Patient Survival Prediction

Capstone Final Report

**Mentored by**

Animesh Tiwari

**Group 1**

Aakriti Gulati

Ciana S Raj

Md Sharib Zeya

Misha Rajagobi

Sheetal Yadav

# Problem Statement

The aim is to build a model that helps to predict whether the patient will survive based on their physiology factors that is collected at the time of patient’s admission. By building the model, it is to help healthcare providers prioritize their patients as well as improve survival outcomes since population is increasing at a faster rate along with increase in comorbidities. The source of the dataset used to help build the model is from Kaggle. The dataset contains 91713 records and 186 attributes.

# About the Dataset

The dataset has 186 attributes which is split up into 9 categories. From these attributes, we are considering hospital\_death as our target variable. There are 170 float, 8 int, and 8 object data types.

* Identifier
* Demographic
* APACHE covariate
* APACHE comorbidity
* APACHE grouping
* APACHE prediction
* Labs
* Labs blood gas
* Vitals

## Attributes

### Identifier

The attributes under the identifier category are to uniquely identify the records.

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Data Type** | **Description** |
| encounter\_id | integer | Unique identifier associated with a patient unit stay |
| hospital\_id | integer | Unique identifier associated with a hospital |
| patient\_id | integer | Unique identifier associated with a patient |
| icu\_id | integer | A unique identifier for the unit to which the patient was admitted |

### Demographic

The demographic attributes tell us the details of patient, as well as information about which unit the patient is being admitted to and whether it was a scheduled surgery.

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Data Type** | **Description** |
| hospital\_death | binary | Whether the patient died during this hospitalization |
| age | numeric | The age of the patient on unit admission |
| bmi | string | The body mass index of the person on unit admission |
| elective\_surgery | binary | Whether the patient was admitted to the hospital for an elective surgical operation |
| ethnicity | string | The common national or cultural tradition which the person belongs to |
| gender | string | The genotypical sex of the patient |
| height | numeric | The height of the person on unit admission |
| hospital\_admit\_source | string | The location of the patient prior to being admitted to the hospital |
| icu\_admit\_source | string | The location of the patient prior to being admitted to the unit |
| icu\_admit\_type | string | The type of unit admission for the patient |
| icu\_stay\_type | string |  |
| icu\_type | string | A classification which indicates the type of care the unit is capable of providing |
| pre\_icu\_los\_days | numeric | The length of stay of the patient between hospital admission and unit admission |
| readmission\_status | binary | Whether the current unit stay is the second (or greater) stay at an ICU within the same hospitalization |
| weight | numeric | The weight (body mass) of the person on unit admission |

### APACHE covariate

The following attributes contains readings that were taken at the time of admission. It contains the diagnosis as well.

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Data Type** | **Description** |
| albumin\_apache | numeric | The albumin concentration measured during the first 24 hours which results in the highest APACHE III score |
| apache\_2\_diagnosis | string | The APACHE II diagnosis for the ICU admission |
| apache\_3j\_diagnosis | string | The APACHE III-J sub-diagnosis code which best describes the reason for the ICU admission |
| apache\_post\_operative | binary | The APACHE operative status; 1 for post-operative, 0 for non-operative |
| arf\_apache | binary | 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 |
| bilirubin\_apache | numeric | The bilirubin concentration measured during the first 24 hours which results in the highest APACHE III score |
| bun\_apache | numeric | The blood urea nitrogen concentration measured during the first 24 hours which results in the highest APACHE III score |
| creatinine\_apache | numeric | The creatinine concentration measured during the first 24 hours which results in the highest APACHE III score |
| fio2\_apache | numeric | The fraction of inspired oxygen from the arterial blood gas taken during the first 24 hours of unit admission which produces the highest APACHE III score for oxygenation |
| gcs\_eyes\_apache | integer | 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 | integer | 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 | binary | Whether the Glasgow Coma Scale was unable to be assessed due to patient sedation |
| gcs\_verbal\_apache | integer | The verbal component of the Glasgow Coma Scale measured during the first 24 hours which results in the highest APACHE III score |
| glucose\_apache | numeric | The glucose concentration measured during the first 24 hours which results in the highest APACHE III score |
| heart\_rate\_apache | numeric | The heart rate measured during the first 24 hours which results in the highest APACHE III score |
| hematocrit\_apache | numeric | The hematocrit measured during the first 24 hours which results in the highest APACHE III score |
| intubated\_apache | binary | Whether the patient was intubated at the time of the highest scoring arterial blood gas used in the oxygenation score |
| map\_apache | numeric | The mean arterial pressure measured during the first 24 hours which results in the highest APACHE III score |
| paco2\_apache | numeric | The partial pressure of carbon dioxide from the arterial blood gas taken during the first 24 hours of unit admission which produces the highest APACHE III score for oxygenation |
| paco2\_for\_ph\_apache | numeric | The partial pressure of carbon dioxide from the arterial blood gas taken during the first 24 hours of unit admission which produces the highest APACHE III score for acid-base disturbance |
| pao2\_apache | numeric | The partial pressure of oxygen from the arterial blood gas taken during the first 24 hours of unit admission which produces the highest APACHE III score for oxygenation |
| ph\_apache | numeric | The pH from the arterial blood gas taken during the first 24 hours of unit admission which produces the highest APACHE III score for acid-base disturbance |
| resprate\_apache | numeric | The respiratory rate measured during the first 24 hours which results in the highest APACHE III score |
| sodium\_apache | numeric | The sodium concentration measured during the first 24 hours which results in the highest APACHE III score |
| temp\_apache | numeric | The temperature measured during the first 24 hours which results in the highest APACHE III score |
| urineoutput\_apache | numeric | The total urine output for the first 24 hours |
| ventilated\_apache | binary | 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 |
| wbc\_apache | numeric | The white blood cell count measured during the first 24 hours which results in the highest APACHE III score |

### APACHE Comorbidity

The comorbidities attributes contains indication of whether the patient has previously or additionally diagnosed with the following diseases.

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Data Type** | **Description** |
| aids | binary | Whether the patient has a definitive diagnosis of acquired immune deficiency syndrome (AIDS) (not HIV positive alone) |
| cirrhosis | binary | 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 | binary | Whether the patient has been diagnosed with diabetes, either juvenile or adult onset, which requires medication. |
| hepatic\_failure | binary | Whether the patient has cirrhosis and additional complications including jaundice and ascites, upper GI bleeding, hepatic encephalopathy, or coma. |
| immunosuppression | binary | 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 | binary | Whether the patient has been diagnosed with acute or chronic myelogenous leukemia, acute or chronic lymphocytic leukemia, or multiple myeloma. |
| lymphoma | binary | Whether the patient has been diagnosed with non-Hodgkin lymphoma. |
| solid\_tumor\_with\_metastasis | binary | Whether the patient has been diagnosed with any solid tumor carcinoma (including malignant melanoma) which has evidence of metastasis. |

### APACHE grouping

The variables in APACHE grouping are all categorical variables. They describe the diagnosis for the patient.

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Data Type** | **Description** |
| apache\_3j\_bodysystem | string | Admission diagnosis group for APACHE III |
| apache\_2\_bodysystem | string | Admission diagnosis group for APACHE II |

### APACHE prediction

The death probability is calculated on the basis of the APACHE covariates.

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Data Type** | **Description** |
| apache\_4a\_hospital\_death\_prob | numeric | 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 | numeric | The APACHE IVa probabilistic prediction of in ICU mortality for the patient which utilizes the APACHE III score and other covariates, including diagnosis |

### Labs

The variables containing information on highest and lowest concentration of certain tests at the first hour(h1) and first 24 hours(d1).

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Data Type** | **Description** |
| d1\_albumin\_max | numeric | The lowest albumin concentration of the patient in their serum during the first 24 hours of their unit stay |
| d1\_albumin\_min | numeric | The lowest albumin concentration of the patient in their serum during the first 24 hours of their unit stay |
| d1\_bilirubin\_max | numeric | The highest bilirubin concentration of the patient in their serum or plasma during the first 24 hours of their unit stay |
| d1\_bilirubin\_min | numeric | The lowest bilirubin concentration of the patient in their serum or plasma during the first 24 hours of their unit stay |
| d1\_bun\_max | numeric | The highest blood urea nitrogen concentration of the patient in their serum or plasma during the first 24 hours of their unit stay |
| d1\_bun\_min | numeric | The lowest blood urea nitrogen concentration of the patient in their serum or plasma during the first 24 hours of their unit stay |
| d1\_calcium\_max | numeric | The highest calcium concentration of the patient in their serum during the first 24 hours of their unit stay |
| d1\_calcium\_min | numeric | The lowest calcium concentration of the patient in their serum during the first 24 hours of their unit stay |
| d1\_creatinine\_max | numeric | The highest creatinine concentration of the patient in their serum or plasma during the first 24 hours of their unit stay |
| d1\_creatinine\_min | numeric | The lowest creatinine concentration of the patient in their serum or plasma during the first 24 hours of their unit stay |
| d1\_glucose\_max | numeric | The highest glucose concentration of the patient in their serum or plasma during the first 24 hours of their unit stay |
| d1\_glucose\_min | numeric | The lowest glucose concentration of the patient in their serum or plasma during the first 24 hours of their unit stay |
| d1\_hco3\_max | numeric | The highest bicarbonate concentration for the patient in their serum or plasma during the first 24 hours of their unit stay |
| d1\_hco3\_min | numeric | The lowest bicarbonate concentration for the patient in their serum or plasma during the first 24 hours of their unit stay |
| d1\_hemaglobin\_max | numeric | The highest hemoglobin concentration for the patient during the first 24 hours of their unit stay |
| d1\_hemaglobin\_min | numeric | The lowest hemoglobin concentration for the patient during the first 24 hours of their unit stay |
| d1\_hematocrit\_max | numeric | The highest volume proportion of red blood cells in a patient's blood during the first 24 hours of their unit stay, expressed as a fraction |
| d1\_hematocrit\_min | numeric | The lowest volume proportion of red blood cells in a patient's blood during the first 24 hours of their unit stay, expressed as a fraction |
| d1\_inr\_max | numeric | The highest international normalized ratio for the patient during the first 24 hours of their unit stay |
| d1\_inr\_min | numeric | The lowest international normalized ratio for the patient during the first 24 hours of their unit stay |
| d1\_lactate\_max | numeric | The highest lactate concentration for the patient in their serum or plasma during the first 24 hours of their unit stay |
| d1\_lactate\_min | numeric | The lowest lactate concentration for the patient in their serum or plasma during the first 24 hours of their unit stay |
| d1\_platelets\_max | numeric | The highest platelet count for the patient during the first 24 hours of their unit stay |
| d1\_platelets\_min | numeric | The lowest platelet count for the patient during the first 24 hours of their unit stay |
| d1\_potassium\_max | numeric | The highest potassium concentration for the patient in their serum or plasma during the first 24 hours of their unit stay |
| d1\_potassium\_min | numeric | The lowest potassium concentration for the patient in their serum or plasma during the first 24 hours of their unit stay |
| d1\_sodium\_max | numeric | The highest sodium concentration for the patient in their serum or plasma during the first 24 hours of their unit stay |
| d1\_sodium\_min | numeric | The lowest sodium concentration for the patient in their serum or plasma during the first 24 hours of their unit stay |
| d1\_wbc\_max | numeric | The highest white blood cell count for the patient during the first 24 hours of their unit stay |
| d1\_wbc\_min | numeric | The lowest white blood cell count for the patient during the first 24 hours of their unit stay |
| h1\_albumin\_max | numeric | The lowest albumin concentration of the patient in their serum during the first hour of their unit stay |
| h1\_albumin\_min | numeric | The lowest albumin concentration of the patient in their serum during the first hour of their unit stay |
| h1\_bilirubin\_max | numeric | The highest bilirubin concentration of the patient in their serum or plasma during the first hour of their unit stay |
| h1\_bilirubin\_min | numeric | The lowest bilirubin concentration of the patient in their serum or plasma during the first hour of their unit stay |
| h1\_bun\_max | numeric | The highest blood urea nitrogen concentration of the patient in their serum or plasma during the first hour of their unit stay |
| h1\_bun\_min | numeric | The lowest blood urea nitrogen concentration of the patient in their serum or plasma during the first hour of their unit stay |
| h1\_calcium\_max | numeric | The highest calcium concentration of the patient in their serum during the first hour of their unit stay |
| h1\_calcium\_min | numeric | The lowest calcium concentration of the patient in their serum during the first hour of their unit stay |
| h1\_creatinine\_max | numeric | The highest creatinine concentration of the patient in their serum or plasma during the first hour of their unit stay |
| h1\_creatinine\_min | numeric | The lowest creatinine concentration of the patient in their serum or plasma during the first hour of their unit stay |
| h1\_glucose\_max | numeric | The highest glucose concentration of the patient in their serum or plasma during the first hour of their unit stay |
| h1\_glucose\_min | numeric | The lowest glucose concentration of the patient in their serum or plasma during the first hour of their unit stay |
| h1\_hco3\_max | numeric | The highest bicarbonate concentration for the patient in their serum or plasma during the first hour of their unit stay |
| h1\_hco3\_min | numeric | The lowest bicarbonate concentration for the patient in their serum or plasma during the first hour of their unit stay |
| h1\_hemaglobin\_max | numeric | The highest hemoglobin concentration for the patient during the first hour of their unit stay |
| h1\_hemaglobin\_min | numeric | The lowest hemoglobin concentration for the patient during the first hour of their unit stay |
| h1\_hematocrit\_max | numeric | The highest volume proportion of red blood cells in a patient's blood during the first hour of their unit stay, expressed as a fraction |
| h1\_hematocrit\_min | numeric | The lowest volume proportion of red blood cells in a patient's blood during the first hour of their unit stay, expressed as a fraction |
| h1\_inr\_max | numeric | The highest international normalized ratio for the patient during the first hour of their unit stay |
| h1\_inr\_min | numeric | The lowest international normalized ratio for the patient during the first hour of their unit stay |
| h1\_lactate\_max | numeric | The highest lactate concentration for the patient in their serum or plasma during the first hour of their unit stay |
| h1\_lactate\_min | numeric | The lowest lactate concentration for the patient in their serum or plasma during the first hour of their unit stay |
| h1\_platelets\_max | numeric | The highest platelet count for the patient during the first hour of their unit stay |
| h1\_platelets\_min | numeric | The lowest platelet count for the patient during the first hour of their unit stay |
| h1\_potassium\_max | numeric | The highest potassium concentration for the patient in their serum or plasma during the first hour of their unit stay |
| h1\_potassium\_min | numeric | The lowest potassium concentration for the patient in their serum or plasma during the first hour of their unit stay |
| h1\_sodium\_max | numeric | The highest sodium concentration for the patient in their serum or plasma during the first hour of their unit stay |
| h1\_sodium\_min | numeric | The lowest sodium concentration for the patient in their serum or plasma during the first hour of their unit stay |
| h1\_wbc\_max | numeric | The highest white blood cell count for the patient during the first hour of their unit stay |
| h1\_wbc\_min | numeric | The lowest white blood cell count for the patient during the first hour of their unit stay |

### Labs Blood Gas

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Data Type** | **Description** |
| d1\_arterial\_pco2\_max | numeric | The highest arterial partial pressure of carbon dioxide for the patient during the first 24 hours of their unit stay |
| d1\_arterial\_pco2\_min | numeric | The lowest arterial partial pressure of carbon dioxide for the patient during the first 24 hours of their unit stay |
| d1\_arterial\_ph\_max | numeric | The highest arterial pH for the patient during the first 24 hours of their unit stay |
| d1\_arterial\_ph\_min | numeric | The lowest arterial pH for the patient during the first 24 hours of their unit stay |
| d1\_arterial\_po2\_max | numeric | The highest arterial partial pressure of oxygen for the patient during the first 24 hours of their unit stay |
| d1\_arterial\_po2\_min | numeric | The lowest arterial partial pressure of oxygen for the patient during the first 24 hours of their unit stay |
| d1\_pao2fio2ratio\_max | numeric | The highest fraction of inspired oxygen for the patient during the first 24 hours of their unit stay |
| d1\_pao2fio2ratio\_min | numeric | The lowest fraction of inspired oxygen for the patient during the first 24 hours of their unit stay |
| h1\_arterial\_pco2\_max | numeric | The highest arterial partial pressure of carbon dioxide for the patient during the first hour of their unit stay |
| h1\_arterial\_pco2\_min | numeric | The lowest arterial partial pressure of carbon dioxide for the patient during the first hour of their unit stay |
| h1\_arterial\_ph\_max | numeric | The highest arterial pH for the patient during the first hour of their unit stay |
| h1\_arterial\_ph\_min | numeric | The lowest arterial pH for the patient during the first hour of their unit stay |
| h1\_arterial\_po2\_max | numeric | The highest arterial partial pressure of oxygen for the patient during the first hour of their unit stay |
| h1\_arterial\_po2\_min | numeric | The lowest arterial partial pressure of oxygen for the patient during the first hour of their unit stay |
| h1\_pao2fio2ratio\_max | numeric | The highest fraction of inspired oxygen for the patient during the first hour of their unit stay |
| h1\_pao2fio2ratio\_min | numeric | The lowest fraction of inspired oxygen for the patient during the first hour of their unit stay |

### Vitals

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Data Type** | **Description** |
| d1\_diasbp\_invasive\_max | numeric | The patient's highest diastolic blood pressure during the first 24 hours of their unit stay, invasively measured |
| d1\_diasbp\_invasive\_min | numeric | The patient's lowest diastolic blood pressure during the first 24 hours of their unit stay, invasively measured |
| d1\_diasbp\_max | numeric | 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 | numeric | 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 | numeric | The patient's highest diastolic blood pressure during the first 24 hours of their unit stay, non-invasively measured |
| d1\_diasbp\_noninvasive\_min | numeric | The patient's lowest diastolic blood pressure during the first 24 hours of their unit stay, non-invasively measured |
| d1\_heartrate\_max | numeric | The patient's highest heart rate during the first 24 hours of their unit stay |
| d1\_heartrate\_min | numeric | The patient's lowest heart rate during the first 24 hours of their unit stay |
| d1\_mbp\_invasive\_max | numeric | The patient's highest mean blood pressure during the first 24 hours of their unit stay, invasively measured |
| d1\_mbp\_invasive\_min | numeric | The patient's lowest mean blood pressure during the first 24 hours of their unit stay, invasively measured |
| d1\_mbp\_max | numeric | 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 | numeric | 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 | numeric | The patient's highest mean blood pressure during the first 24 hours of their unit stay, non-invasively measured |
| d1\_mbp\_noninvasive\_min | numeric | The patient's lowest mean blood pressure during the first 24 hours of their unit stay, non-invasively measured |
| d1\_resprate\_max | numeric | The patient's highest respiratory rate during the first 24 hours of their unit stay |
| d1\_resprate\_min | numeric | The patient's lowest respiratory rate during the first 24 hours of their unit stay |
| d1\_spo2\_max | numeric | The patient's highest peripheral oxygen saturation during the first 24 hours of their unit stay |
| d1\_spo2\_min | numeric | The patient's lowest peripheral oxygen saturation during the first 24 hours of their unit stay |
| d1\_sysbp\_invasive\_max | numeric | The patient's highest systolic blood pressure during the first 24 hours of their unit stay, invasively measured |
| d1\_sysbp\_invasive\_min | numeric | The patient's lowest systolic blood pressure during the first 24 hours of their unit stay, invasively measured |
| d1\_sysbp\_max | numeric | 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 | numeric | 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 | numeric | The patient's highest systolic blood pressure during the first 24 hours of their unit stay, non-invasively measured |
| d1\_sysbp\_noninvasive\_min | numeric | The patient's lowest systolic blood pressure during the first 24 hours of their unit stay, non-invasively measured |
| d1\_temp\_max | numeric | The patient's highest core temperature during the first 24 hours of their unit stay, invasively measured |
| d1\_temp\_min | numeric | The patient's lowest core temperature during the first 24 hours of their unit stay |
| h1\_diasbp\_invasive\_max | numeric | The patient's highest diastolic blood pressure during the first hour of their unit stay, invasively measured |
| h1\_diasbp\_invasive\_min | numeric | The patient's lowest diastolic blood pressure during the first hour of their unit stay, invasively measured |
| h1\_diasbp\_max | numeric | The patient's highest diastolic blood pressure during the first hour of their unit stay, either non-invasively or invasively measured |
| h1\_diasbp\_min | numeric | 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 | numeric | The patient's highest diastolic blood pressure during the first hour of their unit stay, non-invasively measured |
| h1\_diasbp\_noninvasive\_min | numeric | The patient's lowest diastolic blood pressure during the first hour of their unit stay, non-invasively measured |
| h1\_heartrate\_max | numeric | The patient's highest heart rate during the first hour of their unit stay |
| h1\_heartrate\_min | numeric | The patient's lowest heart rate during the first hour of their unit stay |
| h1\_mbp\_invasive\_max | numeric | The patient's highest mean blood pressure during the first hour of their unit stay, invasively measured |
| h1\_mbp\_invasive\_min | numeric | The patient's lowest mean blood pressure during the first hour of their unit stay, invasively measured |
| h1\_mbp\_max | numeric | The patient's highest mean blood pressure during the first hour of their unit stay, either non-invasively or invasively measured |
| h1\_mbp\_min | numeric | 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 | numeric | The patient's highest mean blood pressure during the first hour of their unit stay, non-invasively measured |
| h1\_mbp\_noninvasive\_min | numeric | The patient's lowest mean blood pressure during the first hour of their unit stay, non-invasively measured |
| h1\_resprate\_max | numeric | The patient's highest respiratory rate during the first hour of their unit stay |
| h1\_resprate\_min | numeric | The patient's lowest respiratory rate during the first hour of their unit stay |
| h1\_spo2\_max | numeric | The patient's highest peripheral oxygen saturation during the first hour of their unit stay |
| h1\_spo2\_min | numeric | The patient's lowest peripheral oxygen saturation during the first hour of their unit stay |
| h1\_sysbp\_invasive\_max | numeric | The patient's highest systolic blood pressure during the first hour of their unit stay, invasively measured |
| h1\_sysbp\_invasive\_min | numeric | The patient's lowest systolic blood pressure during the first hour of their unit stay, invasively measured |
| h1\_sysbp\_max | numeric | The patient's highest systolic blood pressure during the first hour of their unit stay, either non-invasively or invasively measured |
| h1\_sysbp\_min | numeric | 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 | numeric | The patient's highest systolic blood pressure during the first hour of their unit stay, non-invasively measured |
| h1\_sysbp\_noninvasive\_min | numeric | The patient's lowest systolic blood pressure during the first hour of their unit stay, non-invasively measured |
| h1\_temp\_max | numeric | The patient's highest core temperature during the first hour of their unit stay, invasively measured |
| h1\_temp\_min | numeric | The patient's lowest core temperature during the first hour of their unit stay |

## Missing Values

The first step after checking out the dataset, is to find the missing values. In this step, we check if there any attributes that can be dropped if it contains more than 70-80% missing data.

|  |  |  |
| --- | --- | --- |
| **Attributes** | **Total missing values** | **Percentage of missing values** |
| h1\_bilirubin\_max | 84619 | 92.27 % |
| h1\_bilirubin\_min | 84619 | 92.27 % |
| h1\_lactate\_max | 84369 | 91.99 % |
| h1\_lactate\_min | 84369 | 91.99 % |
| h1\_albumin\_max | 83824 | 91.4 % |
| h1\_albumin\_min | 83824 | 91.4 % |
| h1\_pao2fio2ratio\_max | 80195 | 87.44 % |
| h1\_pao2fio2ratio\_min | 80195 | 87.44 % |
| h1\_arterial\_ph\_max | 76424 | 83.33 % |
| h1\_arterial\_ph\_min | 76424 | 83.33 % |
| h1\_hco3\_max | 76094 | 82.97 % |
| h1\_hco3\_min | 76094 | 82.97 % |
| h1\_arterial\_pco2\_max | 75959 | 82.82 % |
| h1\_arterial\_pco2\_min | 75959 | 82.82 % |
| h1\_wbc\_max | 75953 | 82.82 % |
| h1\_wbc\_min | 75953 | 82.82 % |
| h1\_arterial\_po2\_max | 75945 | 82.81 % |
| h1\_arterial\_po2\_min | 75945 | 82.81 % |
| h1\_calcium\_max | 75863 | 82.72 % |
| h1\_calcium\_min | 75863 | 82.72 % |
| h1\_platelets\_max | 75673 | 82.51 % |
| h1\_platelets\_min | 75673 | 82.51 % |
| h1\_bun\_max | 75091 | 81.88 % |
| h1\_bun\_min | 75091 | 81.88 % |
| h1\_creatinine\_max | 74957 | 81.73 % |
| h1\_creatinine\_min | 74957 | 81.73 % |
| h1\_diasbp\_invasive\_max | 74928 | 81.7 % |
| h1\_diasbp\_invasive\_min | 74928 | 81.7 % |
| h1\_sysbp\_invasive\_max | 74915 | 81.68 % |
| h1\_sysbp\_invasive\_min | 74915 | 81.68 % |
| h1\_mbp\_invasive\_max | 74844 | 81.61 % |
| h1\_mbp\_invasive\_min | 74844 | 81.61 % |
| h1\_hematocrit\_max | 73420 | 80.05 % |
| h1\_hematocrit\_min | 73420 | 80.05 % |
| h1\_hemaglobin\_max | 73123 | 79.73 % |
| h1\_hemaglobin\_min | 73123 | 79.73 % |
| h1\_sodium\_max | 72617 | 79.18 % |
| h1\_sodium\_min | 72617 | 79.18 % |
| h1\_potassium\_max | 72102 | 78.62 % |
| h1\_potassium\_min | 72102 | 78.62 % |
| fio2\_apache | 70868 | 77.27 % |
| paco2\_apache | 70868 | 77.27 % |
| paco2\_for\_ph\_apache | 70868 | 77.27 % |
| pao2\_apache | 70868 | 77.27 % |
| ph\_apache | 70868 | 77.27 % |
| d1\_lactate\_max | 68396 | 74.58 % |
| d1\_lactate\_min | 68396 | 74.58 % |
| d1\_diasbp\_invasive\_max | 67984 | 74.13 % |
| d1\_diasbp\_invasive\_min | 67984 | 74.13 % |
| d1\_sysbp\_invasive\_max | 67959 | 74.1 % |
| d1\_sysbp\_invasive\_min | 67959 | 74.1 % |
| d1\_mbp\_invasive\_max | 67777 | 73.9 % |
| d1\_mbp\_invasive\_min | 67777 | 73.9 % |
| d1\_pao2fio2ratio\_max | 66008 | 71.97 % |
| d1\_pao2fio2ratio\_min | 66008 | 71.97 % |
| d1\_arterial\_ph\_max | 60123 | 65.56 % |
| d1\_arterial\_ph\_min | 60123 | 65.56 % |
| d1\_arterial\_pco2\_max | 59271 | 64.63 % |
| d1\_arterial\_pco2\_min | 59271 | 64.63 % |
| d1\_arterial\_po2\_max | 59262 | 64.62 % |
| d1\_arterial\_po2\_min | 59262 | 64.62 % |
| bilirubin\_apache | 58134 | 63.39 % |
| d1\_inr\_max | 57941 | 63.18 % |
| d1\_inr\_min | 57941 | 63.18 % |
| h1\_inr\_max | 57941 | 63.18 % |
| h1\_inr\_min | 57941 | 63.18 % |
| albumin\_apache | 54379 | 59.29 % |
| d1\_bilirubin\_max | 53673 | 58.52 % |
| d1\_bilirubin\_min | 53673 | 58.52 % |
| h1\_glucose\_max | 52614 | 57.37 % |
| h1\_glucose\_min | 52614 | 57.37 % |
| d1\_albumin\_max | 49096 | 53.53 % |
| d1\_albumin\_min | 49096 | 53.53 % |
| urineoutput\_apache | 48998 | 53.43 % |
| wbc\_apache | 22012 | 24 % |
| h1\_temp\_max | 21732 | 23.7 % |
| h1\_temp\_min | 21732 | 23.7 % |
| hospital\_admit\_source | 21409 | 23.34 % |
| hematocrit\_apache | 19878 | 21.67 % |
| bun\_apache | 19262 | 21 % |
| creatinine\_apache | 18853 | 20.56 % |
| sodium\_apache | 18600 | 20.28 % |
| d1\_hco3\_max | 15071 | 16.43 % |
| d1\_hco3\_min | 15071 | 16.43 % |
| d1\_platelets\_max | 13444 | 14.66 % |
| d1\_platelets\_min | 13444 | 14.66 % |
| d1\_wbc\_max | 13174 | 14.36 % |
| d1\_wbc\_min | 13174 | 14.36 % |
| d1\_calcium\_max | 13069 | 14.25 % |
| d1\_calcium\_min | 13069 | 14.25 % |
| d1\_hemaglobin\_max | 12147 | 13.24 % |
| d1\_hemaglobin\_min | 12147 | 13.24 % |
| d1\_hematocrit\_max | 11654 | 12.71 % |
| d1\_hematocrit\_min | 11654 | 12.71 % |
| glucose\_apache | 11036 | 12.03 % |
| d1\_bun\_max | 10514 | 11.46 % |
| d1\_bun\_min | 10514 | 11.46 % |
| d1\_sodium\_max | 10195 | 11.12 % |
| d1\_sodium\_min | 10195 | 11.12 % |
| d1\_creatinine\_max | 10169 | 11.09 % |
| d1\_creatinine\_min | 10169 | 11.09 % |
| d1\_potassium\_max | 9585 | 10.45 % |
| d1\_potassium\_min | 9585 | 10.45 % |
| h1\_mbp\_noninvasive\_max | 9084 | 9.9 % |
| h1\_mbp\_noninvasive\_min | 9084 | 9.9 % |
| apache\_4a\_hospital\_death\_prob | 7947 | 8.67 % |
| apache\_4a\_icu\_death\_prob | 7947 | 8.67 % |
| h1\_diasbp\_noninvasive\_max | 7350 | 8.01 % |
| h1\_diasbp\_noninvasive\_min | 7350 | 8.01 % |
| h1\_sysbp\_noninvasive\_max | 7341 | 8 % |
| h1\_sysbp\_noninvasive\_min | 7341 | 8 % |
| d1\_glucose\_max | 5807 | 6.33 % |
| d1\_glucose\_min | 5807 | 6.33 % |
| h1\_mbp\_max | 4639 | 5.06 % |
| h1\_mbp\_min | 4639 | 5.06 % |
| h1\_resprate\_max | 4357 | 4.75 % |
| h1\_resprate\_min | 4357 | 4.75 % |
| age | 4228 | 4.61 % |
| h1\_spo2\_max | 4185 | 4.56 % |
| h1\_spo2\_min | 4185 | 4.56 % |
| temp\_apache | 4108 | 4.48 % |
| h1\_diasbp\_max | 3619 | 3.95 % |
| h1\_diasbp\_min | 3619 | 3.95 % |
| h1\_sysbp\_max | 3611 | 3.94 % |
| h1\_sysbp\_min | 3611 | 3.94 % |
| bmi | 3429 | 3.74 % |
| h1\_heartrate\_max | 2790 | 3.04 % |
| h1\_heartrate\_min | 2790 | 3.04 % |
| weight | 2720 | 2.97 % |
| d1\_temp\_max | 2324 | 2.53 % |
| d1\_temp\_min | 2324 | 2.53 % |
| gcs\_eyes\_apache | 1901 | 2.07 % |
| gcs\_motor\_apache | 1901 | 2.07 % |
| gcs\_verbal\_apache | 1901 | 2.07 % |
| apache\_2\_diagnosis | 1662 | 1.81 % |
| apache\_3j\_bodysystem | 1662 | 1.81 % |
| apache\_2\_bodysystem | 1662 | 1.81 % |
| d1\_mbp\_noninvasive\_max | 1479 | 1.61 % |
| d1\_mbp\_noninvasive\_min | 1479 | 1.61 % |
| ethnicity | 1395 | 1.52 % |
| height | 1334 | 1.45 % |
| resprate\_apache | 1234 | 1.35 % |
| apache\_3j\_diagnosis | 1101 | 1.2 % |
| d1\_diasbp\_noninvasive\_max | 1040 | 1.13 % |
| d1\_diasbp\_noninvasive\_min | 1040 | 1.13 % |
| gcs\_unable\_apache | 1037 | 1.13 % |
| d1\_sysbp\_noninvasive\_max | 1027 | 1.12 % |
| d1\_sysbp\_noninvasive\_min | 1027 | 1.12 % |
| map\_apache | 994 | 1.08 % |
| heart\_rate\_apache | 878 | 0.96 % |
| arf\_apache | 715 | 0.78 % |
| intubated\_apache | 715 | 0.78 % |
| ventilated\_apache | 715 | 0.78 % |
| aids | 715 | 0.78 % |
| cirrhosis | 715 | 0.78 % |
| diabetes\_mellitus | 715 | 0.78 % |
| hepatic\_failure | 715 | 0.78 % |
| immunosuppression | 715 | 0.78 % |
| leukemia | 715 | 0.78 % |
| lymphoma | 715 | 0.78 % |
| solid\_tumor\_with\_metastasis | 715 | 0.78 % |
| d1\_resprate\_max | 385 | 0.42 % |
| d1\_resprate\_min | 385 | 0.42 % |
| d1\_spo2\_max | 333 | 0.36 % |
| d1\_spo2\_min | 333 | 0.36 % |
| d1\_mbp\_max | 220 | 0.24 % |
| d1\_mbp\_min | 220 | 0.24 % |
| d1\_diasbp\_max | 165 | 0.18 % |
| d1\_diasbp\_min | 165 | 0.18 % |
| d1\_sysbp\_max | 159 | 0.17 % |
| d1\_sysbp\_min | 159 | 0.17 % |
| d1\_heartrate\_max | 145 | 0.16 % |
| d1\_heartrate\_min | 145 | 0.16 % |
| icu\_admit\_source | 112 | 0.12 % |
| gender | 25 | 0.03 % |

### Inference

The labs and vitals attributes such as h1\_bilirubin, h1\_lactate, h1\_albumin, h1\_calcium, h1\_creatine, h1\_sodium, d1\_lactate, d1\_arterial, fio2, paco2 and few more contains more than 70% missing data. A few demographic and APACHE attributes contain missing values as well, but they are essential for model building and are in the range of 4.6-0.03% and 64-0.7% respectively. Hence, these attributes will be processed in the later stages.

# Data Preprocessing

## Missing Value Imputation

### Attributes that contain more than 70% missing values

The number of attributes that had more than 70% of missing values is 55. As mentioned, the attributes belonged to the labs, vitals and apache covariate groups. Since a large amount of data is missing, the attributes are removed from the dataset.

### Attributes that contain less than 70% missing values

The missing values are handled by imputing mean/mode depending on their types using sklearn SimpleImputer library.

For categorical attributes, the missing value was replaced with most frequently occurring value, whereas for numerical attributes, depending on the skewness of the attribute, mean or median was added.

### Irrelevant attributes

By researching more into the attributes, we had decided to remove few attributes as they do not contribute to the survival rate of a patient. The attributes were mainly identifiers, apache covariate, labs and lab blood gas variables, such as albumin\_apache, apache\_post\_operative, bilirubin\_apache, encounter\_id, fio2\_apache, hospital\_id, icu\_id, paco2\_apache, paco2\_for\_ph\_apache, pao2\_apache, patient\_id, ph\_apache, pre\_icu\_los\_days, readmission\_status and urineoutput\_apache

## Outlier Detection and Treatment

The outliers were detected using the IQR method and boxplots were used for visualization.

For the treatment, we had initially used IQR and Z scale methods were applied to treat the outliers depending on their skewness. On further research, we used PowerTransform to treat the outliers as it stabilizes the variance or in other words reduces the skewness. By doing so, it also does feature scaling of the data.

# Analysis of Dataset

## Univariate Analysis

Chart, waterfall chart

Description automatically generatedChart, waterfall chart

Description automatically generatedChart, bar chart, histogram

Description automatically generated

|  |  |
| --- | --- |
| Table  Description automatically generated | A picture containing bar chart  Description automatically generated |
| Table  Description automatically generated |  |

### Inferences

* The target variable hospital death has an imbalanced proportion as there are more survivals than deaths.
* From the ethnicity graph it is seen that patients that were admitted were Caucasians in majority. And the least admitted ones are native Americans.
* There isn’t much of a difference between male and female patients.
* In hospital admit source, Emergency Room and Floor are maximum.
* Admission is the most popular one in ICU stay type.
* The unit that had a greater number of patients admitted is Med-Surg.
* Almost all of them have significant number of cases except for Gynecology and Hematology
* The range of age had been admitted ranges from 16 to 90. The age group that is highly likely to be admitted is in the range 50-85, in which the age 85 is the highest number of patients admitted.
* In terms of bmi, it ranges from 15-70. Patients with bmi in the range 20-35 are high in number, especially with bmi 30.
* For height, it ranges from 137 to 195, where the patients with height range 160-180 are higher in number in admissions.
* For weight, patients with weight in the range of 62-100 are more in number.

## Bivariate Analysis

### Numerical vs Categorical

Chart, box and whisker chart

Description automatically generatedChart, bar chart, box and whisker chart

Description automatically generated

### Categorical vs Categorical

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, bar chart

Description automatically generatedChart

Description automatically generated

### Numerical vs Numerical

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generatedChart, line chart

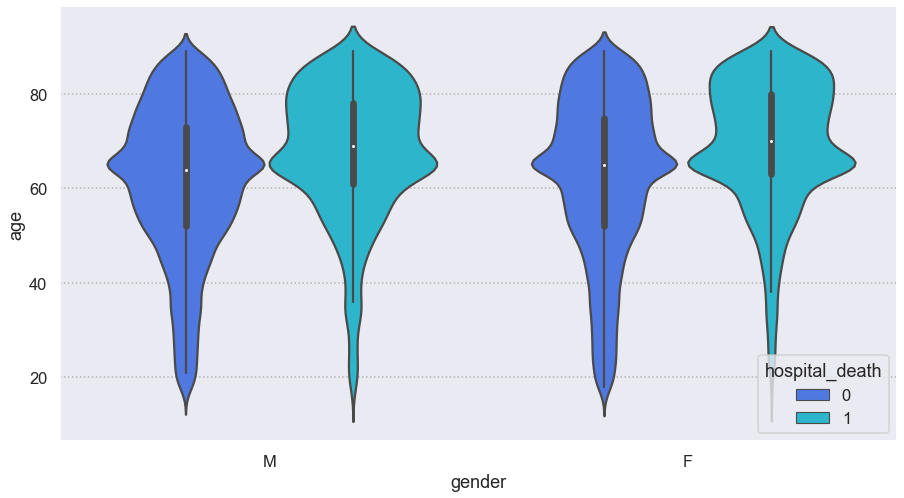
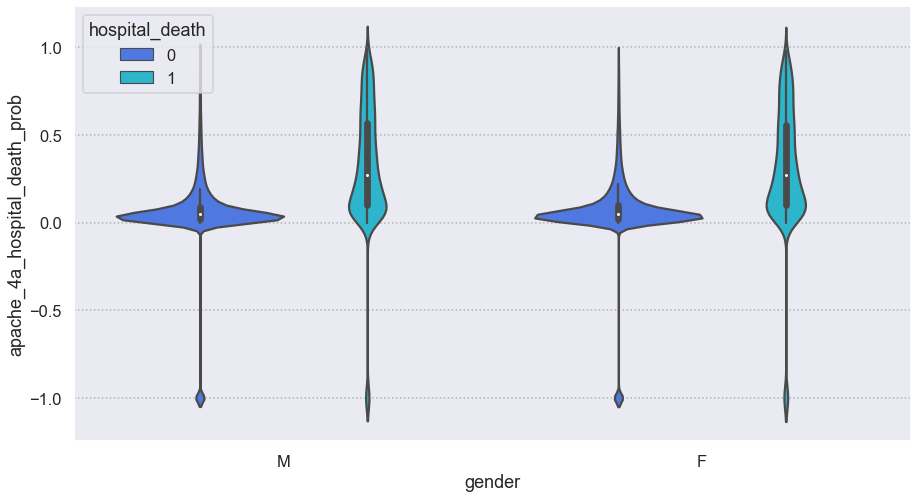
Description automatically generatedChart, line chart

Description automatically generated

### Inferences

* There’s no difference in the age groups for both male and female patients
* For patients with higher heart rate have a lower chance of survival
* It is seen Accident and Emergency cases have higher chances on both survival and death
* Cardiovascular bodysystem has the similar conclusion as above.
* As age increases, the apache death probability also increases.

## Multivariate Analysis



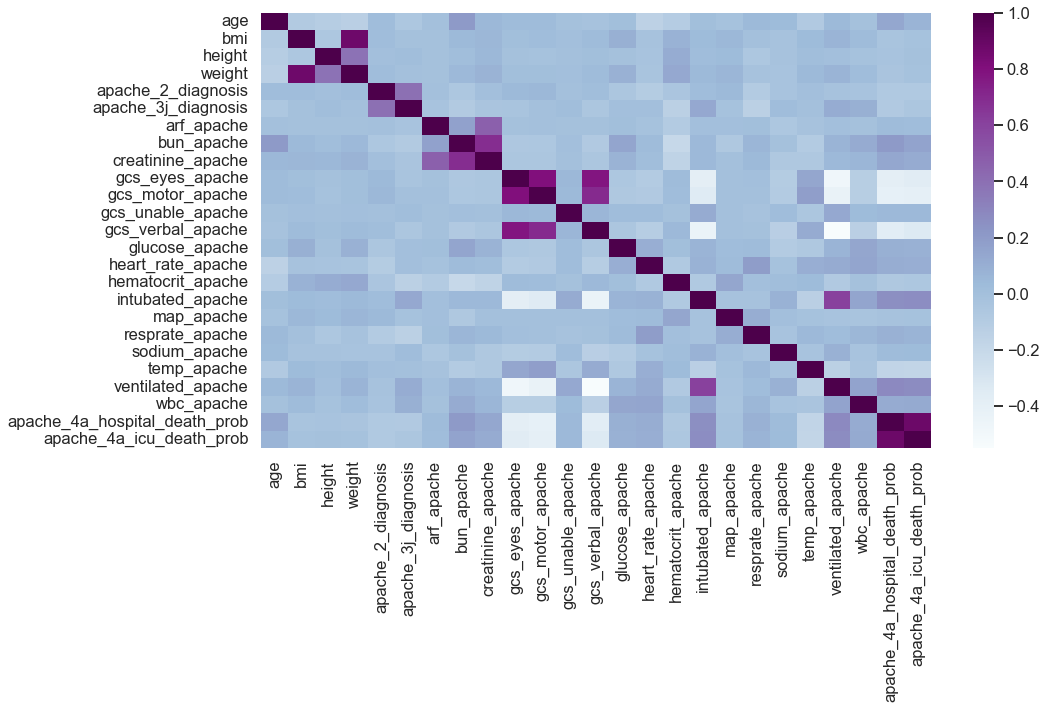
A picture containing table

Description automatically generated

### Inferences

* The apache death probability is a bit higher for Female when it comes patients not surviving.
* The common age for both genders for possible death is the same.
* The death probability is higher in the case of Cardiovascular bodysystem.

## Multicollinearity



### Inferences

There are only a handful of variables that show some collinearity.

# Hypothesis testing

## Chi-Square Test

The chi-square test is used to check whether the variables are dependent on one another. The categorical variables are compared with the target variable, hospital death. From the table, we see that icu\_admit\_source and hospital\_death are independent of one another, and hospital\_death is dependent on the other variables.

## Mann-Whitney U Test

The first step that was taken was to check if the data is normally distributes using the Shapiro Wilk test. From the test, we see that all the variables did not have a normal distribution, hence we used the Mann-Whitney test if there are differences between two independent outcomes, i.e. it checks whether the mean of a variables varies for different outcomes of hospital\_death. From the results of the test, it is seen that all the variables had differences for different outcomes of hospital\_death.

# Feature Engineering

## Feature Encoding

The function get\_dummies from pandas is used to create indicator/dummy attributes except for the target variable hospital\_death.

## Feature Selection

The variation inflation factor method was used to determine the features involved in multicollinearity. These features were removed from the dataset.

## Feature Scaling

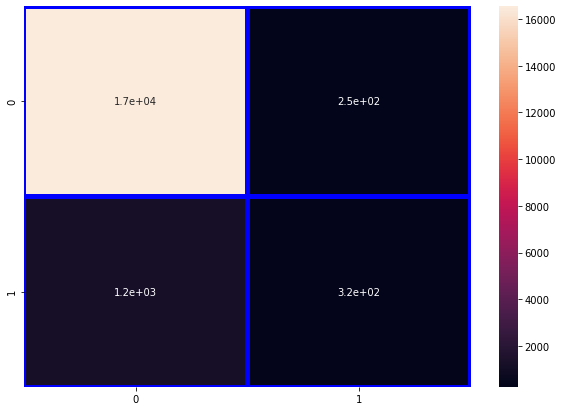
In the outlier treatment, we used Power Transform to reduce the skewness by scaling the data.

# Model Building

We first began trying variety of base models and comparing the results to help choose a model giving a good recall score. In the following models, we will see the confusion matrix, classification report and the ROC curve. The data has been split into train and test sets.

## Logistic Regression

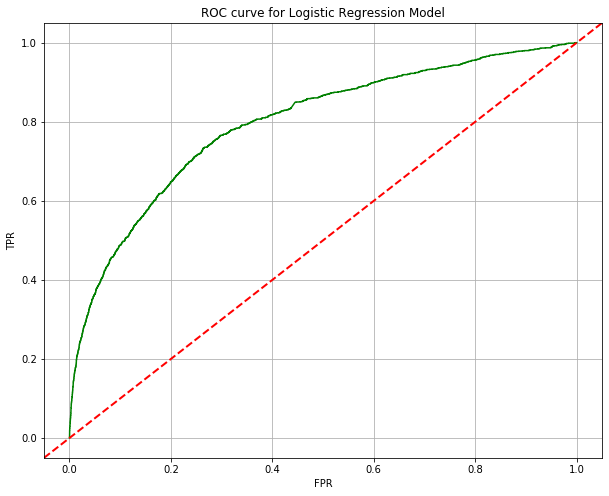
The model building first begins with logistic regression model. After generating the confusion matrix, we calculated the optimal threshold value using the Youden’s Index. After getting the optimal threshold value of 0.51, we generated the following reports once more.



From the confusion matrix, it predicted that 17000 patients will survive, and 320 patients will not survive. However, it incorrectly predicted that 250 patients will not survive and 1200 will survive.

Table

Description automatically generated



### Inferences:

* Cross Entropy for the Logistic Regression Model is 2.77
* ROC AUC Score for the Logistic Regression Model is 79.55
* The Model Accuracy for the Logistic Regression Model is coming out to be around 92%.
* f1 weighted average for the Logistic Regression Model is around 90%.
* Specificity : 98%
* Sensitivity : 20.74%

## K-Nearest Neighbor Classification Model

A screenshot of a computer

Description automatically generated with low confidence

Table

Description automatically generated

Chart, line chart

Description automatically generated

### Inferences:

* Cross Entropy for K-Nearest Neighbours Model is 3.10
* ROC AUC Score for the K-Nearest Neighbours Model is 66.84
* The Model Accuracy for the K-Nearest Neighbours Model is coming out to be around 91%.
* f1 weighted avergae for the K-Nearest Neighbours Model is around 89%.
* Specificity : 98.44%
* Sensitivity : 10%

## Decision Tree Classification Model

Chart, treemap chart

Description automatically generated

Table

Description automatically generated

Chart, line chart

Description automatically generated

### Inferences

* The Cross entropy for the Decision Tree model is 4.19
* ROC AUC Score for the Decision Tree Model is 63.89
* The Model Accuracy for the Decision Tree Model is coming out to be around 88%.
* f1 weighted avg for the Decision Tree Model is around 88%.
* Specificity : 92%
* Sensitivity : 34%

## Naïve Byes Classification Model

Chart, treemap chart

Description automatically generated

Table

Description automatically generated

Chart, line chart

Description automatically generated

## Base Model Comparison

Graphical user interface, application

Description automatically generated with medium confidence

### Inferences

Overall Accuracy Score:

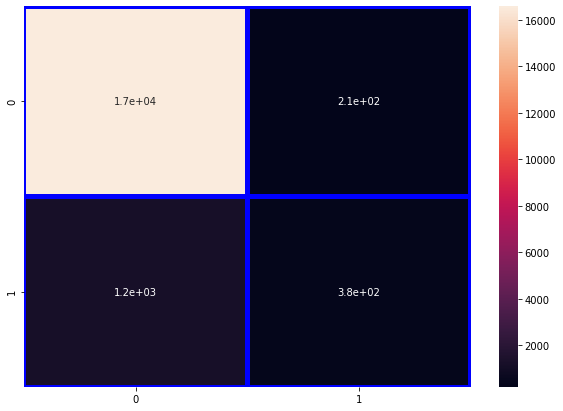
1. Logistic Regression Model has the highest overall accuracy of about 92%.
2. Decision Tree Model yields the lowest overall model accuracy of about 88%.

Overfitting/Underfitting:

1. All the classification models exhibit overfitting of the trained data with respect to the test data.
2. The model accuracy for train data and test data for both Logistic Regression Model and K-Nearest Neighbors Model has very less overfitting.
3. As observed, the model accuracy for train data and test data for the Decision Tree Model has a considerably high difference in accuracies which can be considered a high overfitting condition in comparison to other models.

# Model Optimization

## Random Forest



Table

Description automatically generated

Chart, line chart

Description automatically generated

## Hyperparameter tunning

A good choice of hyperparameters can really make a model succeed in meeting desired metric value or on the contrary it can lead to a unending cycle of continuous training and optimization. Hence, we have used Grid Search Cross Validation technique for Hyperparameter tunning the models where the Cross Validation method considered is 10-Fold Cross Validation.

### Logistic Regression

Table

Description automatically generated

Table

Description automatically generated

### Random Forest Model

Table

Description automatically generated

Table

Description automatically generated

# Imbalanced Target

As seen in univariate analysis, our target variable is imbalanced. Even though it makes sense that there are high survivals than deaths, to improve our model, we use SMOTE analysis it correct the imbalanced data.

## SMOTE Analysis

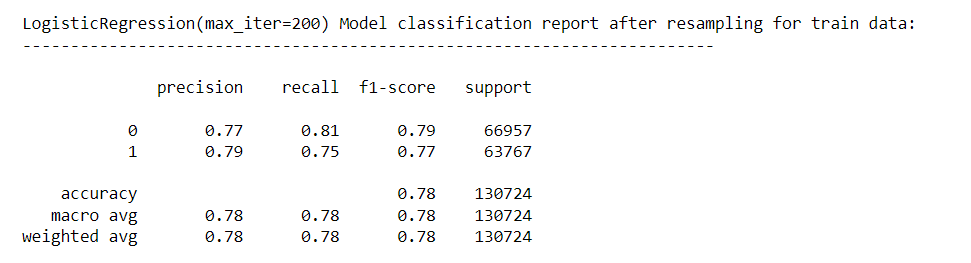
The minority class, i.e., the patients that have not survived, gets oversampled. After doing so, the ratio between survival and death became 51:48.

A picture containing logo

Description automatically generated

We performed the same steps of model building, and compared the scores once again.

## Logistic Regression



Table

Description automatically generated

## Decision Tree

Graphical user interface

Description automatically generated with low confidence

Graphical user interface, application, table

Description automatically generated

## Random Forest

Graphical user interface

Description automatically generated with medium confidence

Table

Description automatically generated

## Inference

From the comparisons above, it is seen that precision and recall has improved for Random Forest compared to the one built with the original data,

# Feature Extraction

The important features were determine with the help of Recursive Feature Elimination using the Random forest model. The following features were selected

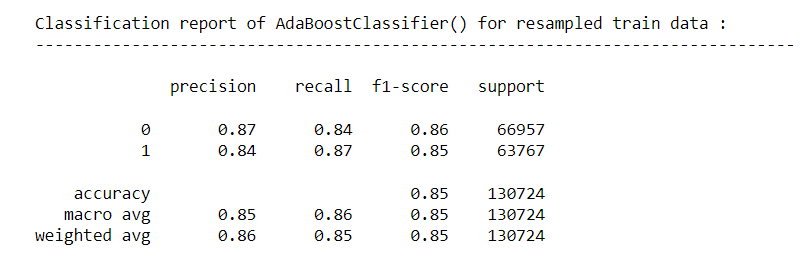
* map\_apache
* apache\_4a\_icu\_death\_prob
* apache\_2\_diagnosis
* apache\_3j\_diagnosis
* apache\_4a\_hospital\_death\_prob
* bun\_apache
* creatinine\_apache
* wbc\_apache
* gcs\_verbal\_apache
* glucose\_apache
* diabetes\_mellitus
* resprate\_apache

## Inference

* The Sensitivity and Specificity of both Logistic Regression model and Decision Tree model has increased significantly from the base models resulting in an improved weighted-avgerage of Precision and Recall rates.
* The Decision Tree Model has a better overall performance than the Logistic Model for the new dataset obtained after recursive feature elimination and resampling the data.
* The Random Forest Model has not shown any major improvement from the previous model but has the best overall performance as compared to the Logistic Regression and Decision Tree model.
* Random forest has been considered.

# Bagging and Boosting Algorithms

## AdaBoost Model



Table

Description automatically generated

Graphical user interface, text, application

Description automatically generated

## GradientBoost Model

A picture containing table

Description automatically generated

Table

Description automatically generated

Graphical user interface, text, application

Description automatically generated

## XGBoost Model

A picture containing table

Description automatically generated

Text

Description automatically generated

# Model Comparison

Graphical user interface

Description automatically generated

Based on the comparison we can conclude that:

* XGBoost Model and Random Forest Model have the approximately overall accuracy of 95% on test data.
* XGBoost has less overfitting of train data as compared to Random Forest Model.
* XGBoost has a slightly lesser Bias error and variance error than Random Forest Model.
* f1-score weighted for XGBoost Model: 96%
* AUC score for XGBoost Model: 0.98

# Final Model

From the comparisons, we have decided to use XGBoost Model. The following are it evaluation metrics

Table

Description automatically generated

Chart, line chart

Description automatically generated

Graphical user interface, text

Description automatically generated

# Conclusion

We have decided on XGBoost Model as our final model. The important features gives some idea on what factors to look at more closely when determining the survival rate of the patient. Some of the features are diabetes, respirate, glucose, Apache diagnosis. The limitation is that there it is still not very precision in determining the survival rate but out of all the models, XGBoost showed the best result.