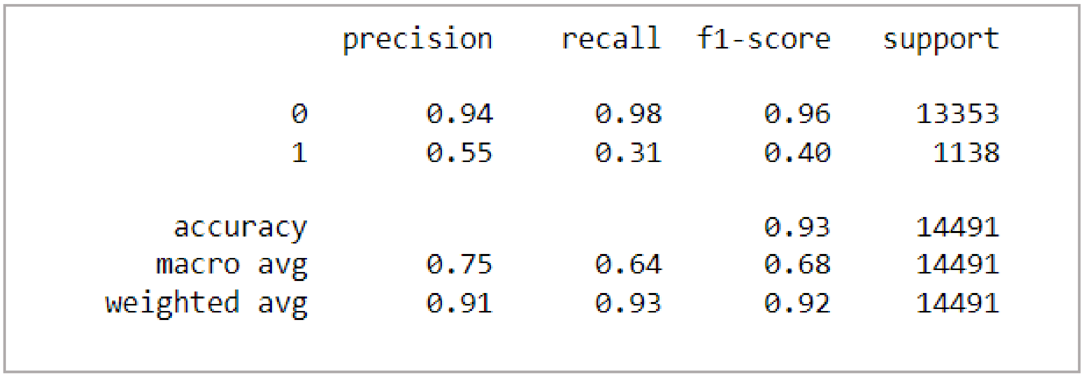
```
              precision    recall  f1-score   support

           0       0.98      0.79      0.88     13353
           1       0.25      0.79      0.38      1138

    accuracy                           0.79     14491
   macro avg       0.61      0.79      0.63     14491
weighted avg       0.92      0.79      0.84     14491
```

- **AdaBoost – 25 % Down-Sampling + Up-Sampling :**



```
              precision    recall  f1-score   support

           0       0.95      0.88      0.91     13353
           1       0.26      0.51      0.34      1138

    accuracy                           0.85     14491
   macro avg       0.61      0.69      0.63     14491
weighted avg       0.90      0.85      0.87     14491
```

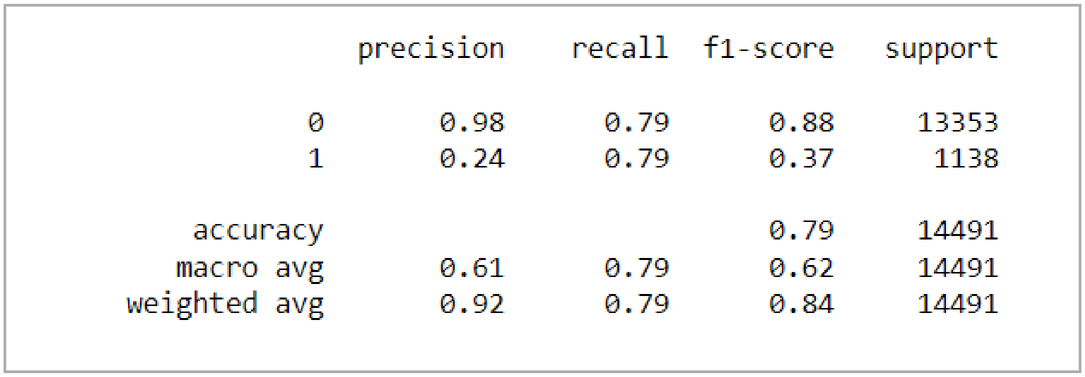
- **XGBoost - 25 % Down-Sampling + Up-Sampling :**



```
              precision    recall  f1-score   support

           0       0.94      0.98      0.96     13353
           1       0.55      0.31      0.40      1138

    accuracy                           0.93     14491
   macro avg       0.75      0.64      0.68     14491
weighted avg       0.91      0.93      0.92     14491
```

With this approach, our recall score has increased significantly for the minority class reaching up to 0.79 in the case of Logistic Regression. We have experimented the same approach but here instead of down-sampling to 25 %, we have considered down-sampling to 50 %. And for Logistic Regression, we have obtained similar results. The classification report is as follows :

- **Logistic Regression – 50 % Down-Sampling + Up-Sampling :**



```
              precision    recall  f1-score   support

           0       0.98      0.79      0.88     13353
           1       0.24      0.79      0.37      1138

    accuracy                           0.79     14491
   macro avg       0.61      0.79      0.62     14491
weighted avg       0.92      0.79      0.84     14491
```

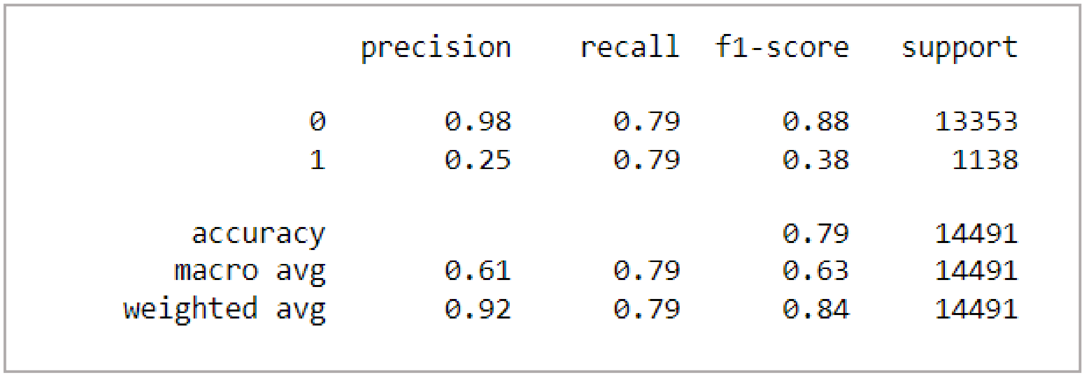
Using the above approach, we are getting good results with Logistic Regression Algorithm. We have also experimented with Balanced Class Weights approach where we have balanced the class weights so that both the classes get equal representation and reduce the bias of the model towards the majority class. The results have been displayed in the next section.

In our model we are focusing on Recall Score as the main evaluation metric. As our model belongs to the Healthcare Domain, Type-II error is considered to be riskier in this case. Type-II error happens when the model predicts that a patient is likely to survive when in reality, he is highly likely to die. This is a big risk because this will mislead the doctors and the medical staff into believing that the patient is not in a critical condition but in reality, he is likely to die and needs immediate medical attention. The cases which are likely to die but the model predicts otherwise, such cases are termed as False Negatives (FN). The higher the number of False Negatives (FN) reported, higher is the Type-II error. Recall Score considers FNs and higher is the Recall Score, minimum is the number of FNs. So, we are considering Recall Score as our main evaluation metric.

USING BALANCED CLASS WEIGHTS APPROACH

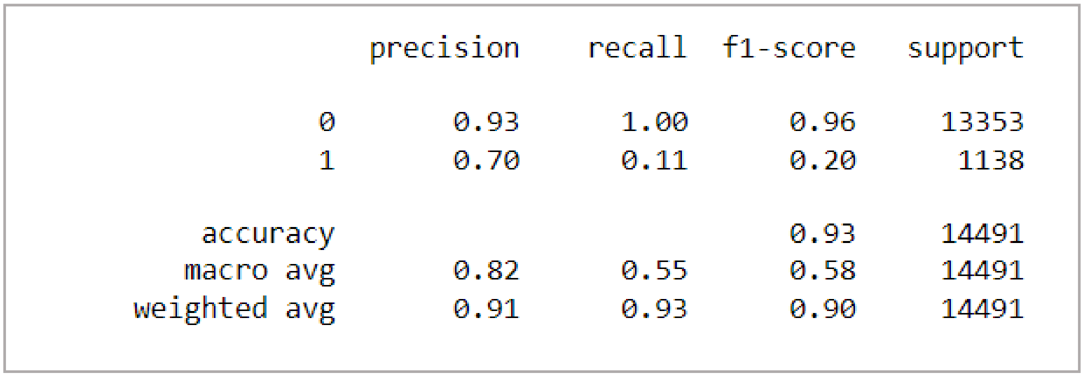
In this approach, we have used balanced class weights so that both the classes get equal weightage in the model in spite of the imbalance. The results are as follows :

- **Logistic Regression – Balanced Class Weights :**



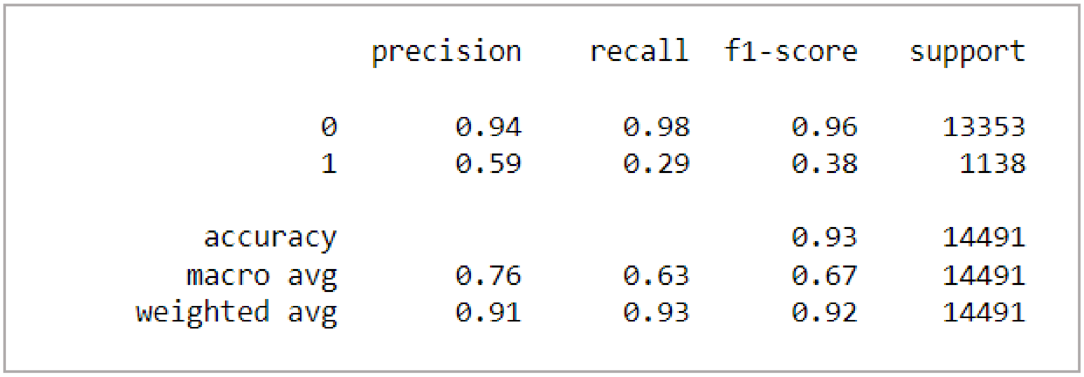
```
              precision    recall  f1-score   support

           0       0.98      0.79      0.88     13353
           1       0.25      0.79      0.38      1138

    accuracy                           0.79     14491
   macro avg       0.61      0.79      0.63     14491
weighted avg       0.92      0.79      0.84     14491
```

- **Random Forest – Balanced Class Weights :**



```
              precision    recall  f1-score   support

           0       0.93      1.00      0.96     13353
           1       0.70      0.11      0.20      1138

    accuracy                           0.93     14491
   macro avg       0.82      0.55      0.58     14491
weighted avg       0.91      0.93      0.90     14491
```

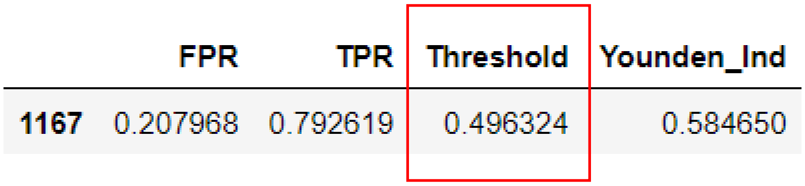
- **XGBoost – Balanced Class Weights :**



```
              precision    recall  f1-score   support

           0       0.94      0.98      0.96     13353
           1       0.59      0.29      0.38      1138

    accuracy                           0.93     14491
   macro avg       0.76      0.63      0.67     14491
weighted avg       0.91      0.93      0.92     14491
```

With this approach we are getting good Recall Score for Logistic Regression. With Random Forest and XGBoost, we are not getting good results.

Considering the Logistic Regression Model above, we have tried to find the best threshold value using Youden's Index approach and the threshold giving the highest score for Youden's Index is considered the best threshold value.



	FPR	TPR	Threshold	Younden_Ind
1167	0.207968	0.792619	0.496324	0.584650

We are getting 0.496 as the best threshold value which is nearly equal to 0.5 which is the default threshold for Logistic Regression. We consider this model as our final model as among all the models, it is giving us the most reliable results. Also, the Logistic Regression model with 25 % down-sampling and up-sampling approach is equally a good model.

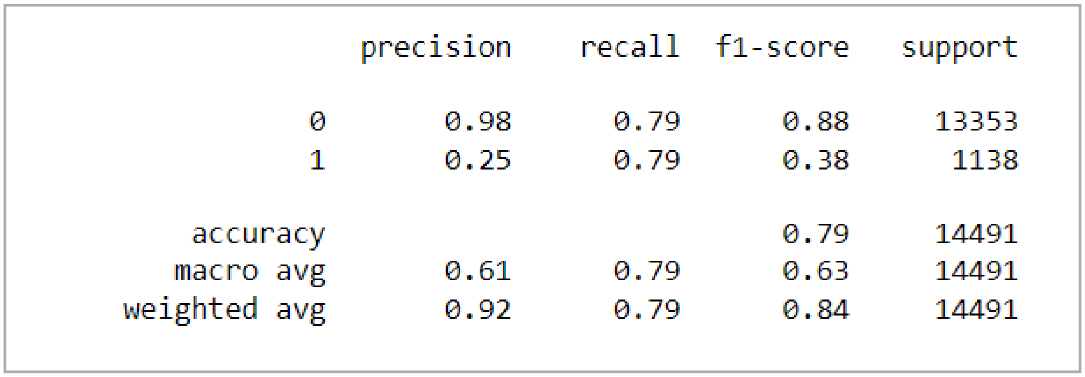
FINAL MODEL – LOGISTIC REGRESSION

BALANCED CLASS WEIGHTS APPROACH

We are considering the Logistic Regression Model with Balanced Class Weights as our Final Model. The results are as follows :

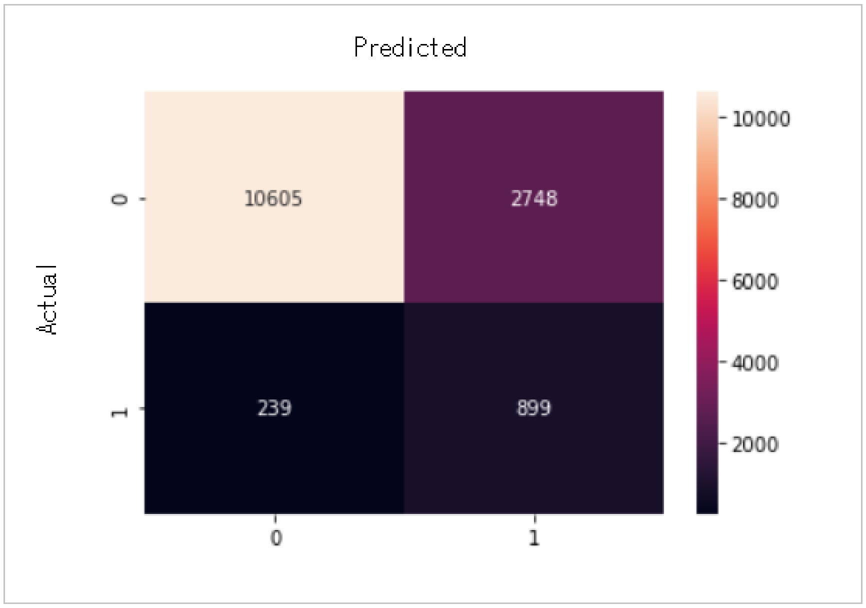
Logistic Regression – Balanced Class Weights

- **Classification Report :**

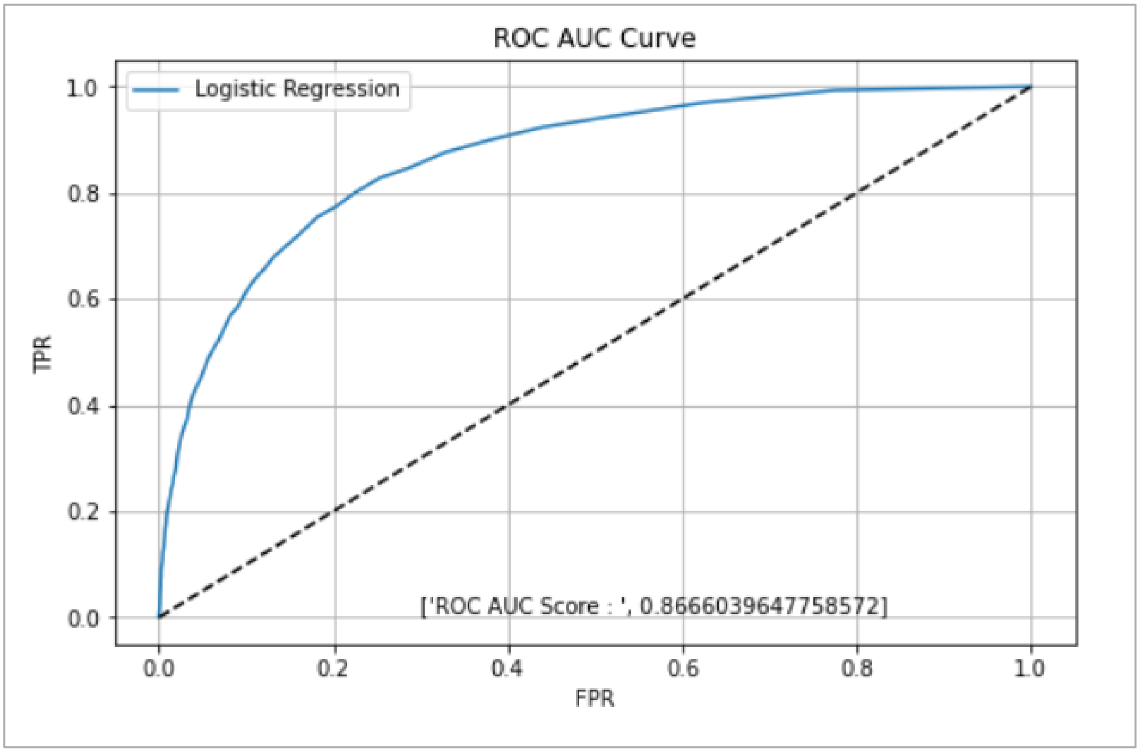


```
              precision    recall  f1-score   support

           0       0.98      0.79      0.88     13353
           1       0.25      0.79      0.38      1138

    accuracy                           0.79     14491
   macro avg       0.61      0.79      0.63     14491
weighted avg       0.92      0.79      0.84     14491
```

- **Confusion Matrix :**



- **ROC-AUC Curve :**



Here, AUC Score = 0.866

RESULTS & DISCUSSION

In our model, we are considering Recall Score as our main evaluation metric. Here, in the results also, we have focused on improving the Recall Score for the minority class i.e. hospital death reported. Because Recall Score takes into account the FN (False Negative) values. FNs are those cases which are in reality likely to die but our model predicts them as not likely to die. If the value of FNs is very high, our model can be dangerously misleading. If the value of FNs is high, this may lead the doctors and the medical staff to neglect the patients which are highly likely to die and need immediate attention. Thus, by focusing on improving the recall value, we are effectively avoiding the above error which is also termed as Type-II error.

Based on the value of Recall Score, we consider the above model as a good model.

Another concern in our model is that the precision value is not high enough which means that the False Positive (FP) are high in number. FPs are those cases which are not likely to die but our model predicts them as likely to die. If we consider the healthcare domain, there are many medical cases which are highly critical and the chances of survival are low, but due to the efforts of expert doctors and trained medical staff, many patients survive critical medical conditions. Maybe our model is not able to capture this fact. Due to this, the precision score might be low i.e. the cases were in reality not likely to survive looking at the independent variable values, but because of the efforts and close monitoring of the doctors and medical staff and timely medical intervention, many patients survived which lead to an increase in the value of FPs which in turn decreased the precision score.

IMPLICATIONS & SCOPE

Our model is having low Precision Score and high Recall Score. This means that the number of FPs is high and number of FNs is low i.e. our model is more likely to make Type-I error and very less likely to make Type-II error which is more riskier. That means our model is effectively reducing the risk of Type-II error which is highly misleading and riskier but it is prone to raising false alarms in cases of non-critical patients by classifying them as critical.

Our model is not a highly accurate model, but at the same time it is a safe model that can find application in the real world for patient survival prediction. Our model might not be as highly accurate as expected, but the predictions of our model when clubbed with medical expertise of doctors can definitely be helpful in prioritizing which cases are more severe in terms or survival ability and need immediate attention.

There are a few limitations of our model which include low Precision Score. This limitation can be improved upon in the future and if addressed rightly, can lead to significant improvement in the reliability of our model for patient survival prediction.

CLOSING REFLECTIONS

Throughout this process of Capstone Project right from the selection of topic to model building and evaluating the final model, we have got a glimpse into the real world of Data Science. We understood how real-world Data Science Projects are conducted and the multiple challenges that one might encounter during the different stages of the project.

By working on this project, we got an opportunity to apply the many concepts we have learnt in our course curriculum and understood which technique to use at which place. Throughout this project, we extensively used the Jupyter Notebook and are now very comfortable using it. We learnt the application of different classification algorithms in detail and how to tweak the parameters just enough to get improved results. We learnt to apply multiple approaches to the same problem statement and experiment with different techniques.

Also, the guidance from our mentor at every stage of the project helped us to overcome the many difficulties we faced as this was our first hands-on real-world Data Science Project. We are very thankful to our mentor Ms. Vibha Santhanam for guiding and helping us through all the stages of the project. We would also like to express our gratitude towards Great Learning for assigning us such knowledgeable mentor and giving us the chance to learn and grow into better students of Data Science though the medium of this project. Thank You.

REFERENCES

1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4999179/
2. https://towardsdatascience.com/how-to-create-a-simple-cancer-survival-prediction-model-with-eda-629dfa45d98b
3. https://theconversation.com/medical-ai-can-now-predict-survival-rates-but-its-not-ready-to-unleash-on-patients-127039
4. https://en.wikipedia.org/wiki/APACHE_II
5. https://www.nature.com/articles/s41598-020-73558-3
6. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0233678
7. https://pubmed.ncbi.nlm.nih.gov/11064783/
8. https://www.sciencedirect.com/science/article/pii/B9781437710151000461
9. https://towardsdatascience.com/stop-using-smote-to-handle-all-your-imbalanced-data-34403399d3be
10. https://towardsdatascience.com/the-right-way-of-using-smote-with-cross-validation-92a8d09d00c7
11. https://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_sample_weight.html