In Patient Survival Prediction

Submitted by Group 01 [Batch: Nov 2021]

Group Members:

Shiva sai kumar
Sparsh Singhal
Darshan GV
Vinod B

Research Supervisor

Mr. Srikar Muppidi

Submitted towards partial fulfilment of the criteria

for the award of PGP-DSE by Great Lakes Institute of Management

GREAT LAKES INSTITUTE OF MANAGEMENT

ABSTRACT

The project aims to develop and validate a prediction model for all-cause in-hospital mortality among admitted patients.ICU admission are useful for mortality prediction. Disease history can be used to differentiate mortality risk between patients with similar vital signs with more precision than SAPS II and APACHE II scores. Machine learning models can be deconvoluted to generate novel understandings of how ICU patient features from long-term and short-term events interact with each other. Explainable machine learning models are key in clinical settings, and our results emphasise how to progress towards the transformation of advanced models into actionable, transparent, and trustworthy clinical tools.

Techniques:

- Predictive Modelling
- Supervised Machine Learning Models
- Ensemble Techniques
- SMOTE Tomek

- Tools:
  - Python
  - IDE - Jupyter Notebook
- Domain:
  - Data Analytics
  - Health care

ACKNOWLEDGEMENT

Any endeavour in a specific field requires the guidance and support of many people for successful completion. The sense of achievement on completing anything remains incomplete if the people who were instrumental in its execution are not properly acknowledged. We would like to take this opportunity to verbalize our deepest sense of indebtedness to our project mentor, Mr. Muppidi Srikar, who was a constant pillar of support and continually provided us with valuable insights to improve upon our project and make it a success. Further, we would like to thank our parents for encouraging us and providing us a platform wherein we got an opportunity to design our own project.

Date:

Place:

CERTIFICATE OF COMPLETITION

This is to certify that the project titled **"In Patient Survival Prediction"** for case resolution was undertaken and completed under the supervision of Mr. Srikar Muppidi for Post Graduate Program in Data Science and Engineering (PGP – DSE)

Name of the Mentor:

Mr. Srikar Muppidi

Signature of the Mentor:

Date:

Place:

DECLARATION

We hereby declare, that this submission is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Date:

Place:

TABLE OF CONTENT

(I)

LIST OF ABBREVIATIONS

S. No.	Abbreviation	Meaning
1.	CRISP-DM	cross-industry process for data mining
2.	GCS	Glasgow Coma Scale
3.	SP02	Oxygen Saturation
4.	SMOTE-Tomek	Synthetic Minority Oversampling Technique Tomek lines
5.	VIF	Variance influence factor
6.	RFE	Recursive factor elimination
7.	PCA	Principal component analysis
8.	ADABoost	Adaptive boost
9.	GBM	Gradient boosting machine
10.	XG Boost	Extreme Gradient Boosting
11.	PCA	Principal Component Analysis

(II)

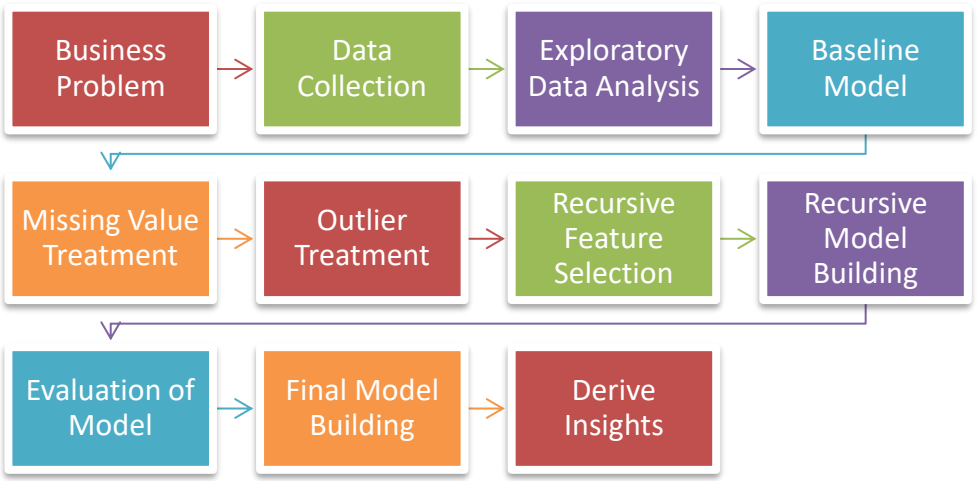
EXECUTIVE SUMMARY

The objective of the project is to predict whether a patient survives when admitted in an hospital with any health condition or in an emergency case. Dataset receives data submissions from intensive care units (ICUs) throughout multiple hospitals in different ethnicity, biochemical, physiological and demographic information required for the calculation of severity and diagnosis scores, together some data on therapies received during the ICU stay and information about patient outcomes.

First a base line model was performed using Logistic Regression with using one hot encoding and ordinal encoding for the categorical columns, which has given us a Training accuracy of 92% and Test accuracy of 92.7%. Later Data standardization was done, such as Outlier treatment, missing values treatment, checking for skewness, dropping columns based on correlation and feature selection using statistical tests that are chi-square and mutual information test for categorical variables, annova for numerical variables. Multiple imputation was used for the treatment of the missing values in the data, capping was used for the outlier treatment on individual variables based on the skewness and domain range.

For the final model Logistic regression with XG boost and hyper parameter tuning the accuracy was reduced but the target recall has been able to improve with a F-Beta score of 0.55 and accuracy was 87.24 %, for the Decision tree with Hyper parameter tuning.

CHAPTER 1
PROJECT OVERFLOW



The process we chose to work on is an iterative process.

- We define the business problem.
- We collect data from different relevant sources.
- In Exploratory Data Analysis, we try to understand the structure of the data, the nature of the variables. The information present in the data. Statistical summary of the variables. A special attention is paid to the target variable distribution
- We build baseline models to get the idea of the minimum amount of performance that we can get from each of the models we have decided to use in the future.
- The first step of data preparation is Missing Value Treatment which means that If we have missing values in our data, then according to the percentage of the missing values in a row or column, we decide whether to drop a row or a column or replace those values with one of the central tendency measures as studied in statistics. Any method that will cause minimum impact on the data so we can get the actual pattern within the data for our models to learn.
- The other step of data preparation is Outlier Treatment. It is important to treat the outliers as they can affect the mean of the data in question and our evaluation of the data could be way off than the actual values.
- Recursive Feature Selection is a method to select features which have the maximum impact on the metrics of the models. The features which don't influence the performance of the model much are dropped.
- Recursive model building is one of the key highlights of our methodology as we are going to build models, evaluate their performance, find out the features or hyper parameters behind the performance of the model and then fine tune the possible causes and make our model better up to a certain threshold which in real world will be decided by the actual business requirements.

- Final Model building will be done when we are equipped with the information as to what drives the performance of the models and what we want from our model, such as, interpretability, explain ability, performance etc.
- We can also derive insights from the selected features, like, which features to focus on the most in the future.

1.1 PROCESS OVERVIEW:

A stepwise process has been followed in implementation of the supervised learner.

1. Once the data is ready, it has been imported as a pandas data frame and the structure of the data is noted.
2. A recursive approach has been followed while checking the data for missing values, variable data types and outliers.
3. Exploratory data analysis has been performed including the univariate and bivariate analysis for the continuous variables and for the categorical variables as far as possible.
4. Data has been visualized using different charts including distance plots, histograms, bar charts and box plots. Five-point summary statistics and correlation of data is also noted.
5. The target variable has been completely analysed and checked for any impbalance in the data and observed that the target variable is moderately imbalanced.
6. Baseline models have been designed using 2 approaches – first by completely dropping the missing values which is a biased approach ,second scaled the numerical columns .An accuracy of 92% is observed.
7. Based on conclusions on baseline model, different Missing Value Imputation and Outlier Treatment techniques are used to prepare data for application of models.
8. Based on conclusions on baseline model, different Missing Value Imputation such as KNN imputation or MICE and Outlier Treatment techniques such as Winsorization and capping is used to prepare data for application of models in order to preserve the atomicity and authenticity of the data. In the view that data in health predictions plays a significant role.
9. After processing the data, Chi-square , Mutual information and Anova tests has been performed and important features have been identified using Recursive Feature Engineering, feature importance, p-values, variation inflation factor.
10. Accordingly apache_4a_hospital_death_prob ,apache_4a_icu_death_prob ,d1_temp_min,d1_sysbp_min,d1_mbp_min,d1_diasbp_min,temp_apache,d1_heartrate_max,h1_sysbp_min,h1_mbp_min,h1_diasbp_min,h1_resprate_max , h1_resprate_min ,age ,d1_potassium_max,h1_heartrate_max,heart_rate_apache,d1_resprate_max,apache_3j_diagnosis,apache_2_diagnosis are the important features .
11. Multiple classification models including Logistic Regression, Decision Tress, Random Forest, K-Neighbours, and AdaBoost Classifier etc. are implemented recursively in the process to compare the results of each.
12. Finally models have been cross validated using K-fold Cross Validation so as to learn a highly sensitive classifier which maps or classifies the image data to ‘Ad’ or ‘Non-Ad’ with high accuracy.

CHAPTER -2
INTRODUCTION

Intensive-care units (ICUs) treat the most critically ill patients, which is complicated by the heterogeneity of the diseases that they encounter. Severity scores based mainly on acute physiology measures collected at ICU admission are used to predict mortality, but are non-specific, and predictions for individual patients can be inaccurate. We investigated whether inclusion of long-term disease history before ICU admission improves mortality predictions.

According to WHO, Heart Diseases and cancer are a leading cause of death worldwide. It is quite difficult to identify the cardiovascular disease (CVD) and cancer because of some contributory factors which contribute to CVD like high blood pressure, cholesterol level, diabetics, abnormal pulse rate, and many other factors.

Researchers have been exploring a wide range of techniques to predict patient survival but the disease prediction at an early stage is not very efficient due to many factors, including but not limited to complexity, execution time, and accuracy of the approach. As such, proper treatment and diagnosis can save many lives.

2.1 Overview of the Dataset

The Patient survival Dataset receives data submissions from intensive care units (ICUs) throughout patients admitted in multiple hospitals in Accident & Emergency/Surgery/ICU care/other.

Records available are patients from different ethnicity and their previous health diagnosis records with current diagnosis in real time are provided with different diagnosis techniques and treatment methods.

Submitted Data includes biochemical, physiological and demographic information required for the calculation of severity and diagnosis scores, together some data on therapies received during the ICU stay and information about patient outcomes.

Data source url : https://www.kaggle.com/datasets/mitishaagarwal/patient/code

The data consists of 91713 rows and 85 columns. Out of these we have 7 categorical columns and 78 Numerical columns.

Column Name	Data type	Description
encounter_id	int	Identifier associated with patient's unit stay

patient_id	int	Unique identifier associated with a patient
hospital_id	int	Unique identifier associated with a hospital
age	float	The age of the patient on unit admission
bmi	float	The body mass index of the person on unit admission
elective_surgery	Int	Whether the patient was admitted to the hospital for an elective surgical operation
ethnicity	object	The common national or cultural tradition which the person belongs to
gender	object	Sex of the patient
height	Float	The height of the person on unit admission
icu_admit_source	object	The location of the patient prior to being admitted to the unit
icu_id	Int	A unique identifier for the unit to which the patient was admitted
icu_stay_type	Object	
icu_type	Object	A classification which indicates the type of care the unit is capable of providing
pre_icu_los_days	Float	The length of stay of the patient between hospital admission and unit admission
weight	Float	The weight (body mass) of the person on unit admission
apache_2_diagnosis	Float	The APACHE II diagnosis for the ICU admission.APACHE II score is a general measure of disease severity based on current physiologic measurements, age & previous health conditions. The score can help in the assessment of patients to determine the level & degree of diagnostic & therapeutic intervention.
apache_3j_diagnosis	Float	APACHE III discriminates poorly between survivors and non-survivors of patients admitted to the ICU after OLT. Though APACHE III has been shown to be valid in heterogenous populations and in certain groups of patients with specific diagnoses, it should be used with caution – if used at all – in recipients

		of liver transplantation.
apache_post_operative	Int	The APACHE operative status; 1 for post-operative, 0 for non-operative
arf_apache	Float	Whether the patient had acute renal failure during the first 24 hours of their unit stay, defined as a 24 hour urine output <410ml, creatinine >=133 micromol/L and no chronic dialysis
gcs_eyes_apache	Float	The eye opening component of the Glasgow Coma Scale measured during the first 24 hours which results in the highest APACHE III score
gcs_motor_apache	Float	The motor component of the Glasgow Coma Scale measured during the first 24 hours which results in the highest APACHE III score
gcs_unable_apache	Float	Whether the Glasgow Coma Scale was unable to be assessed due to patient sedation
gcs_verbal_apache	Float	The verbal component of the Glasgow Coma Scale measured during the first 24 hours which results in the highest APACHE III score
heart_rate_apache	Float	The heart rate measured during the first 24 hours which results in the highest APACHE III score
intubated_apache	Float	Whether the patient was intubated at the time of the highest scoring arterial blood gas used in the oxygenation score
map_apache	Float	The mean arterial pressure measured during the first 24 hours which results in the highest APACHE III score
resprate_apache	Float	The respiratory rate measured during the first 24 hours which results in the highest APACHE III score
temp_apache	Float	The temperature measured during the first 24 hours which results in the highest APACHE III score
ventilated_apache	Float	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

d1_diasbp_max	Float	The patient's highest diastolic blood pressure during the first 24 hours of their unit stay, either non-invasively or invasively measured
d1_diasbp_min	Float	The patient's lowest diastolic blood pressure during the first 24 hours of their unit stay, either non-invasively or invasively measured
d1_diasbp_noninvasive_max	Float	The patient's highest diastolic blood pressure during the first 24 hours of their unit stay, non-invasively measured
d1_diasbp_noninvasive_min	Float	The patient's lowest diastolic blood pressure during the first 24 hours of their unit stay, non-invasively measured
d1_heartrate_max	Float	The patient's highest heart rate during the first 24 hours of their unit stay
d1_heartrate_min	Float	The patient's lowest heart rate during the first 24 hours of their unit stay
d1_mbp_max	Float	The patient's highest mean blood pressure during the first 24 hours of their unit stay, either non-invasively or invasively measured
d1_mbp_min	Float	The patient's lowest mean blood pressure during the first 24 hours of their unit stay, either non-invasively or invasively measured
d1_mbp_noninvasive_max	Float	The patient's highest mean blood pressure during the first 24 hours of their unit stay, non-invasively measured
d1_mbp_noninvasive_min	Float	The patient's lowest mean blood pressure during the first 24 hours of their unit stay, non-invasively measured
d1_resprate_max	Float	The patient's highest respiratory rate during the first 24 hours of their unit stay
d1_resprate_min	Float	The patient's lowest respiratory rate during the first 24 hours of their unit stay
d1_spo2_max	Float	The patient's highest peripheral oxygen saturation during the first 24 hours of their unit stay
d1_spo2_min	Float	The patient's lowest peripheral oxygen saturation during the first 24 hours of their unit stay

d1_sysbp_max	Float	The patient's highest systolic blood pressure during the first 24 hours of their unit stay, either non-invasively or invasively measured
d1_sysbp_min	Float	The patient's lowest systolic blood pressure during the first 24 hours of their unit stay, either non-invasively or invasively measured
d1_sysbp_noninvasive_max	Float	The patient's highest systolic blood pressure during the first 24 hours of their unit stay, invasively measured
d1_sysbp_noninvasive_min	Float	The patient's lowest systolic blood pressure during the first 24 hours of their unit stay, invasively measured
d1_temp_max	Float	The patient's highest core temperature during the first 24 hours of their unit stay, invasively measured
d1_temp_min	Float	The patient's lowest core temperature during the first 24 hours of their unit stay
h1_diasbp_max	Float	The patient's highest diastolic blood pressure during the first hour of their unit stay, either non-invasively or invasively measured
h1_diasbp_min	Float	The patient's lowest diastolic blood pressure during the first hour of their unit stay, either non-invasively or invasively measured
h1_diasbp_noninvasive_max	Float	The patient's highest diastolic blood pressure during the first hour of their unit stay, invasively measured
h1_diasbp_noninvasive_min	Float	The patient's lowest diastolic blood pressure during the first hour of their unit stay, invasively measured
h1_heartrate_max	Float	The patient's highest heart rate during the first hour of their unit stay
h1_heartrate_min	Float	The patient's lowest heart rate during the first hour of their unit stay
h1_mbp_max	Float	The patient's highest mean blood pressure during the first hour of their unit stay, either non-invasively or invasively measured
h1_mbp_min	Float	The patient's lowest mean blood pressure during the first hour of their unit stay, either non-

		invasively or invasively measured
h1_mbp_noninvasive_max	Float	The patient's highest mean blood pressure during the first hour of their unit stay, non-invasively measured
h1_mbp_noninvasive_min	Float	The patient's lowest mean blood pressure during the first hour of their unit stay, non-invasively measured
h1_resprate_max	Float	The patient's highest respiratory rate during the first hour of their unit stay
h1_resprate_min	Float	The patient's lowest respiratory rate during the first hour of their unit stay
h1_spo2_max	Float	The patient's highest peripheral oxygen saturation during the first hour of their unit stay
h1_spo2_min	Float	The patient's lowest peripheral oxygen saturation during the first hour of their unit stay
h1_sysbp_max	Float	The patient's highest systolic blood pressure during the first hour of their unit stay, either non-invasively or invasively measured
h1_sysbp_min	Float	The patient's lowest systolic blood pressure during the first hour of their unit stay, either non-invasively or invasively measured
h1_sysbp_noninvasive_max	Float	The patient's highest systolic blood pressure during the first hour of their unit stay, non-invasively measured
h1_sysbp_noninvasive_min	Float	The patient's lowest systolic blood pressure during the first hour of their unit stay, non-invasively measured
d1_glucose_max	Float	The highest glucose concentration of the patient in their serum or plasma during the first 24 hours of their unit stay
d1_glucose_min	Float	The lowest glucose concentration of the patient in their serum or plasma during the first 24 hours of their unit stay
d1_potassium_max	Float	The highest potassium concentration for the patient in their serum or plasma during the first 24 hours of their unit stay

d1_potassium_min	Float	The lowest potassium concentration for the patient in their serum or plasma during the first 24 hours of their unit stay
apache_4a_hospital_death_prob	Float	The APACHE IVa probabilistic prediction of in-hospital mortality for the patient which utilizes the APACHE III score and other covariates, including diagnosis.
apache_4a_icu_death_prob	Float	The APACHE IVa probabilistic prediction of in ICU mortality for the patient which utilizes the APACHE III score and other covariates, including diagnosis
aids	Float	Whether the patient has a definitive diagnosis of acquired immune deficiency syndrome (AIDS) (not HIV positive alone)
cirrhosis	Float	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.
diabetes_mellitus	Float	Whether the patient has been diagnosed with diabetes, either juvenile or adult onset, which requires medication.
hepatic_failure	Float	Whether the patient has cirrhosis and additional complications including jaundice and ascites, upper GI bleeding, hepatic encephalopathy, or coma.
immunosuppression	Float	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).
leukemia	Float	Whether the patient has been diagnosed with acute or chronic myelogenous leukemia, acute or chronic lymphocytic leukemia, or multiple myeloma.
lymphoma	Float	Whether the patient has been diagnosed with non-Hodgkin lymphoma.

solid_tumor_with_metastasis	Float	Whether the patient has been diagnosed with any solid tumor carcinoma (including malignant melanoma) which has evidence of metastasis.
apache_3j_bodysystem	Object	Admission diagnosis group for APACHE III
apache_2_bodysystem	Object	Admission diagnosis group for APACHE II
hospital_death	Int	Whether the patient died during this hospitalization

Table 1: Dataset variables

2.2 Problem Statement :

The predictors of in-hospital mortality for admitted patients remain poorly characterized. We aimed to develop and validate a prediction model for all-cause in-hospital mortality among admitted patients.

ICU admission are useful for mortality prediction. Disease history can be used to differentiate mortality risk between patients with similar vital signs with more precision than SAPS II and APACHE II scores. Machine learning models can be deconvoluted to generate novel understandings of how ICU patient features from long-term and short-term events interact with each other. Explainable machine learning models are key in clinical settings, and our results emphasise how to progress towards the transformation of advanced models into actionable, transparent, and trustworthy clinical tools.

Mortality risk estimates based on acute physiology scores—such as the Simplified Acute Physiology Score (SAPS) and the Acute Physiologic Assessment and Chronic Health Evaluation (APACHE)—are sometimes used in clinical practice to assess disease severity

Our study shows the importance of previous disease history in predictions of mortality in ICU patients and thus, the importance of these data in clinical decision making. The predictive value of long-term disease history was stable over time compared with that of physiology measures, is independent of ICU care, and can be made available at admission to the ICU.

CHAPTER-3
VISUALIZATION

3.1 Target variable :

The target dependent variable for above data set is ‘hospital_death’ where we are going to predict whether the patient will survive after admitting into the hospital.

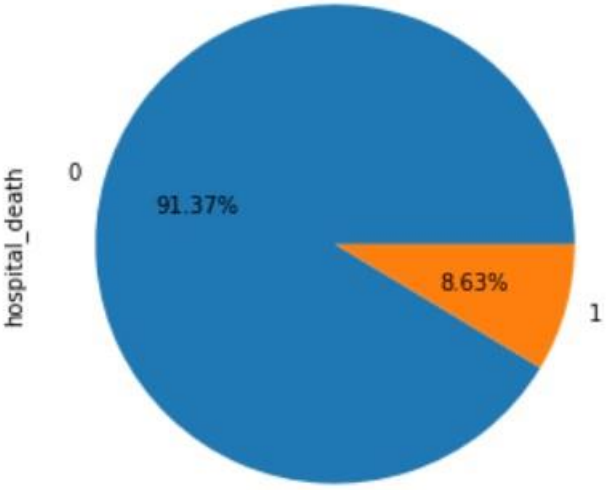
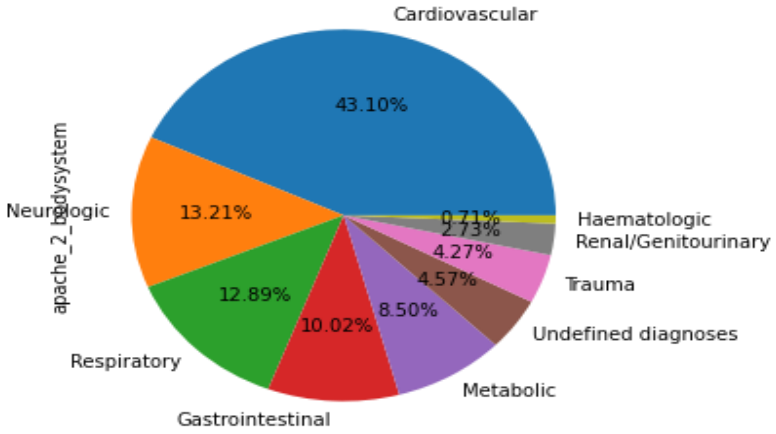


Figure 1

In the above data set for the target variable ‘hospital_death’ shows ‘1’ as patient dies during the hospitalization and ‘0’ as patient didn’t die during the hospitalization.
It implies that 91.37 % of patients didn’t die during the hospitalization and 8.63% of patients died during the hospitalization. **We observe that there is there is presence of moderate amount of class imbalance**.

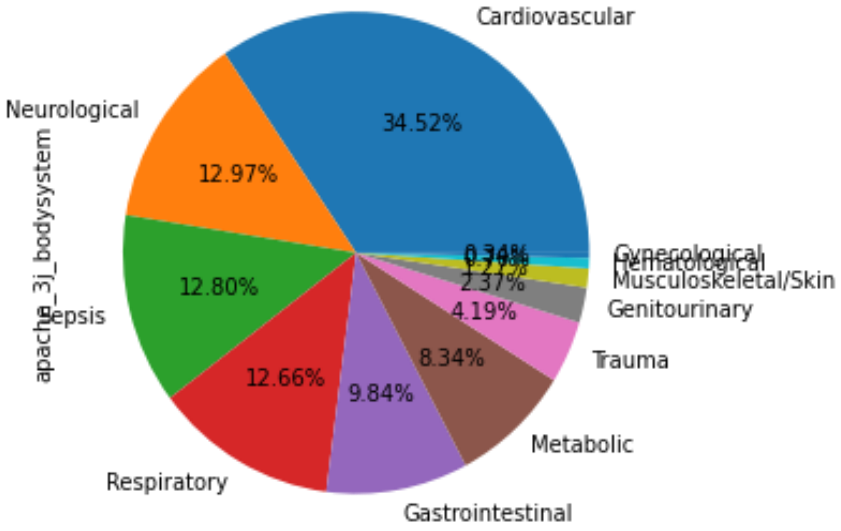
3.2 Univariate Analysis

3.2.1 Df-Apache-System :



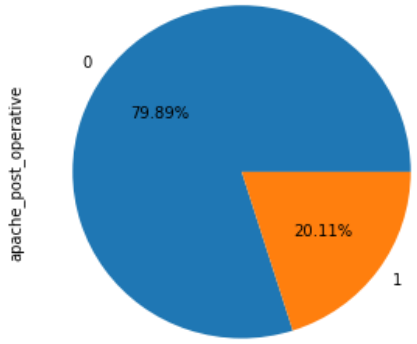
- The Major portion is on cardiovascular category
- Hence more direct relationship is related with heart rate ,bloodpressure, Spo2

3.2.2 Apache 3j Body system :



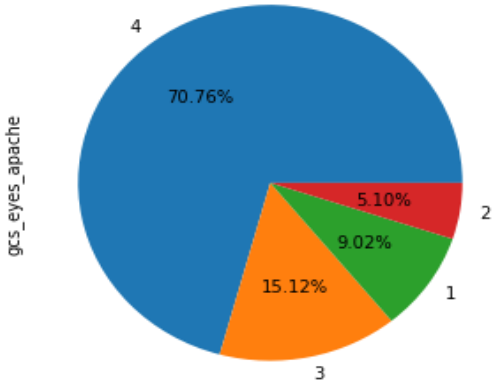
- The Major portion is on cardiovascular category again on apache 3j.
- Hence more direct relationship is related with heart rate ,bloodpressure, Spo2

3.2.3 Apache Post Operative :



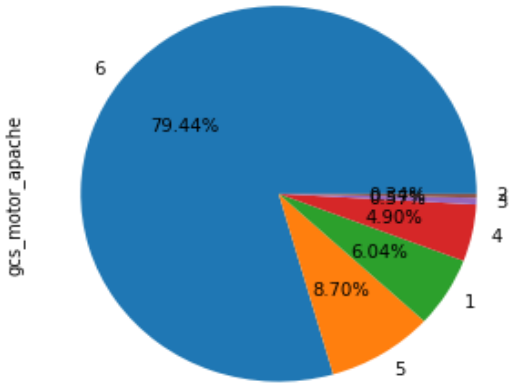
- Data on apache post operative is not much imbalanced percent is more than 5%.

3.2.4 GCS-EYE-Apache:



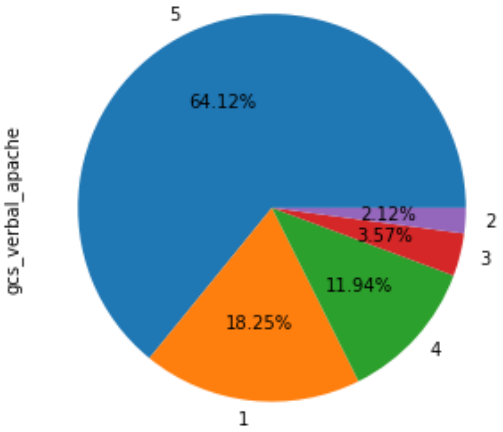
- Critical Value is being taken for about 70%
- Hence it can give a conclusions on criticality on deaths and no deaths.

3.2.5 GCS-Motor-Apache:



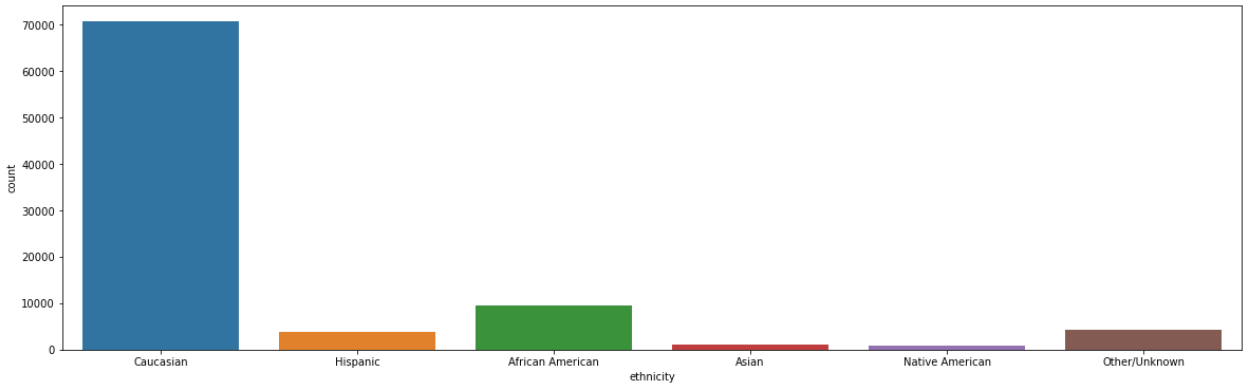
- Critical Value of 6 is being takenby major portions
- Around 80% share is from crtical value 6.

3.2.6 GCS-Verbal-Apache:



- Critical Value of 5 is majorly taken
- Hence a relationship can be inferred for deaths

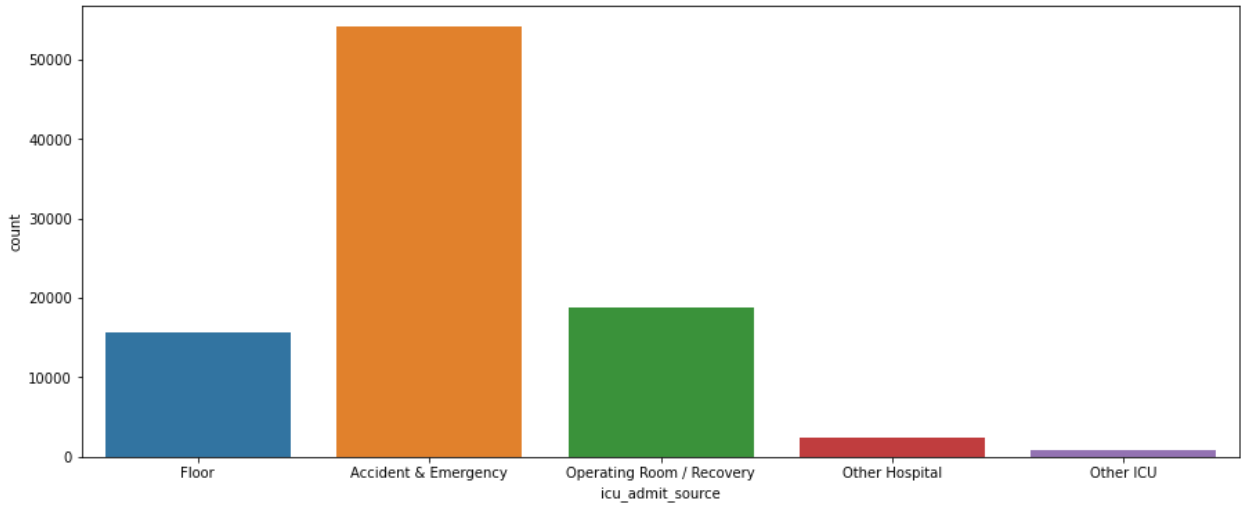
3.2.7 Ethnicity :



Since Ethnicity is one category that might have effect on diagnosis differently.

But major data collected is Caucasian and others are few a prime conclusion is vague and hence ethnicity is not give major priority.

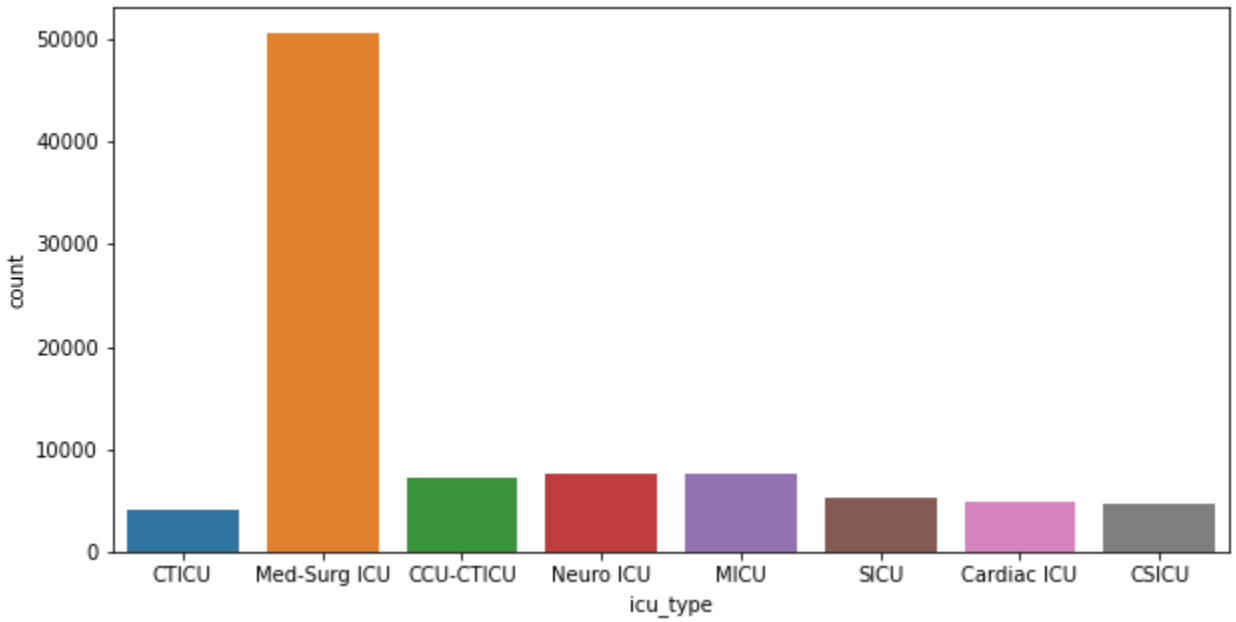
3.2.8 ICU-Admit_Source:



It is clear from the data that major admitted subjects belong to Accident and emergency

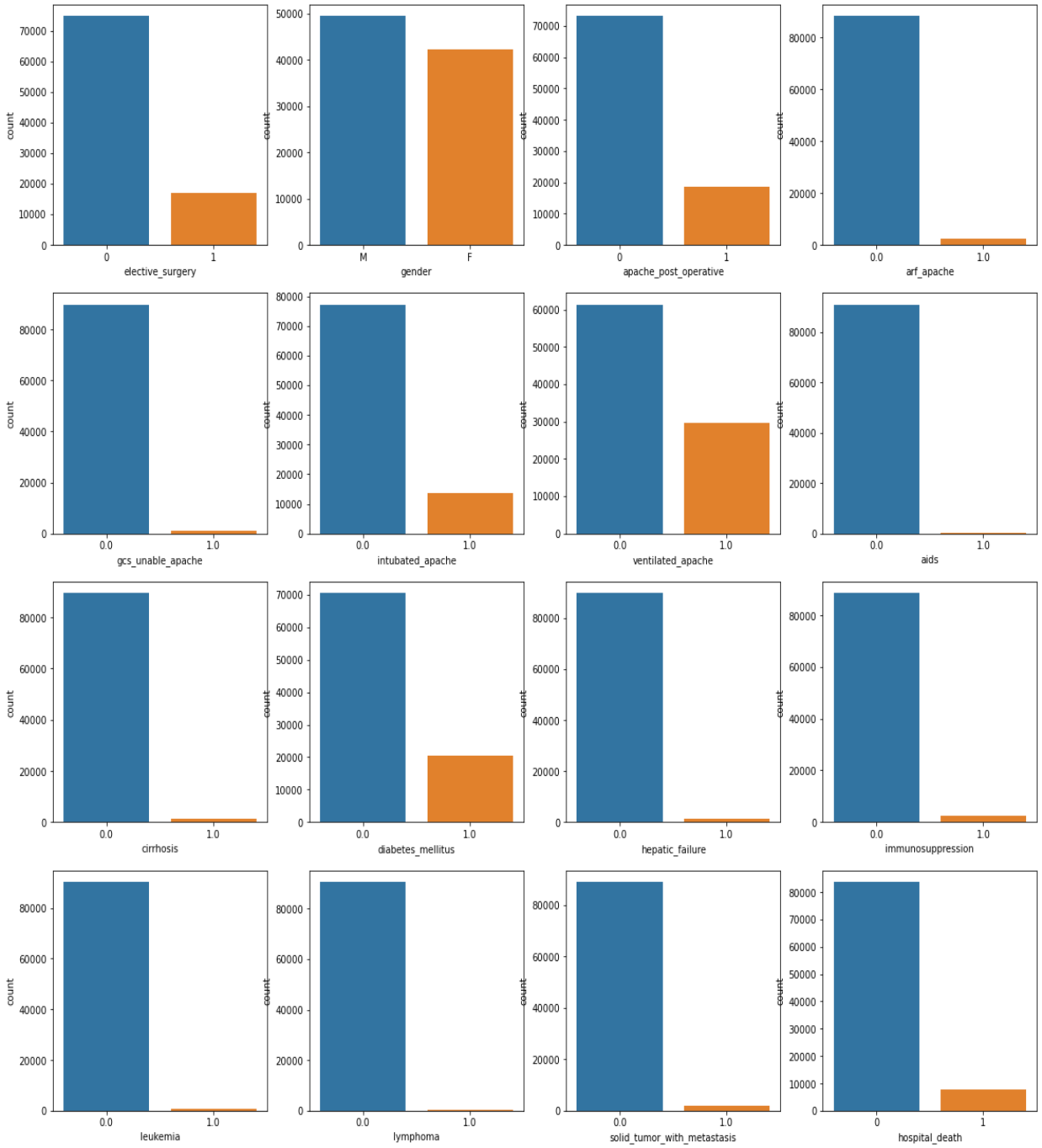
It is the critical scenario for patients survival.

3.2.9 ICU-Type :



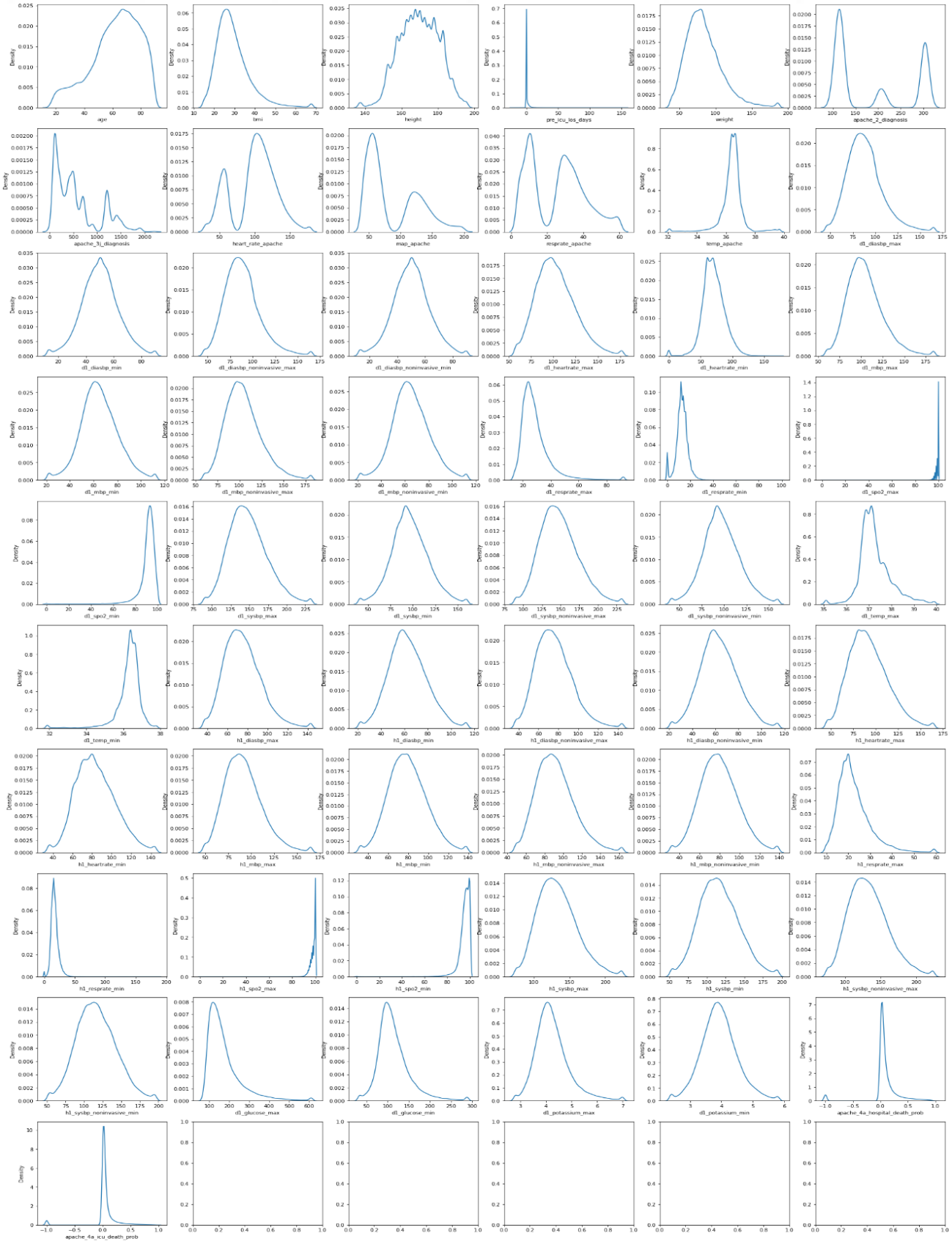
From the previous pie-chart it was clear that heart related ailments had a major reasons for deaths. However to be much clearer Medical -Surgery has an impact.

3.2.10 Categorical Variables count:



3.2.11 KDE plot for Numerical vriables :

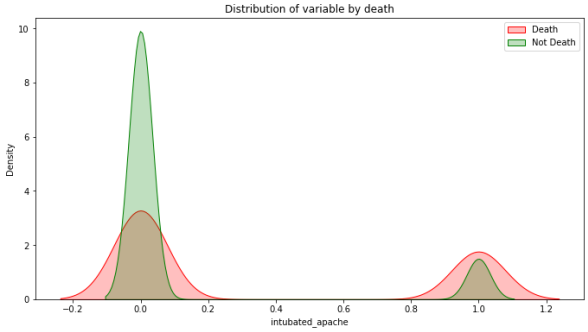
To know the distribution of numerical variables a **KDE_plot** is plotted in one single frame .

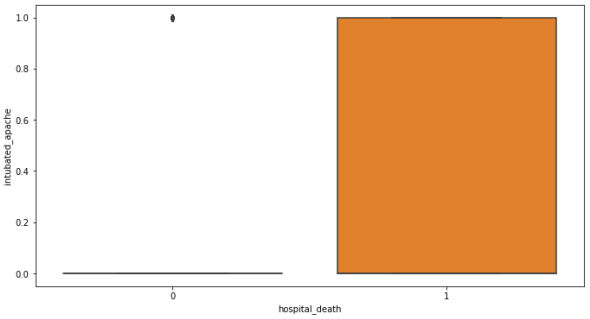


Many of the variables are close to normal distribution. However few must undergo with basic fit transformation.

3.3 Bi-Variate Analysis

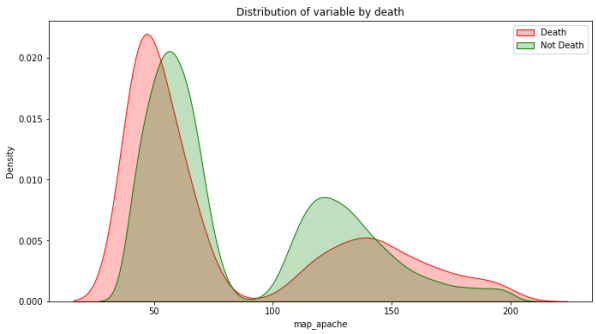
3.3.1 Intubated-Apache vs Hospital-Death :

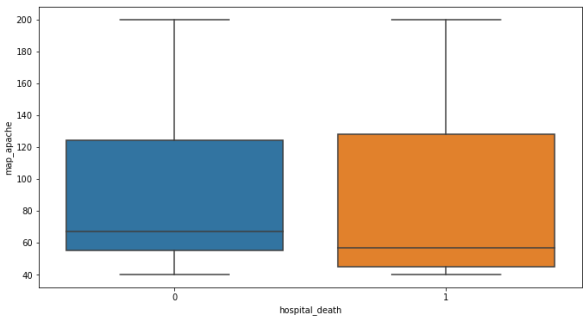




- It has a hospital death covering the no-deaths
- Apache has an effect on death

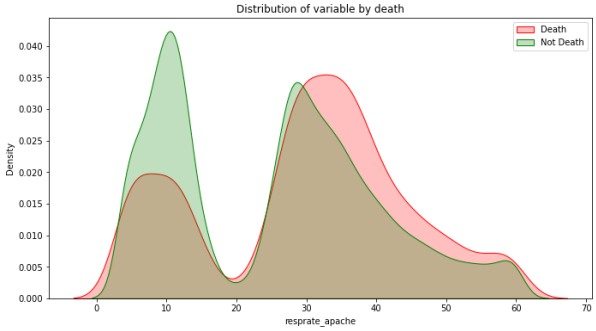
3.3.2 The mean arterial pressure:

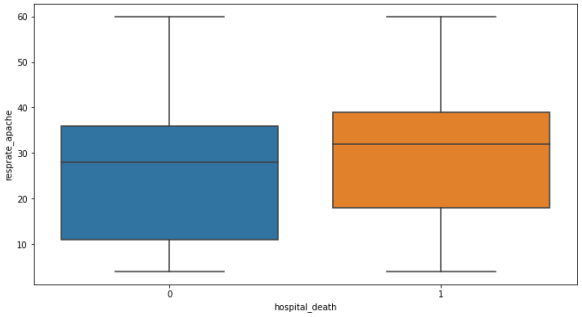




- It has a hospital death almost as same as no deaths
- MAP has little effect on death ,as the distribution is not normal correct conclusions cannot be made

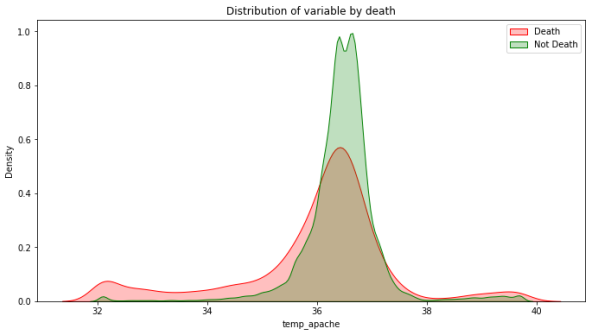
3.3.3 Resperate_Apache :

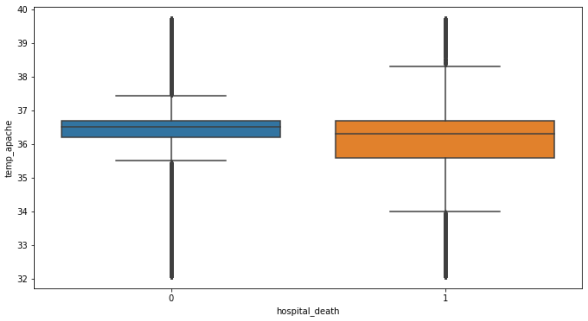




- Resperate Apache signifies that death has just more a bit than non resperate apache
- The distribution is not normal and it is a merger of two normal ranges
- Since deaths has effect on respiration prime conclusion cannot be drawn

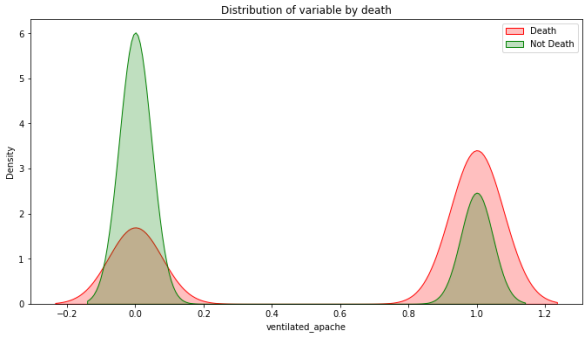
3.3.4 Temp Apache :

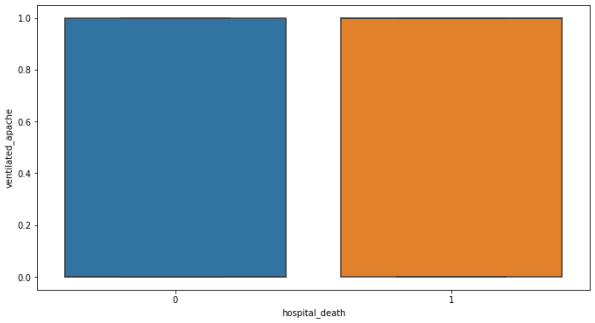




- The distribution is between 34 to 38 and it is a normal range.
- So deaths has less related to temp

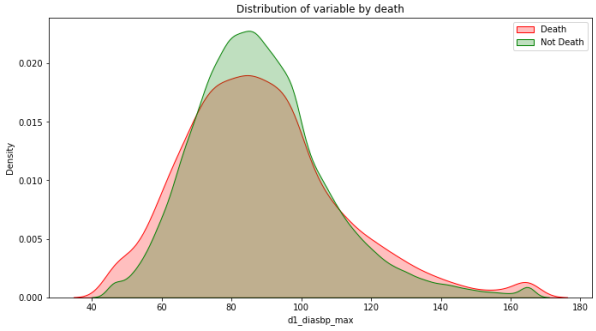
3.3.5 Ventilated Apache:

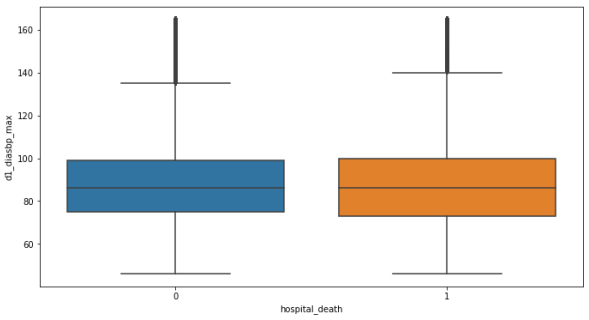




- Deaths has covered the alive .
- No proper relation can be drawn.

3.3.6 D1_Diaspb Max: Diastolic



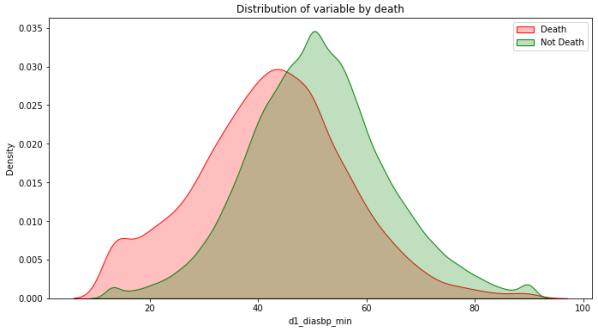


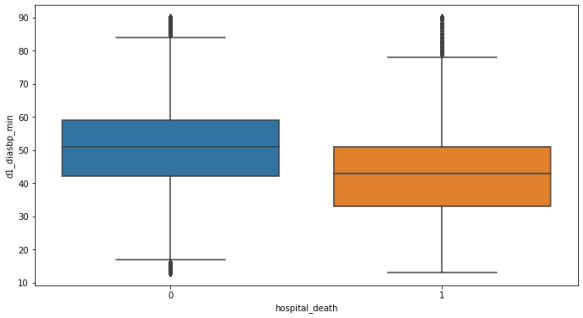
3.3.7 blood pressure :

Range is 40 to 140

Deaths are almost same as no-deaths

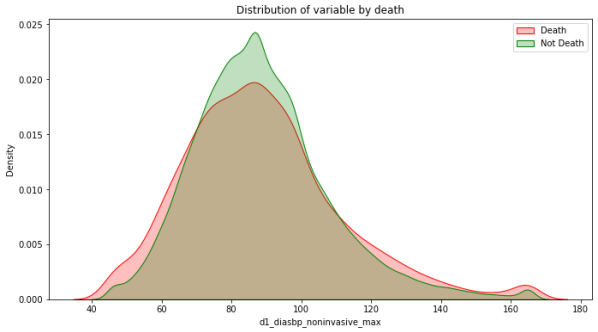
3.3.8 D1_Diaspb Min: Diastolic blood pressure :

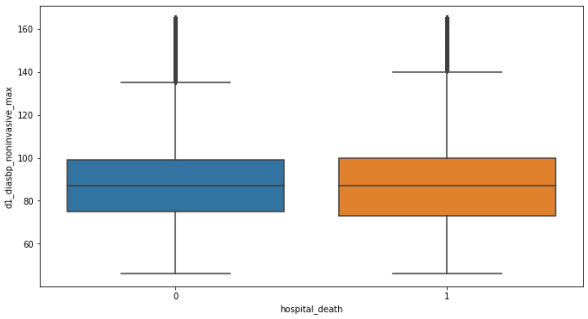




- Range is 20 to 80
- Deaths are lesser than no-deaths.

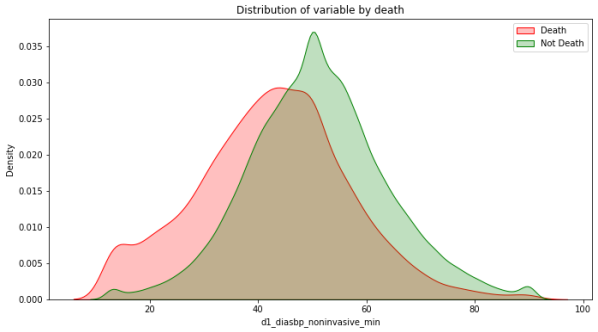
3.3.9 D1_diasbp_noninvasive_max:

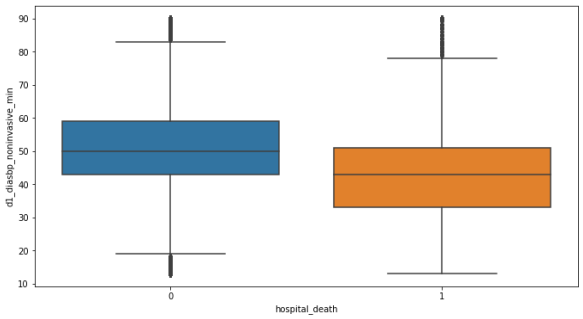




- Range is 40 to 140
- Deaths are same as no-deaths

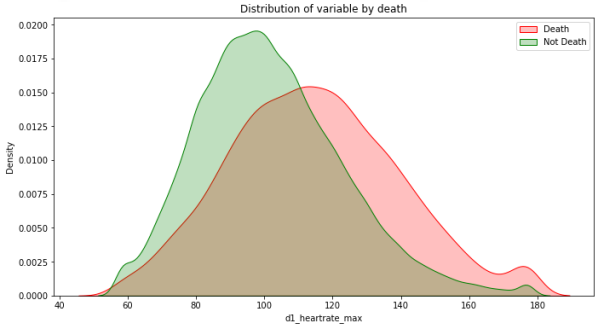
3.3.10 D1_diasbp_noninvasive_min :

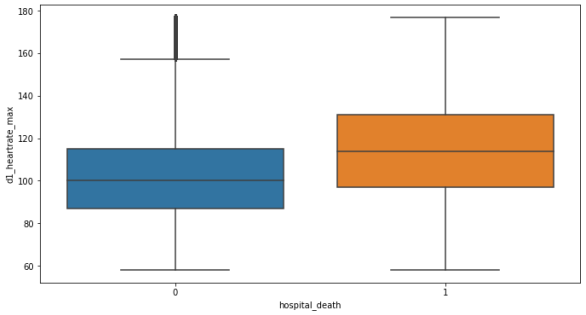




- Range is 40 to 140
- Deaths are less than no-deaths

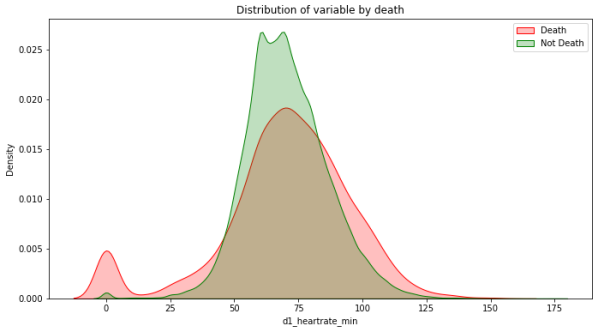
3.3.11 D1_heartrate_max:

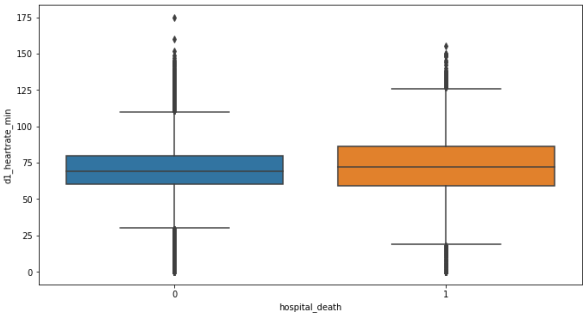




- Range is 60 to 160
- Deaths are more than no-deaths

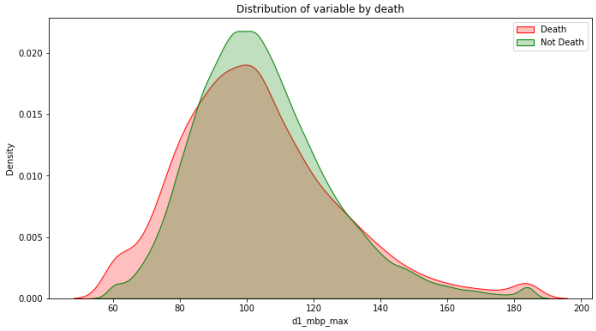
3.3.12 D1_heartrate_min:

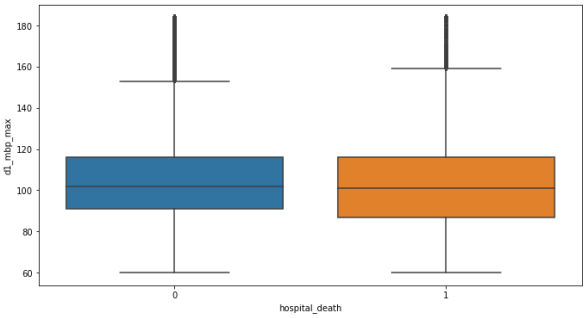




- Range is 25 to 100
- Deaths has same effect as no deaths

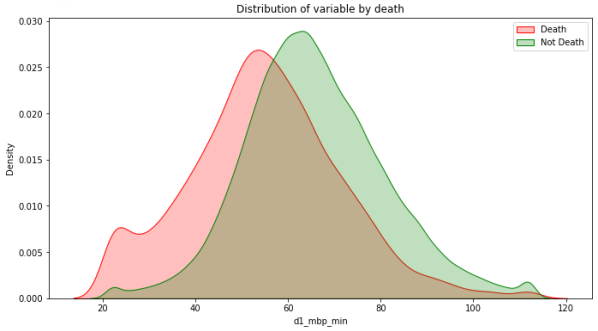
3.3.13 D1_mbp_max :

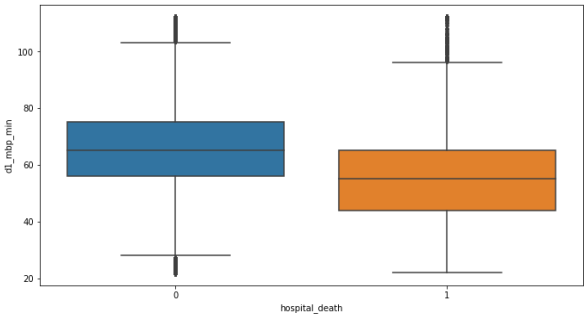




- Range is 60 to 160
- Deaths has same effects as no-deaths

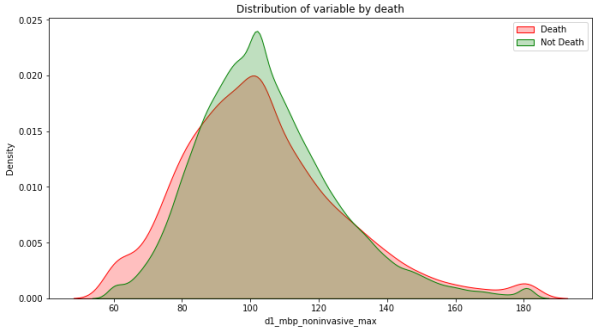
3.3.14 D1_mbp_min :

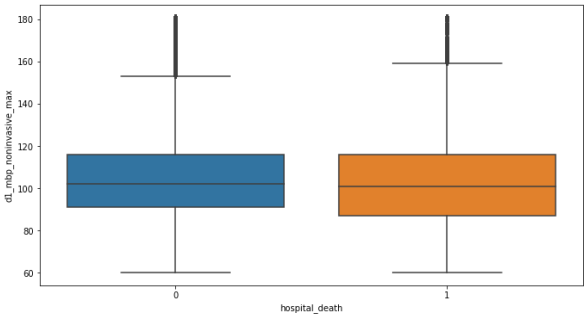




- Range Is 40 to 100
- Deaths has minimal effect as no deaths is more than deaths.

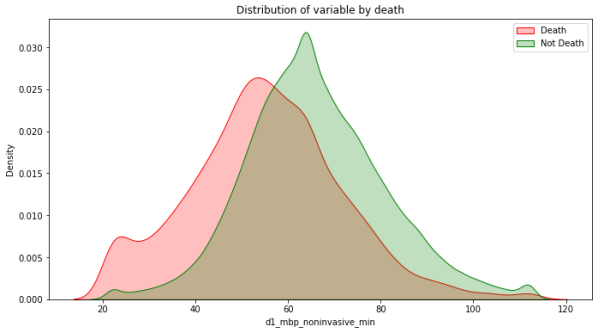
3.3.15 D1_mpb_noninvasive max :

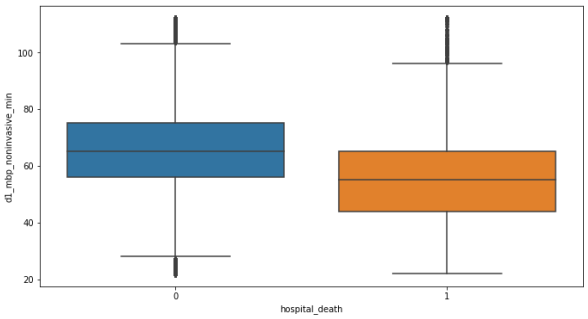




- Range is 60 to 160
- Deaths has same effect as no deaths
-

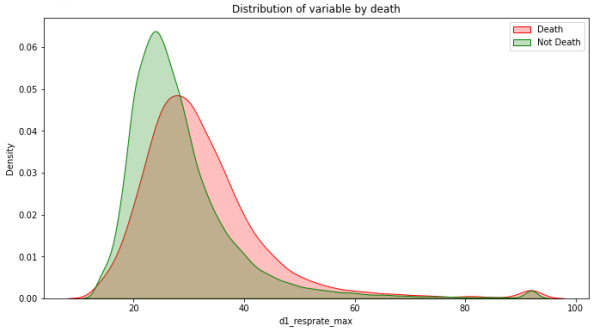
3.3.16 D1_mpb_noninvasive min :

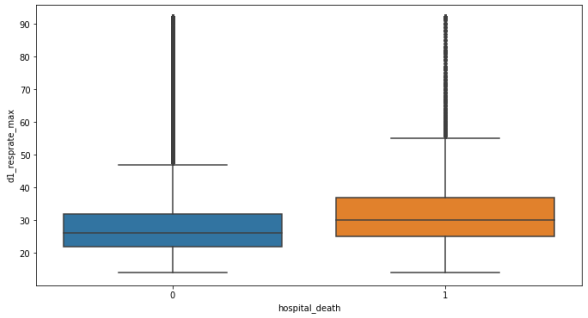




- Range is 40 to 100
- Deaths has less effect as no deaths

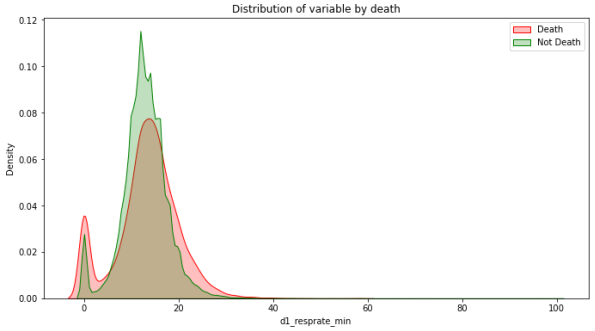
3.3.17 D1-Resperate Max:

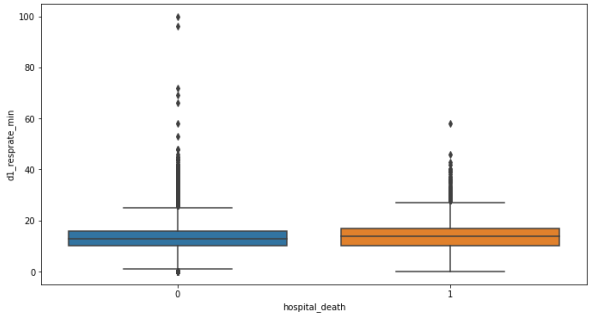




- Range is 20 to 60 . Deaths has little effect as no deaths

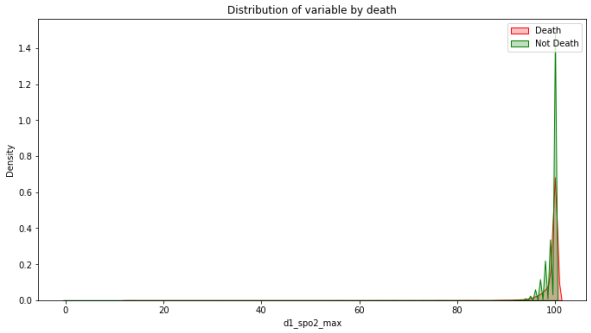
3.3.18 D1-Resperate Min:

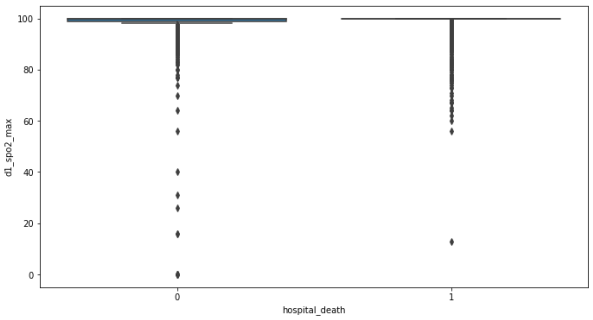




- Range is around 15 Deaths has a little effect as no deaths

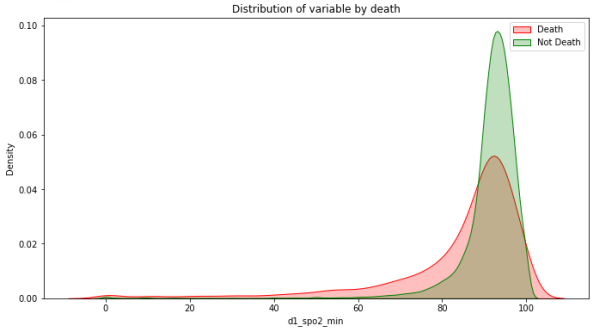
3.3.19 D1_spo2_max:

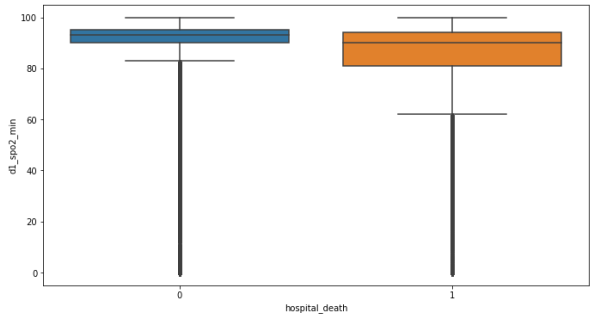




- This is right tight skewed.
- As the normal healthy person must have spo2 of >95%
- Box-Plot cannot be inferred to check for deaths.
- As its tightly skewed

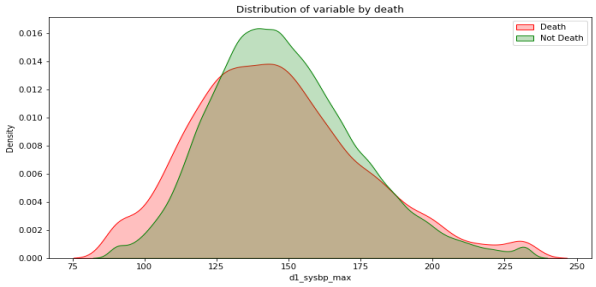
3.3.20 D1_spo2_min:

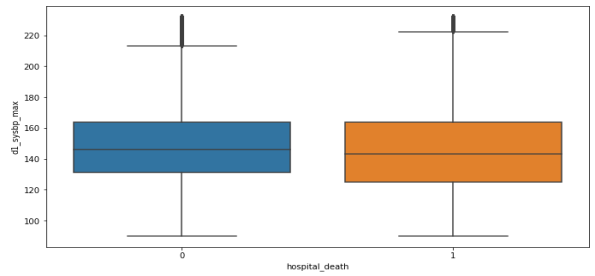




- The min spo2 is clear and hence there are deaths as spo2 has direct relationship
- As the range went lower and lower below 95% the deaths has increased
- The same can be inferred from the box-plot

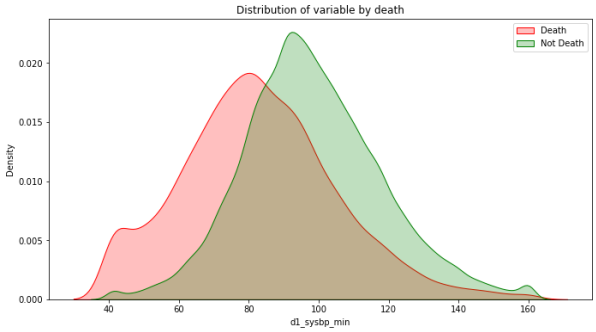
3.3.21 D1_syspb_max :

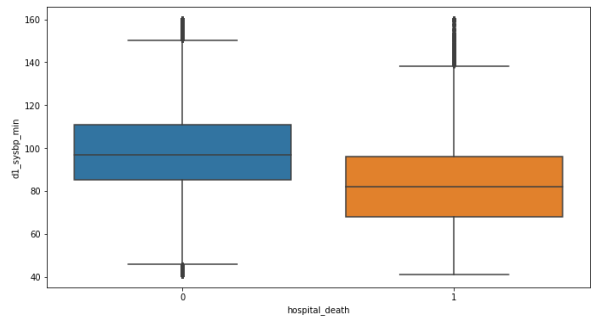




- Deaths are more compared to no deaths
- Systolic blood pressure has effect on the target that is deaths.

3.3.22 D1_syspb.min:





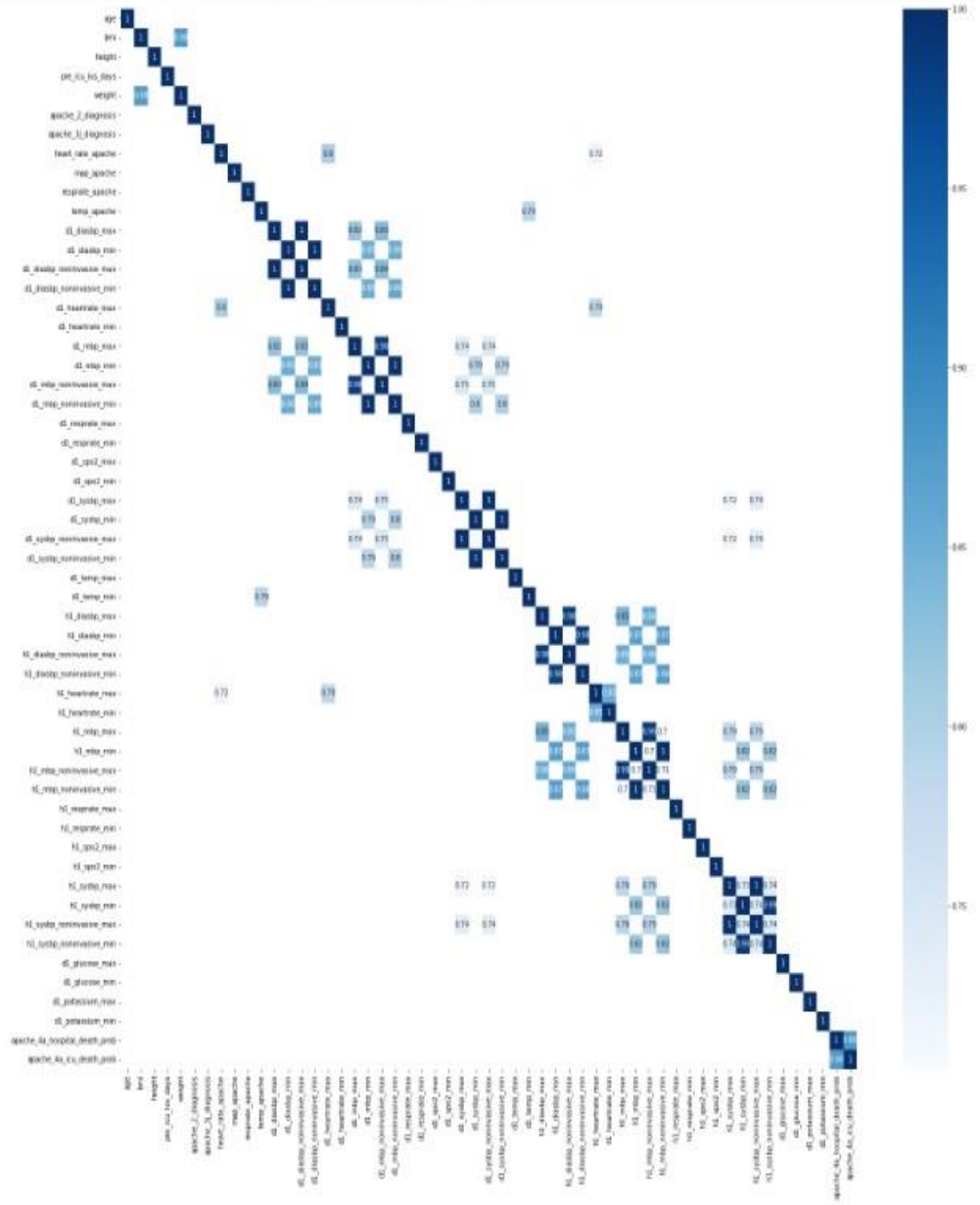
- Deaths are less as min systolic pressure occupied a normal range

3.4 Multivariate Analysis:

- Heat Map is used for Multivariate Analysis.
- Pair plots are not used because of machine limitations on the data set size
-

3.4.1 Heatmap:

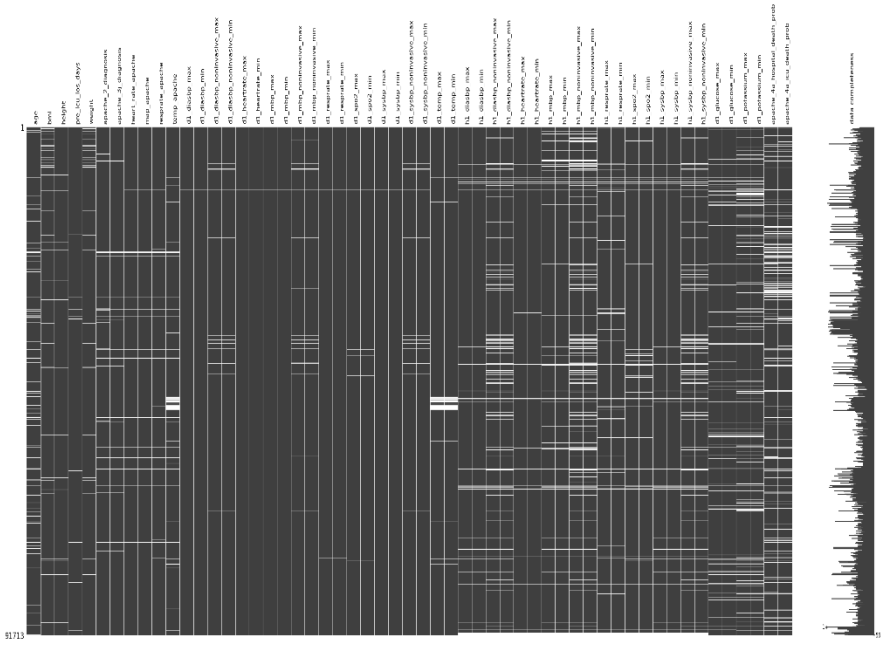
```
plt.figure(figsize=(25,25))
corr = df[numerical_columns].corr()
pearsonmap=sns.heatmap(corr[corr>=0.7],cmap='Blues',annot=True)
```



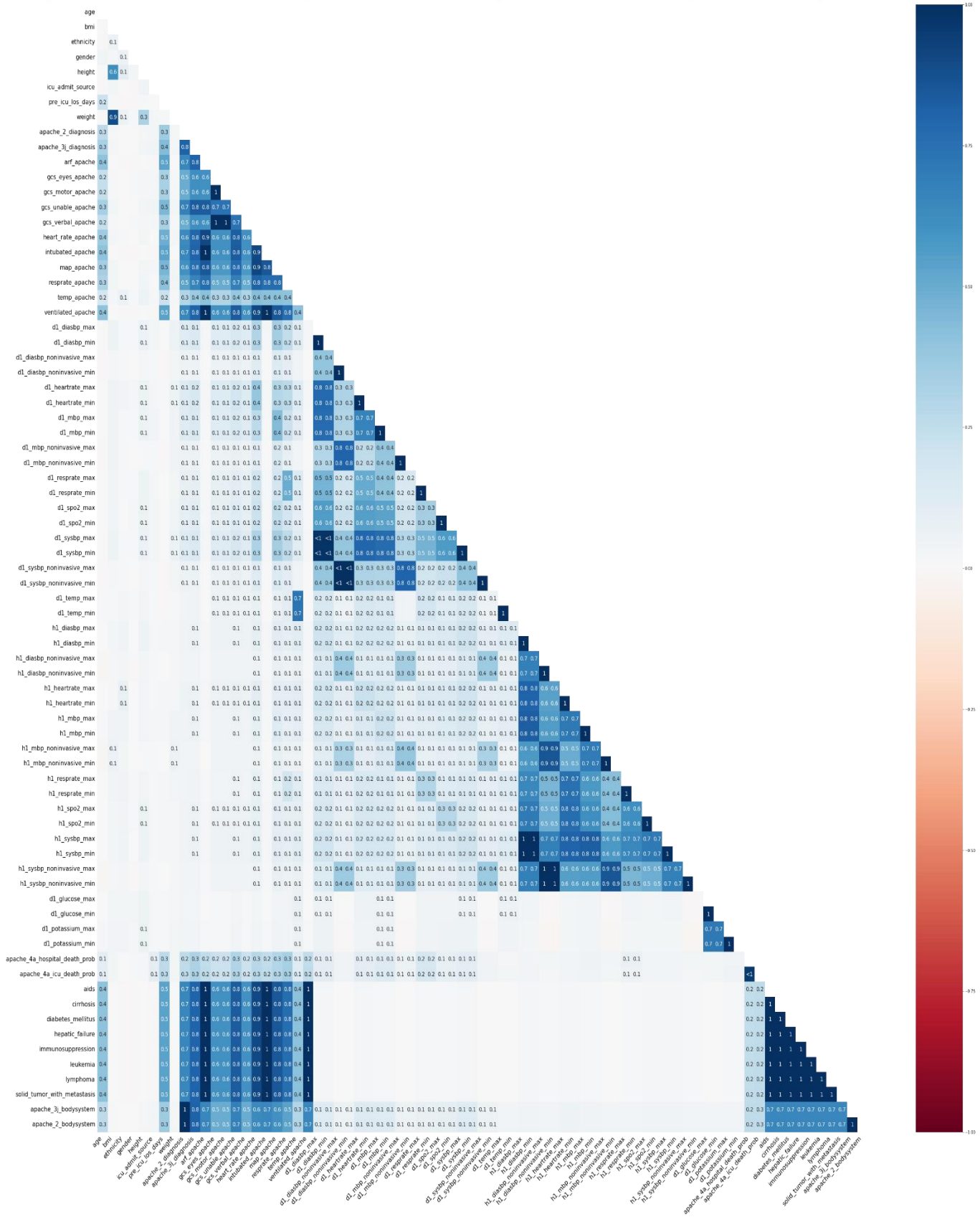
CHAPTER – 4
EXPLORATORY DATA ANALYSIS

- Our data consists of 91713 entries (rows) and 85 variables(columns).
- But arbitrary columns which doesn't play any role in relation with target were dropped making it 80 variables (73 numercial and 7 categorical) and 91713 entries.
- Out of 73 numerical types, few columns had unique values more than 2, this seems like they are kind of ordinal values or nominal values. But they didn't have high cardinality and accordingly were changed to categorical types.
- Missing data: Missing data were not treated initially and model performance was checked using Logarithmic regression only in an effort to know the impact of these
- The Target was also not treated with imbalance to know the effect on performance
- The outliers are treated very carefully and winsorization and capping was implemented. Each column and its relation with respect to target was kept in consideration and its significance was understood using health journals.
- Certain Columns have been grouped into bins because transformations had minimal impact.
- Columns showing non normality were transformed using scaling techniques such as square root, power transformation and these were considered based on kdeplot.
- After every treatment is done the next process of training and testing is carried out.

Missing Values Representation Using Missing no Matrix(MSNO) :



Heat map for MSNO :

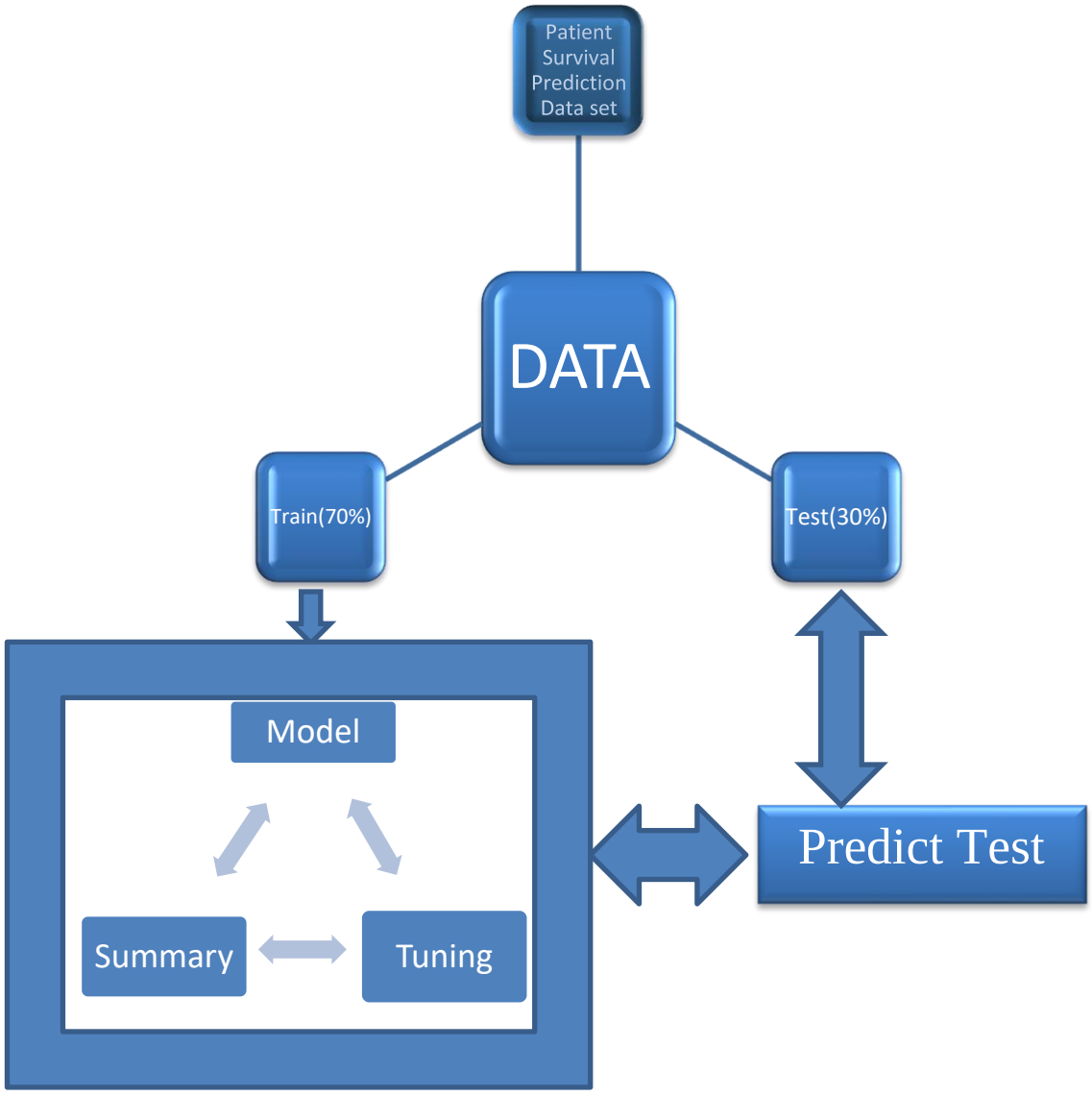

age
bmi
ethnicity
gender
height
icu_admit_source
pre_icu_los_days
weight
apache_2_diagnosis
apache_3j_diagnosis
arf_apache
gcs_eyes_apache
gcs_motor_apache
gcs_unable_apache
gcs_verbal_apache
heart_rate_apache
intubated_apache
map_apache
resprate_apache
temp_apache
ventilated_apache
d1_diasbp_max
d1_diasbp_min
d1_diasbp_noninvasive_max
d1_diasbp_noninvasive_min
d1_heartrate_max
d1_heartrate_min
d1_mbp_max
d1_mbp_min
d1_mbp_noninvasive_max
d1_mbp_noninvasive_min
d1_resprate_max
d1_resprate_min
d1_spo2_max
d1_spo2_min
d1_sysbp_max
d1_sysbp_min
d1_sysbp_noninvasive_max
d1_sysbp_noninvasive_min
d1_temp_max
d1_temp_min
h1_diasbp_max
h1_diasbp_min
h1_diasbp_noninvasive_max
h1_diasbp_noninvasive_min
h1_heartrate_max
h1_heartrate_min
h1_mbp_max
h1_mbp_min
h1_mbp_noninvasive_max
h1_mbp_noninvasive_min
h1_resprate_max
h1_resprate_min
h1_spo2_max
h1_spo2_min
h1_sysbp_max
h1_sysbp_min
h1_sysbp_noninvasive_max
h1_sysbp_noninvasive_min
d1_glucose_max
d1_glucose_min
d1_potassium_max
d1_potassium_min
apache_4a_hospital_death_prob
apache_4a_icu_death_prob
aids
cirrhosis
diabetes_mellitus
hepatic_failure
immunosuppression
leukemia
lymphoma
solid_tumor_with_metastasis
apache_3j_bodysystem
apache_2_bodysystem

CHAPTER - 5
OVERVIEW OF THE MODEL

5 Model Evaluation :

1. Data now is all set to be trained on the model in order to predict the survival.
2. Data is split into train and test and all the model building is done on train data set and the performance is predicted for actual test data

Splitting the data into training & validation datasets, i.e. training dataset : the actual dataset that we use to train the model, model sees and learns from this data; validation dataset : the sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration.

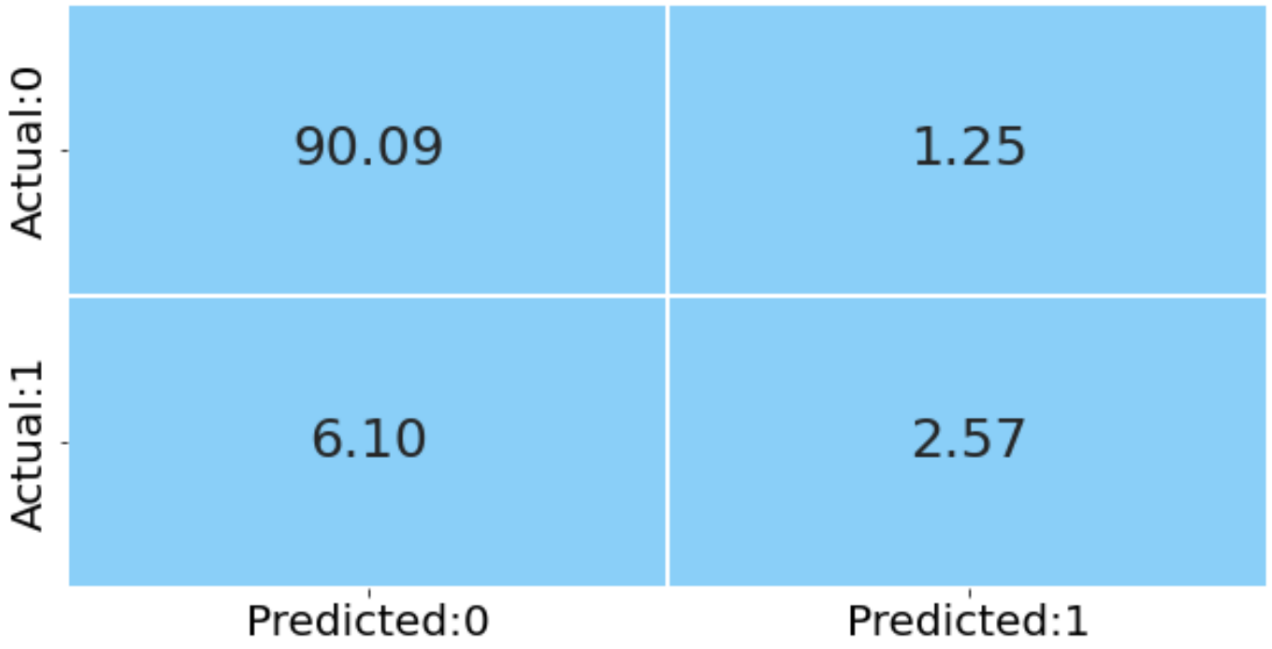


3. Building the base line logistic regression model taking into consideration all the variables. Logistic Regression is a traditional regression model and is preferred when working on target having classes in two's.
4. After building the base line model, the performance was recorded using classification report summary , AUC score and ROC curve.In this stage the selection of import features was still ambiguous since no threshold can be fixed. Hence Feature Selection process was the next step. Howeever to know if the model is fit enough ensemble models were also used.

Train results

```
ROC AUC: 87.8126
Accuracy: 92.6541 %
Avg Precision Score: 0.51
```

	precision	recall	f1-score	support
0	0.94	0.99	0.96	58637
1	0.67	0.30	0.41	5562
accuracy			0.93	64199
macro avg	0.80	0.64	0.69	64199
weighted avg	0.91	0.93	0.91	64199

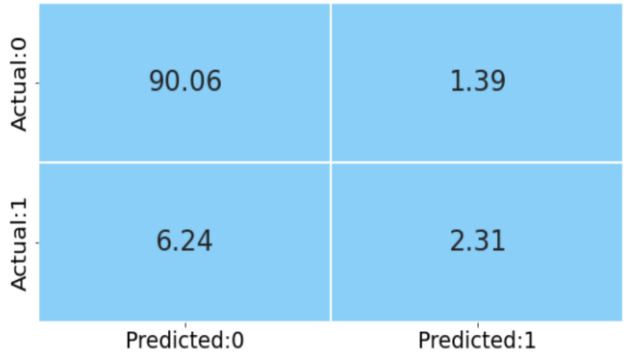


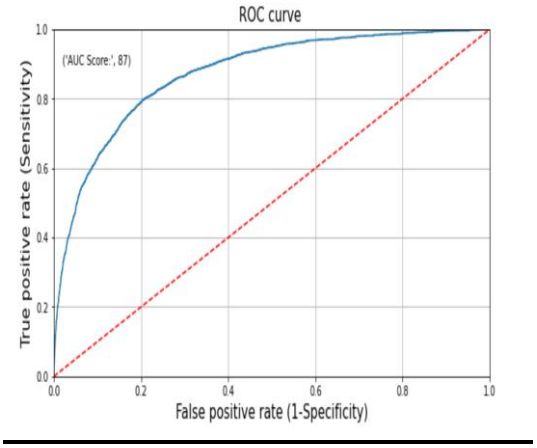
Test results :

Classification Report

Receiver Operating Characteristics curve

```
ROC AUC: 87.3107
Accuracy: 92.3675 %
Avg Precision Score: 0.47
              precision    recall  f1-score   support

           0       0.94      0.98      0.96     25161
           1       0.62      0.27      0.38      2353

    accuracy                           0.92     27514
   macro avg       0.78      0.63      0.67     27514
weighted avg       0.91      0.92      0.91     27514
```





6 We have an ensemble technique which provides us with the relative importance of the variables i.e. Random Forest. Feature importance is used commonly & tells us that which variable is contributing the most to a model & is critical to interpreting the results.The technique resulted in providing some visualization & insights from which we were able to recognize the important variables & tried removing the insignificant variables, taking into consideration the relative importance values.

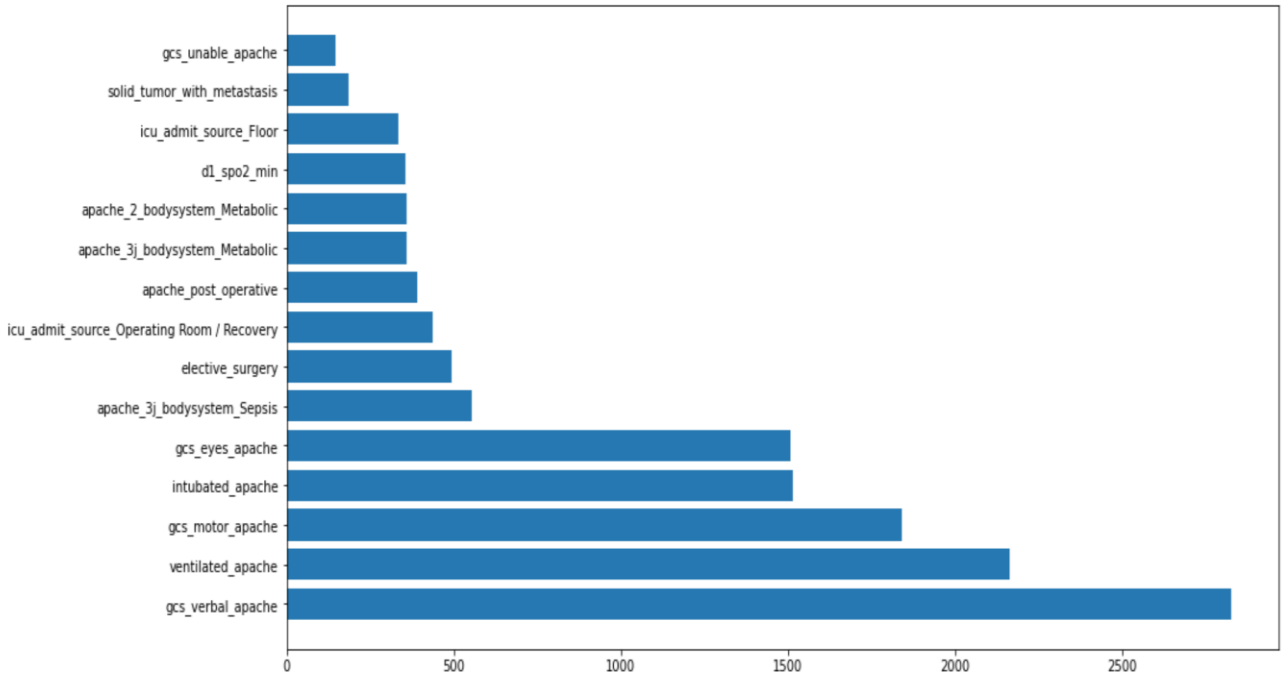
7 Now, we had our important features in the bucket. So, we used those features to build the Logistic Regression model again. But, the problem of multicollinearity still remains the same which the condition number in the model summary shows very well.

8 Multicollinearity occurs when independent variables in a regression model are correlated. This correlation is a problem because independent variables should be independent. If the degree of correlation between variables is high enough, it can cause problems when you fit the model and interpret the results. We used Correlation coefficient to remove the multicollinearity in the data by removing the variables having higher values for the correlation Many columns are highly correlated .The threshold for correlation is being considered as 0.98 or 98% for this dataset. Accordingly, there will be 12 highly correlated columns which will be dropped from the data frame.

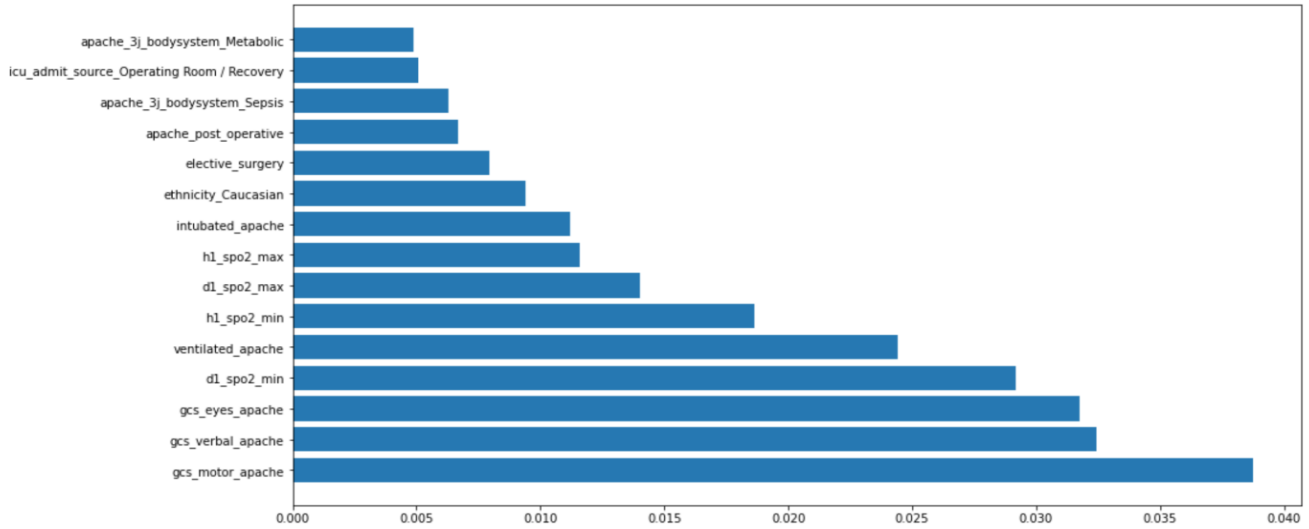
Statistical Test For Feature Selection :

Though highly correlated columns were removed there was a need to select most significant features . So statistics makes pave for this. We used Chi-Square test mutual information and f_classifier for feature selection.

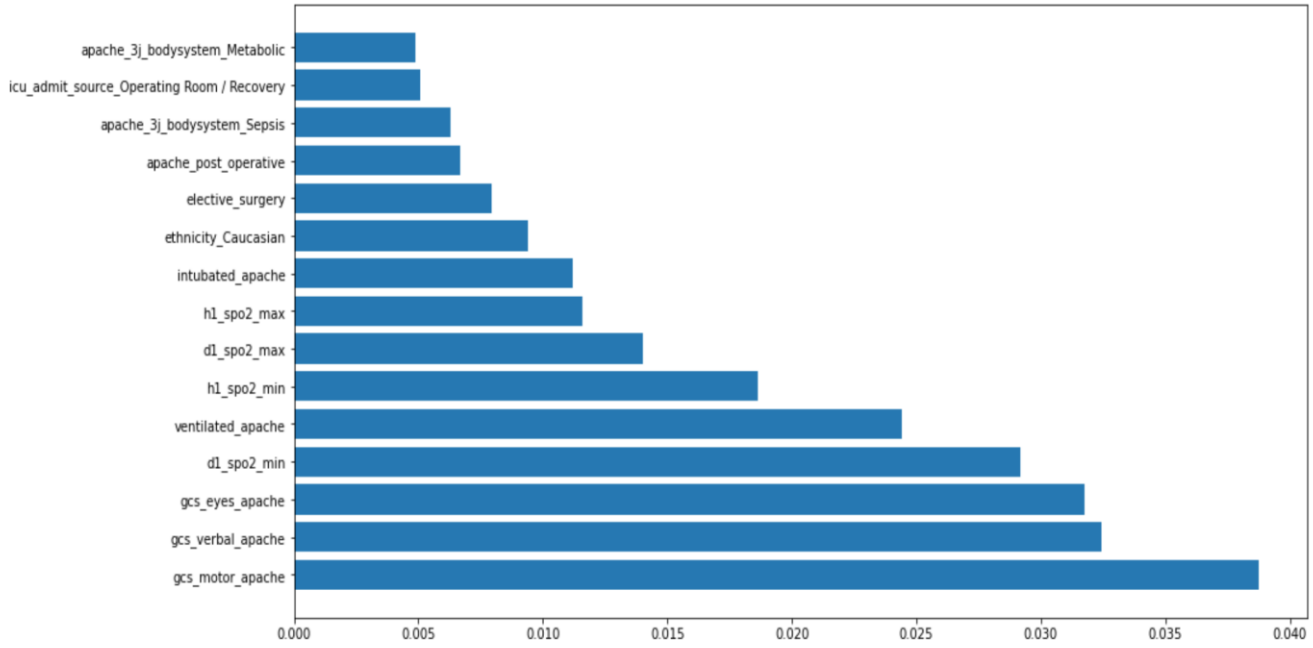
Chi Square results :



Mutual Information feature selection :



F_classifier :



9 It is evident that feature importance resulted almost similar results when various statistical tests were performed.

Missing Value Treatment 2.0 :

1. The missing values in the health data is very critical cause one wrong prediction may results in several problems. So randomly filling the data with mean ,median , mode may not give true sense of actual data.
2. So missing value treatment is done using KNN-Imputation or Multiple imputers. Both imputations fetched almost same results meaning the model performance had minimal impact when these two imputation techniques were used.

Target Class Imbalance :

1. It was evivdent that the recall becomes an important parameter when dealing with health predictions.
2. The base models performance on precision and accuracy was good but total recall was very poor. In order to improve these SMOTE+Tomek was used to balance out these class imbalance problem.

Precision Recall Trade-off:

Youden index is used to arrive at a conclusion in fixing threshold value to balance this tradeoff.

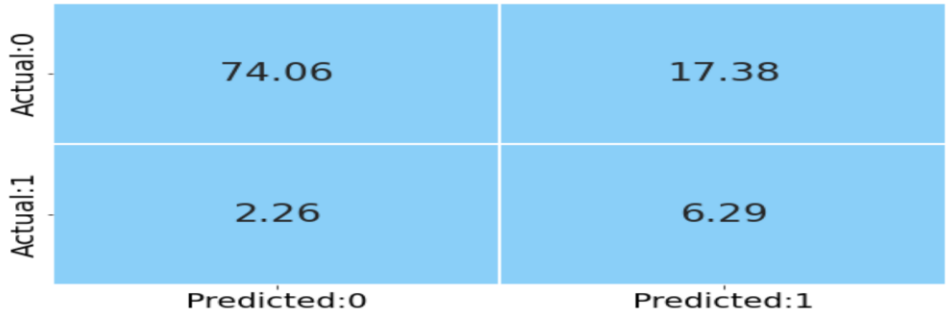
	TPR	FPR	Threshold	Difference
0	0.796430	0.203569	0.075895	0.592861
1	0.796855	0.204125	0.075707	0.592730
2	0.797280	0.204761	0.075482	0.592519
3	0.796005	0.203529	0.075898	0.592476
4	0.796005	0.203569	0.075895	0.592436

TPR : True Positive Rate
FPR: False Positive Rate

Model Building After Class Imbalance treatment and missing value Treatment:

A. Logistic Regression Model :

```
ROC AUC: 85.1138
Accuracy: 80.3555 %
Avg Precision Score: 0.42
              precision    recall  f1-score   support

           0       0.97      0.81      0.88     25161
           1       0.27      0.74      0.39      2353

    accuracy                           0.80     27514
   macro avg       0.62      0.77      0.64     27514
weighted avg       0.91      0.80      0.84     27514
```



Cross Validation :

FBeta score is mean of Precicion and recall . The value depends on how the threshold is considered . It strikes a balance between these two giving more importance to one among them depending on threshold value(0.5).

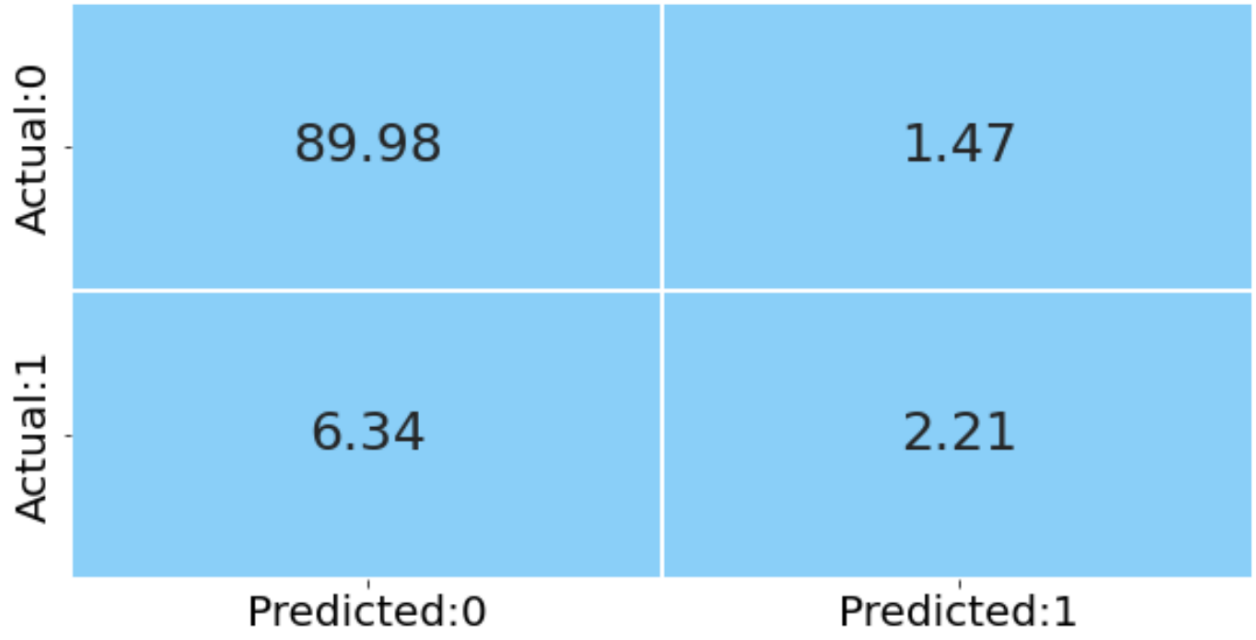
F-Beta Score : 0.34
F1-Score : 0.37

Model with Top features:

A Logistic Regression :

```
ROC AUC: 86.3889
Accuracy: 92.1931 %
Avg Precision Score: 0.45
```

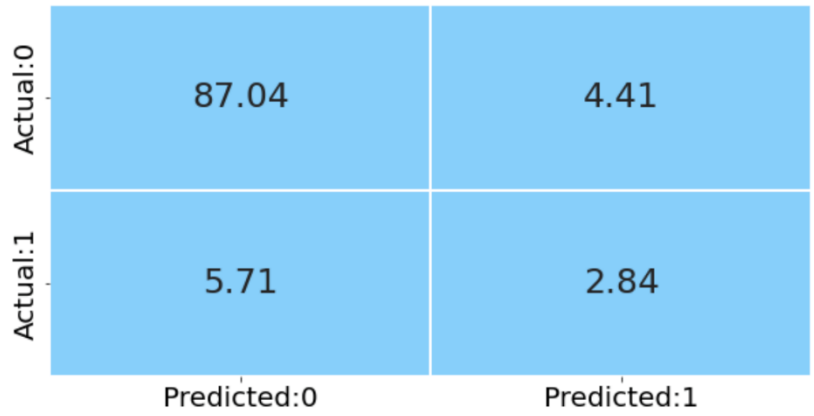
	precision	recall	f1-score	support
0	0.93	0.98	0.96	25161
1	0.60	0.26	0.36	2353
accuracy			0.92	27514
macro avg	0.77	0.62	0.66	27514
weighted avg	0.91	0.92	0.91	27514



B: Decision Tree :

```
ROC AUC: 65.5616
Accuracy: 89.8815 %
Avg Precision Score: 0.24
```

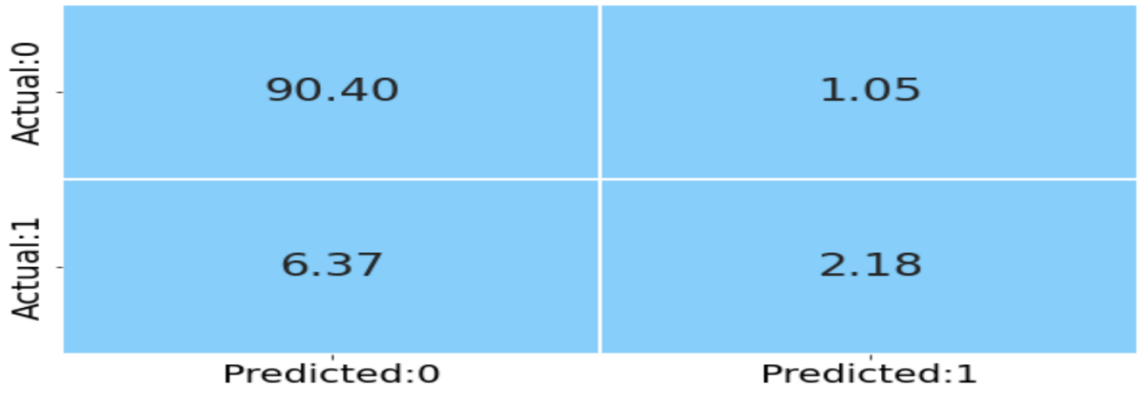
	precision	recall	f1-score	support
0	0.94	0.95	0.95	25161
1	0.39	0.33	0.36	2353
accuracy			0.90	27514
macro avg	0.67	0.64	0.65	27514
weighted avg	0.89	0.90	0.90	27514



C : Random Forest :

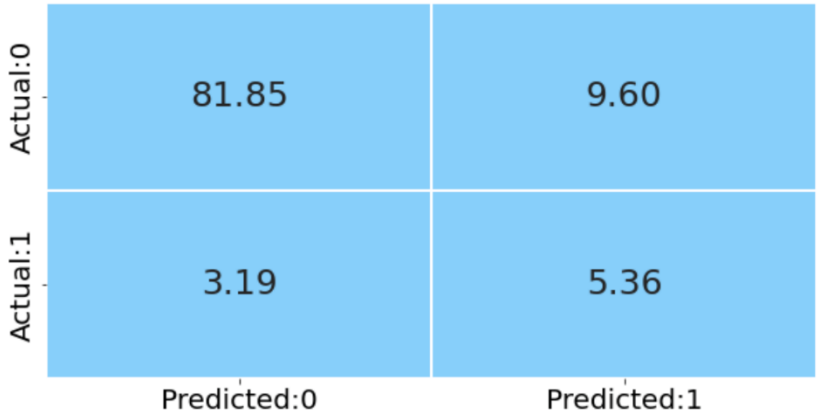
```
ROC AUC: 88.0636
Accuracy: 92.5783 %
Avg Precision Score: 0.49
```

	precision	recall	f1-score	support
0	0.93	0.99	0.96	25161
1	0.68	0.25	0.37	2353
accuracy			0.93	27514
macro avg	0.80	0.62	0.67	27514
weighted avg	0.91	0.93	0.91	27514



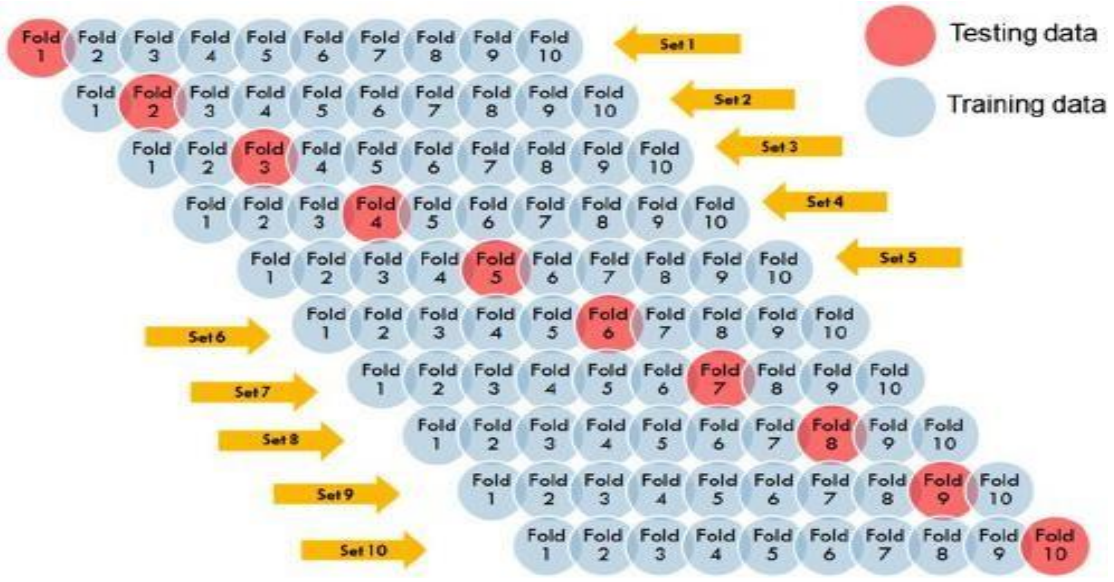
D : XG Boost :

```
ROC AUC: 86.6068
Accuracy: 87.2065 %
Avg Precision Score: 0.48

              precision    recall  f1-score   support

           0       0.96      0.89      0.93     25161
           1       0.36      0.63      0.46      2353

    accuracy                           0.87     27514
   macro avg       0.66      0.76      0.69     27514
weighted avg       0.91      0.87      0.89     27514
```



Model Building Using Hyperparameter Tuning :

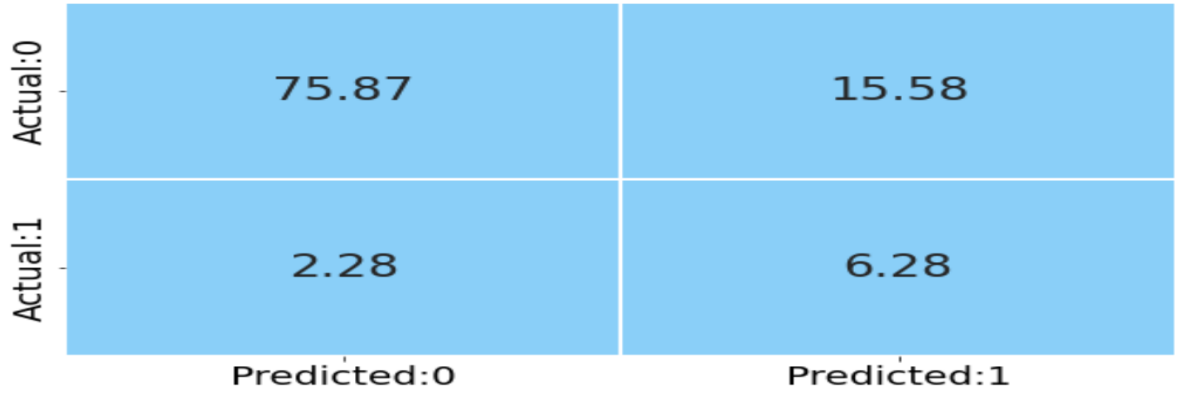
K-fold cross validation technique to improve the results of our model so that we can predict the target variable better



A: Logistic Regression :

```
f2 score: 0.5597692207960585
ROC AUC: 86.2505
Accuracy: 82.1436 %
Avg Precision Score: 0.44
```

	precision	recall	f1-score	support
0	0.97	0.83	0.89	25161
1	0.29	0.73	0.41	2353
accuracy			0.82	27514
macro avg	0.63	0.78	0.65	27514
weighted avg	0.91	0.82	0.85	27514



Classification Report

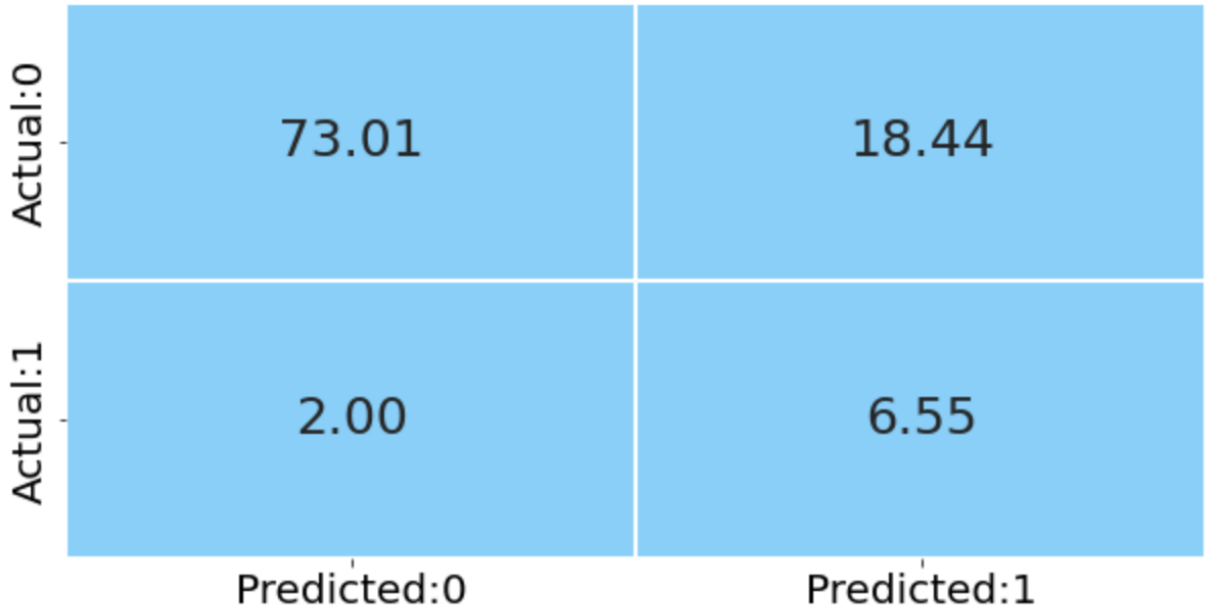
Best Parameters for Logistic Regression :

ranks	scores_mean	scores0	params	C	class_weight	max_iter	penalty	solver
1	0.558281	0.568420	{'C': 0.01, 'class_weight': 'balanced', 'max_i...	0.01	balanced	200	l1	saga
2	0.557422	0.565410	{'C': 0.01, 'class_weight': 'balanced', 'max_i...	0.01	balanced	200	l2	lbfgs
2	0.557422	0.565410	{'C': 0.01, 'class_weight': 'balanced', 'max_i...	0.01	balanced	200	l2	sag
4	0.557407	0.565410	{'C': 0.01, 'class_weight': 'balanced', 'max_i...	0.01	balanced	200	l2	newton-cg
5	0.557224	0.564796	{'C': 0.01, 'class_weight': 'balanced', 'max_i...	0.01	balanced	200	l2	saga
6	0.556240	0.561751	{'C': 0.1, 'class_weight': 'balanced', 'max_it...	0.10	balanced	200	l2	saga
6	0.556240	0.561751	{'C': 0.1, 'class_weight': 'balanced', 'max_it...	0.10	balanced	200	l2	newton-cg
8	0.556224	0.561751	{'C': 0.1, 'class_weight': 'balanced', 'max_it...	0.10	balanced	200	l2	sag
9	0.556224	0.561674	{'C': 0.1, 'class_weight': 'balanced', 'max_it...	0.10	balanced	200	l2	lbfgs
10	0.556209	0.561976	{'C': 1.0, 'class_weight': 'balanced', 'max_it...	1.00	balanced	200	l1	saga
11	0.556194	0.561898	{'C': 1.0, 'class_weight': 'balanced', 'max_it...	1.00	balanced	200	none	newton-cg
11	0.556194	0.561898	{'C': 0.01, 'class_weight': 'balanced', 'max_i...	0.01	balanced	200	none	newton-cg
11	0.556194	0.561898	{'C': 100, 'class_weight': 'balanced', 'max_it...	100.00	balanced	200	none	newton-cg
11	0.556194	0.561898	{'C': 0.1, 'class_weight': 'balanced', 'max_it...	0.10	balanced	200	none	newton-cg
11	0.556194	0.561898	{'C': 10, 'class_weight': 'balanced', 'max_ite...	10.00	balanced	200	l2	newton-cg
11	0.556194	0.561898	{'C': 10, 'class_weight': 'balanced', 'max_ite...	10.00	balanced	200	none	newton-cg
11	0.556194	0.561898	{'C': 100, 'class_weight': 'balanced', 'max_it...	100.00	balanced	200	l2	newton-cg
18	0.556179	0.561898	{'C': 10, 'class_weight': 'balanced', 'max_ite...	10.00	balanced	200	l1	saga
19	0.556178	0.561898	{'C': 0.01, 'class_weight': 'balanced', 'max_i...	0.01	balanced	200	none	sag
19	0.556178	0.561898	{'C': 0.1, 'class_weight': 'balanced', 'max_it...	0.10	balanced	200	none	sag

B: Decision Tree :

```
f2 score: 0.5534409724353859
ROC AUC: 86.0915
Accuracy: 79.5595 %
Avg Precision Score: 0.44
```

	precision	recall	f1-score	support
0	0.97	0.80	0.88	25161
1	0.26	0.77	0.39	2353
accuracy			0.80	27514
macro avg	0.62	0.78	0.63	27514
weighted avg	0.91	0.80	0.84	27514

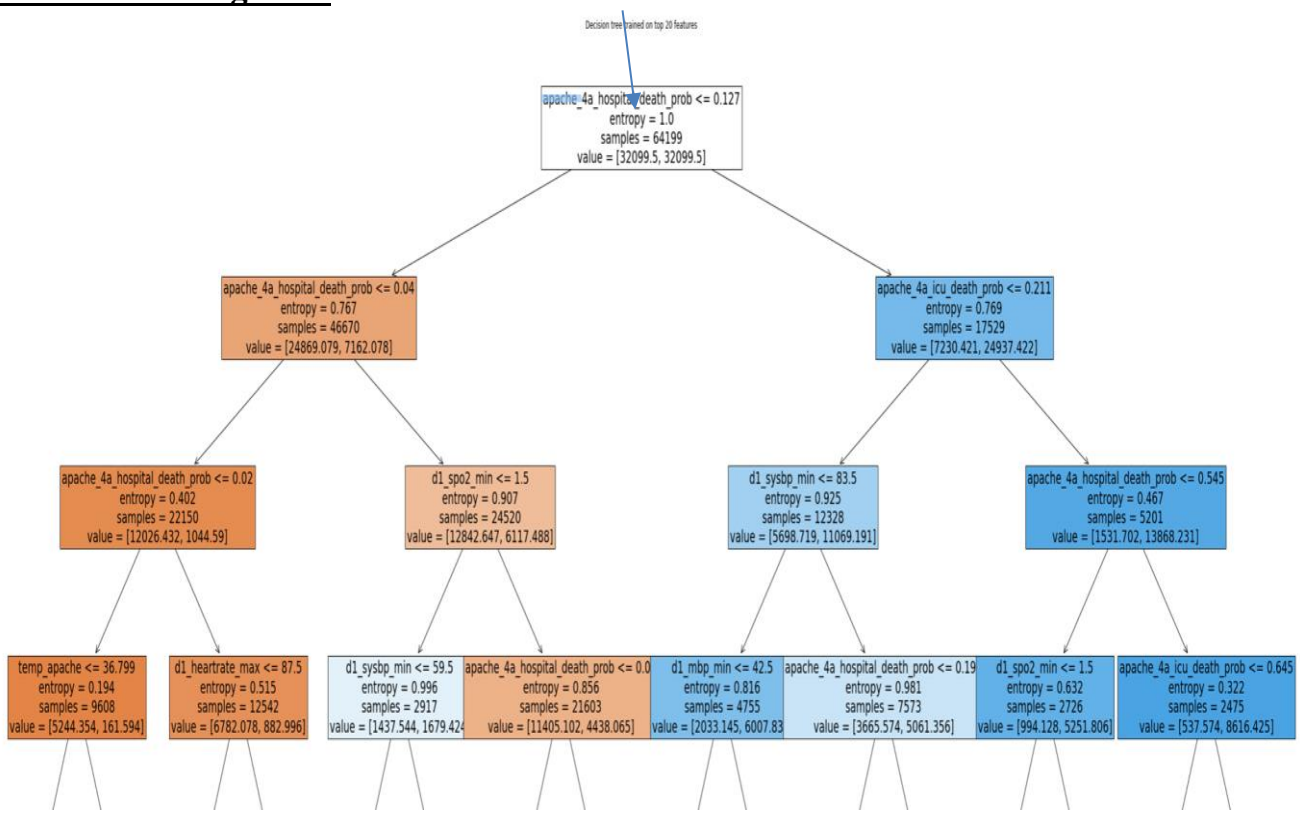


Classification Report

scores0	params	class_weight	criterion	max_depth	min_samples_leaf	min_samples_split	f2score	auc_roc	accuracy	prec	recall	f1score
0.552147	0.550314	{'class_weight': 'balanced', 'criterion': 'ent...	balanced	entropy	7	30	20	0.553441	78.229727	0.795595	0.262178	0.766256
0.552147	0.550314	{'class_weight': 'balanced', 'criterion': 'ent...	balanced	entropy	7	30	5	0.553441	78.229727	0.795595	0.262178	0.766256
0.552147	0.550314	{'class_weight': 'balanced', 'criterion': 'ent...	balanced	entropy	7	30	2	0.553441	78.229727	0.795595	0.262178	0.766256
0.552147	0.550314	{'class_weight': 'balanced', 'criterion': 'ent...	balanced	entropy	7	30	10	0.553441	78.229727	0.795595	0.262178	0.766256
0.550820	0.549666	{'class_weight': 'balanced', 'criterion': 'ent...	balanced	entropy	7	20	20	0.553855	78.285685	0.794505	0.261454	0.768806
0.550820	0.549666	{'class_weight': 'balanced', 'criterion': 'ent...	balanced	entropy	7	20	10	0.553855	78.285685	0.794505	0.261454	0.768806
0.550820	0.549666	{'class_weight': 'balanced', 'criterion': 'ent...	balanced	entropy	7	20	5	0.553855	78.285685	0.794505	0.261454	0.768806
0.550820	0.549666	{'class_weight': 'balanced', 'criterion': 'ent...	balanced	entropy	7	20	2	0.553855	78.285685	0.794505	0.261454	0.768806
0.550255	0.555059	{'class_weight': 'balanced', 'criterion': 'gin...	balanced	gini	5	5	2	0.544201	77.744872	0.782147	0.249691	0.771781
0.550255	0.555059	{'class_weight': 'balanced', 'criterion': 'gin...	balanced	gini	5	5	5	0.544201	77.744872	0.782147	0.249691	0.771781

Best Parameters for Decision Tree:

Decision Tree diagram:



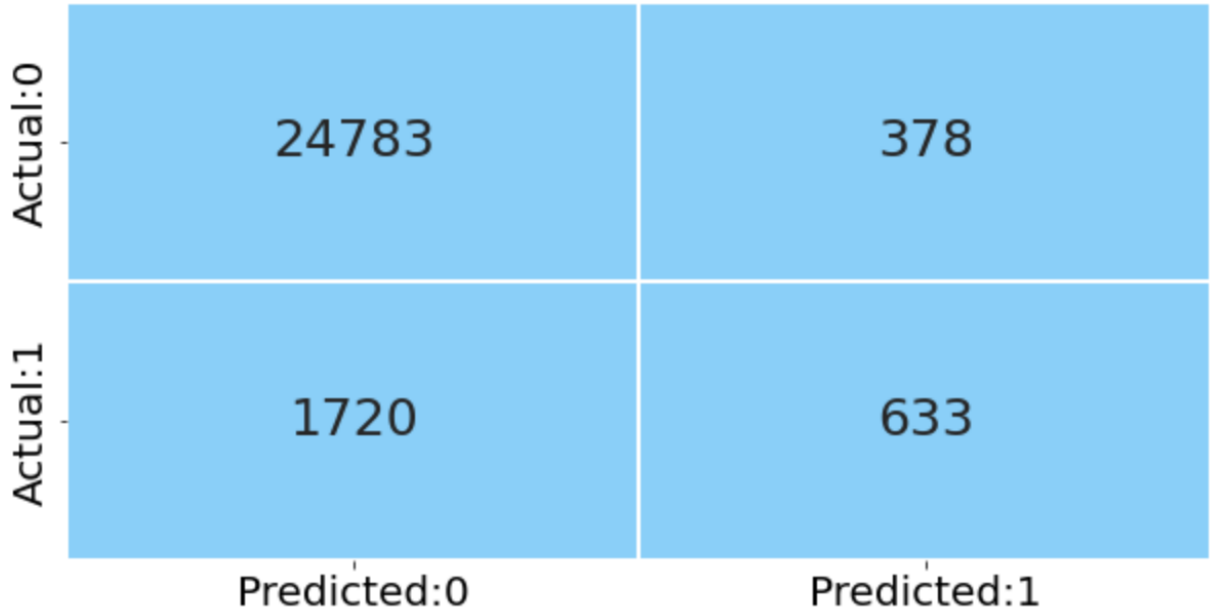
Principal Component Analysis:

PCA is often used when there is dimensionality curse. Our data consisted of 80 columns but after data transformation and feature improvements (one-Hot-Encoding) the columns increased . Therefore in order to reduce these dimensionality curse PCA is performed and Logistic Regression model is built on it.

Logistic Regression on PCA :

```
ROC AUC: 87.2696
Accuracy: 92.3748 %
Avg Precision Score: 0.47
```

	precision	recall	f1-score	support
0	0.94	0.98	0.96	25161
1	0.63	0.27	0.38	2353
accuracy			0.92	27514
macro avg	0.78	0.63	0.67	27514
weighted avg	0.91	0.92	0.91	27514



It is clear here that the performance is good but not better than previous models. Therefore PCA is not carried out further.

Chapter 6

Summary

Chronic diseases are generally considered less important for the prediction than the other types of variables. Possible explanations for this are that chronic diseases are not affecting mortality to a large extent (independent of the model used), or due to a lack of samples since chronic diseases are rare. It is important to be aware that a feature that has a large impact on mortality in real life may be considered unimportant by a model. Some of the other binary variables are considered very important by many models, such as *vent*, whether the patient was intubated, and if the patient had elective surgery. The admission diagnosis does not seem to be considered very important by any of the models. The unit admit source is also not considered very important unless the source is the Operating room.

It is interesting to note how different models consider different features most important. This is particul arly evident with the high importance, which all other models consider significantly less important. Suc h as gcs_motor_apache,gcs_eyes_apache,gcs_verbal_apache,d1_spo2_min, ventilated_apache,h1_spo2 _min,d1_spo2_max,h1_spo2_max,intubated_apache,ethnicity_Caucasian,icu_admit_source_Operating Room / Recovery,elective_surgery,apache_2_bodysystem_Metabolic,apache_3j_bodysystem_Metaboli c,apache_post_operative . These particular columns are considered important by the model but many m edical technicians might consider other columns important.

Decision Tree plot is more accurate than the other models. Features contribute more towards increased mortality than increased chances of survival. The summary looks different from the other models because of more discrete values rather than continuous, due to the mathematical fundamentals of the model. The Decision Tree summary looks less discrete if the number of decision trees used to build the model increases. However, more trees do not necessarily lead to improved performance.

CHAPTER 7
LIMITATIONS AND SCOPE

7.1 Limitations:

Although the classifier gives a highly accurate result, there are certain limitations to the project:

- In data science, prediction models can never be 100% accurate. For example in our dataset it could be because of the unidentified features which we fail to gather or observe may also contribute to the classification of the image. This could make our model even more robust. Searching and identifying such features might improve our model performance.
- In spite of being highly relevant data, since most of our data is a set of binary class variables, thus we see less scope of visualization and bivariate analysis in terms of categorical variables.
- Statistical measures such as Pearson's correlation coefficient matrix to find out multi-collinearity among variables was applicable for continuous variables and not categorical.

7.2 Scope:

- AutoML tool could be used to verify for the model performance.
- LightGBM Classifier can be adopted as one of the model and check for the parameters. LightGBM is a gradient boosting framework based on decision trees to increases the efficiency of the model and reduces memory usage.
- Since it is health data,collecting more data in various stages and several hospitals gives better results.

References:

- **https://www.kaggle.com/datasets/mitishaagarwal/patient/code**
- https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/
- https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/
- https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e
- https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/
- https://medium.com/ibm-data-science-experience/missing-data-conundrum-exploration-and-imputation-techniques-9f40abe0fd87
- https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4