University of Essex

Department of Mathematical Sciences

MA981 DISSERTATION

Prediction of Sepsis from Clinical Data using Machine Learning Algorithms: Gaussian Naive Bayes, Random Forest and Decision Tree

Archit Dubey
2211582

Supervisor: **Dr Andrew Harrison**

November 24, 2023
Colchester

Contents

List of Figures

List of Tables

Chapter 1

Abstract

A comprehensive review for Machine Learning (ML) model over anticipating sepsis which requires timely intervention is studied in this dissertation. Three Machine Learning algorithms; Gaussian Naive Bayes, Random Forest, and Decision Tree classifiers are assessed. Every method is investigated employing a range of datasets which include Imbalanced, Undersampled and SMOTE oversampled data. Performance indicators used are accuracy, recall, F1 score, precision, and Area Under the ROC Curve (AU ROC) which will help to to assess the successful these models are.

Although Gaussian Naive Bayes works exceptionally well in unbalanced datasets it has problems with recall and AU ROC particularly in circumstances with low sample counts. High precision as well as accuracy can be observed from Random Forest especially for negative scenarios, however oversampling may lead to excessive fitting. Decision trees have an adaptable choice as they regularly show excellent accuracy and precision when using multiple methods of sampling. It highlights the importance to choose a classifier and sampling plan which is suitable for the task at on hand. As a way to enable greater adoption of such models in medical practise subsequent studies will focus on creating advanced machine learning algorithms over real-time risk evaluations in Intensive Care Unit (ICU) while considering confidentiality and ethical issues.

Keywords: Machine Learning, Sepsis anticipation, Gaussian Naive Bayes, Random Forest, Decision Tree classifiers, Imbalanced, Undersampled, SMOTE oversampled, Accuracy, Recall, F1 Score, Precision, AU ROC

CHAPTER 2

Introduction

Sepsis as a term is used for describing a pathological infection and physiological alterations called as the Systemic Inflammatory Response Syndrome (SIRS). This reaction leads to physiological changes happening at the capillary endothelial level. This process's initial signs continue to be ambiguous and because of this reason, medical care usually undervalues it. Nevertheless, lowering sepsis mortality needs early diagnosis about this disease. According to clinical research sepsis starts with a infection and progresses to sepsis, severe sepsis which result organ failure and septic shock in some situation. Sepsis has become much more common over the past few decades, and it seems to be becoming worse with time as more people die from it even though total in-hospital mortality is declining. Increased mortality in severe sepsis and is linked to advanced age, underlying comorbidity and organ failure. [1]

Severity received the label of "the quintessential medical disorder of our time" as it was becoming a major cause of mortality for the patients in hospitals, it also regularly leads to notable advancements in the treatment of patients with a variety of ailments. [2] Throughout numerous years, the term of sepsis has changed. Sepsis and septic shock are at present among the 13 most common causes of death in the United States of America. Each year, about 500,000 new cases take place with mortality around of 35 percent. [3]

Severe causes organ failure by a uncontrolled host response to the viral infection is now recognised as sepsis. Severe disruptions with the metabolic, and circulatory systems are the signs of sepsis known as "septic shock." It's more likely to end in death as compared to sepsis alone. Over the previous 40 years, there has been a steady rise in the age-adjusted death rate between 0.5 to 7 per 100,000 cases of sepsis. [3] Based on the more thorough Global Burden of Diseases survey, sepsis persists as a widespread disease having roughly 50 million cases worldwide per year. The common causes of sepsis are bacterial infections of the digestive and respiratory systems, though the location and maturity of the diseased site and pathogenic organism can vary. [2]

In 2017, sepsis was projected to result in an alarming 11 million deaths, or a age-standardized mortality of 148 per 100,000 people. The following amounts to over 20 percent of the fatalities globally. The statistics are alarming: a third of those in sepsis requiring to be temporarily admitted to critical care facilities cannot remain alive for 30 days. The number and kind of organ dysfunctions, age, and comorbid condition all affect mortality rates. Sepsis survivors face several difficulties. One in six fails to survive their initial year, and more than 50 percent have to return to the hospital a minimum of once a year. [2]

In addition, although more than 200 Randomized Controlled Trail (RCT) as well as more than 30 years for research, no treatment has ever been demonstrated to continually extend the survival of sepsis patients. The cornerstone of treating sepsis today includes offering psychological assistance in alongside early antibiotic delivery, organ dysfunction the leadership team, source control, and resuscitation. The emergency care community is right now investigating the possibility of categorised patients with sepsis into smaller groups according to identifiable features or particular pathobiological deviations. [2]

Three clinical syndromes that characterise a progressive rise in the systemic inflammatory response to infection have replaced the relatively nebulous term "sepsis" in recent years. The first diagnosis is systemic inflammatory response syndrome (SIRS) which is followed by septic shock, severe sepsis, and sepsis. Providing more accurate definitions for clinical and scientific purposes was the intention behind the introduction of these words. [3]

Researchers are right now emphasising on understanding immunological dysregulations with sepsis. They are looking extensively into the mechanisms which maintain homeostasis during as well as after bacterial contacts, particularly with a focus on immunological imbalances which are typified by simultaneous aberrations which are proinflammatory and immune suppressive. This is important to remember that infections with viruses, fungi, or parasites can cause sepsis. The idea of dysregulated host reaction in sepsis encompasses such non-immune changes. Because sepsis continues to be an ill-defined illness, real-time immunological profiling could grow possible with the combination of cutting-edge technologies and bedside computing support, which could influence patient outcomes as well as treatment modalities. [2]

CHAPTER 3

Objective and challenges of project

Developing a model using the Machine Learning (ML) which shall predict Sepsis for persons in Intensive Care Unit (ICU) is the goal of this project. Here 3 algorithms like Gaussian Naive Bayes, Random Forest and Decision Tree are going to be used seeking to enhance the treatment result of this condition by allowing more exact sepsis prediction.

3.1 Supervised Machine Learning Algorithms

Supervised machine learning (ML) models are utilized for Healthcare industry where data is clearly labelled with a established final result. For example, the existence of medical issue like diabetes or hypertension. The precision and dependability for the outcome flags the data is crucial parts among these models. Any bias of these labels have the chance to negatively impact the models efficacy along with restricting the application beyond original training dataset. The precise description of this problem statement was the first stage in the well established method for constructing a supervised machine learning model in the Healthcare sector. Following the identification of relevant information this phase is important because it involve processing of data operations like data manipulation, picking variables, and redundant data removal. [4] A sample that has data is chosen a algorithm is selected then the model is trained and the efficiency is determined using methods like Accuracy of model and AU - ROC. Given the repetitive nature of this method choosing and adjusting procedures is often required in order to identify optimal set of variables that shall enhance the models predictive power. In a bid to shorten the period and computational demands associated with hyperparameter tuning study in this area recommends to set values in various parameters. [4]

Supervised ML Algorithm generate a function through analysing the input data that has defined result value then it will be used to identify outcome of new input. Supervisor here is the teacher who is available when algorithm is being taught how to predict the Classifier. In this method of machine learning algorithm gives prediction based on initial data and accept changes from teacher. The procedure to learn is then repeated unless learning algorithm achieves desired degree of accuracy. [5]

Supervised Machine Learning Algorithm are classified in 2 categories ; Classification and Regression. Some of them popular are Linear Regression, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Naive Bayes and K Nearest Neighbor. [5]

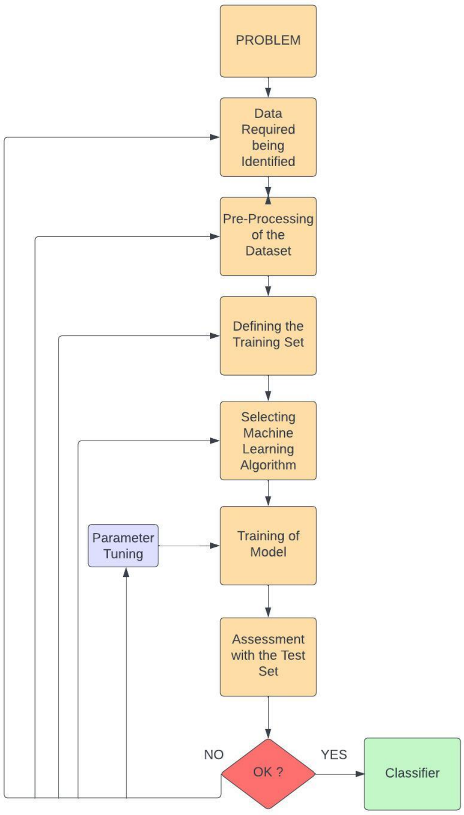

PROBLEM
Data Required being Identified
Pre-Processing of the Dataset
Defining the Training Set
Selecting Machine Learning Algorithm
Parameter Tuning
Training of Model
Assessment with the Test Set
NO
OK ?
YES
Classifier

Figure 3.1: The Development of Supervised Machine Learning Code [6]

3.1.1 Supervised Machine Learning for Sepsis Prediction

This explores supervised algorithms which are going to be used for prediction of sepsis namely Gaussian Naive Bayes, Random Forest and Decision Tree. For selecting model which is most suitable in sepsis prediction it becomes essential us to assess the performance of these models. The model predictive accuracy, precision, recall, F measure and Area under the Receiver Operating Characteristic (ROC) curve are among parameters upon which the best fitting algorithm will be analyzed.

3.2 Building a Machine Learning Model to Predict Sepsis

Errors, like missing values and wrong spelling, may appear with patient data gathered by the bedside. The dataset included quite a few missing values, largely as a result of the human filling those records throughout continued treatment [7]. So as to make sure the model built is accurate and suitable for predicting the sepsis, additional efforts involved feature selection and figuring out the columns in the dataset were most relevant to it.

CHAPTER 4

Literature Review

4.1 Defining Sepsis

Sepsis that can be described as an inflammatory response to infections that is dysregulated which ends in organ damage, is a global cause of mortality. Results are substantially enhanced by prompt action, especially if antibiotics are employed. The difficulty is in recognising it earlier since symptoms tend to be ambiguous and may contribute to higher mortality rates. Sepsis typically gets diagnosed and treated in urgent care centres, where busy and sometimes understaffed situations make it hard to recognise and manage this severe illness in an time effective way. It remains essential to handle these problems if sepsis is to have substantially treated. [8]

Sepsis 1.0: Based on two or more Systemic Inflammatory Response Syndrome (SIRS) criteria along with a possible or confirmed infection source sepsis has been defined as inflammatory response to infection. Persistent hypotension or hyperlactatemia despite fluid resuscitation was indicative of septic shock. [9]

Sepsis 2.0: The modification maintained original definitions however introduced the requirements for Sequential Organ Failure Assessment (SOFA) for identifying organ dysfunction. The modification shifted the definition of "severe sepsis" to incorporate organ dysfunction. [9]

Addressing the problems associated with Sepsis 1.0, Sepsis 3.0 gave attention to the little specificity and sensitivity of SIRS criteria. The focus of Sepsis 3.0 became primarily upon the existence of organ dysfunction as well as the necessity of identifying sepsis properly from additional serious illnesses which additionally indicate evidence of inflammation. [9]

Table 4.1: Comparison of older and new definitions for the spectrum of sepsis, severe sepsis and septic shock [9]

	Sepsis-2 Definitions	**Sepsis 3.0 Definitions**
Sepsis	Sepsis 2 SIRS criteria	Increase in SOFA score ≥ 2 from baseline OR qSOFA ≥ 2
Severe Sepsis	Severe sepsis: Sepsis AND Organ Faliure	(Not applied here)
Septic Shock	Septic shock: Sepsis AND Hypotension despite fluid resuscitation	Sepsis AND Vasopressor requirement despite fluid resuscitation

4.2 Clinical Scoring system for Sepsis

Clinical scoring systems depend on vital signs and data and are frequently used to identify sepsis. These techniques can assess disease severity quickly in Emergency Departments via only just a few of clinical markers. A selection of scoring systems, including Modified Early Warning Score (MEWS) score, the National Early Warning Score (NEWS) and Quick Sequential Organ Failure Assessment (qSOFA) have been developed and proven effective in emergency situations. These systems evaluate the condition generally by identifying deviations in crucial parameters. [10]

4.2.1 Sequential Organ Failure Assessment (SOFA) Score

A advanced scoring method largely utilised in the Intensive Care Units (ICU) is the SOFA score. It analyses each of the six organ systems: cardiovascular, neurological, hepatic, coagulation, renal and respiratory. It employs the ratio of the arterial oxygen partial pressure (PaO2) to fractional inspired oxygen (FiO2) in the respiratory system, the serum bilirubin level over the evaluation for liver function, the number of platelets for coagulation, the serum creatinine level or the urine output for renal function, and the

Glasgow Coma Scale score for neurological status. The total score may range between 0 to 24, having each system getting a score between 0 (normal) and 4 (severe degree of dysfunction) raised scores have been linked to increased mortality as well as more chronic organ dysfunction. [11]

4.2.2 Quick Sequential Organ Failure Assessment (qSOFA) Score

The purpose of Quick Sequential Organ Failure Assessment (qSOFA) score is to quickly identify patient with suspected infections who are likely to experience adverse outcomes such as sepsis. Because this scoring system is simple to use and does not require laboratory testing it is especially helpful outside of Intensive Care Unit (ICU). It evaluates 3 parameters: elevated breathing rate, hypotension, and changed mental state. [12]

Glasgow Coma Scale (GCS) is used to assess altered mentation; a score of less than 15 suggests possible neurological dysfunction which is frequently an early indicator of sepsis induced organ failure. Systolic blood pressure of 100 mmHg or less is considered hypotension and indicates potential cardiovascular compromise which is common in sepsis and severe infections. The respiratory rate is last factor to be examined. A rate of 22 breaths per minute or higher indicates a possible case of respiratory distress or metabolic acidosis both of which are linked to sepsis. [12]

Patients who meet at least 2 of these criteria and have positive qSOFA score are at risk of longer ICU stays or higher death rates. Instead of being a diagnostic tool it is a screening tool that encourages additional assessment of patients who are at risk. Compared to other complex scoring systems qSOFA is less sensitive but its simplicity facilitates speedy decision making. [12]

4.2.3 Systemic Inflammatory Response Syndrome (SIRS) criteria

A systemic inflammatory response that may arise from a number of infectious as well as non-infectious causes, such as sepsis, is identified employing the criteria of Systemic Inflammatory Response Syndrome (SIRS). At least 2 of the 4 requirements listed below must be met for SIRS to be present:
(a) an elevated heart rate exceeding 90 beats per minute is one of the key indicator. This increase shall be a response for the body's inflammatory state.
(b) the respiratory system's involvement is marked by either an increased respiratory

rate of more than 20 breaths/min.
(c) the criteria consider with white blood cell count abnormalities, either an elevation above 12,000 cells, a decrease below 4,000 cells, or presence of more than 10 percent immature neutrophils.
(d) abnormal body temperature is a criterion which encompasses both fever (temperature above 38 C or 100.4 F) and hypothermia (temperature below 36 C or 96.8 F) indicative of the body systemic response to the inflammation. This criteria of SIRS is helpful to detect sepsis at its onset but they aren't distinctive to this illness and are additionally brought on by autoimmune disorders. SIRS should therefore be interpreted within the context of a wider clinical setting incorporating note of whether or not a infectious source is in place. [13]

4.3 Research Paper review on Sepsis Prediction

[14] This paper primarily uses data from 2019 PhysioNet Competition but it also refers to MIMIC-III dataset in some parts, especially when evaluating the effectiveness of a sepsis prediction model on an infection-prone ICU patients that only uses objective features. A total of 40,336 ICU patients' clinical data of two different hospitals make up the Dataset. 3,932 of these patients had sepsis. A "nan" (not-a-number) marker is present in dataset to indicate the absence of any values. The classification of patients as septic based on specific guidelines, particularly those related to infection diagnosis and modifications within the SOFA score, had a component of this Dataset.
Going forward with the approach main machine learning algorithm (ML) used in this study is eXtreme Gradient Boosting(XGBoost). 5-fold cross-validation was the method utilized for assessing the model. 85 percent of the data distribution shall be split among training and testing tasks with rest 15 percent designated only for testing.Examining the results more closely, three modelsâthose with 168 features, 102 features, and 37 feature were examined. There are noticeable differences in performance between these 3 models. Reducing the number of features revealed a substantial decrease in operational metrics. It is found that 168-feature model held a specificity of 0.7687 and a sensitivity of 0.7449, the 102-feature model had a 0.715 for both, and the 37-feature model had a 0.643 for both. These results assist to further clarify the corresponding specificities as

well as sensitivities for each of these models. 168-feature model, the 102-feature model, and the 37-feature model had Positive Predictive Values (PPV) of 0.0560, 0.044 and 0.032, respectively. Major drops in performance metrics including AU-ROC, AUPRC, F1-Score, and Utility, were seen when features were over time eliminated. That was another important finding. [14]

According to the paper XGBoost model performs far less well if it solely uses vital signs, especially when applied to a diverse group of ICU patients. It highlights the challenge of getting lab data and opinions from experts beyond intensive care unit (ICU) and speculates that models that depend solely on objective measurements might be less effective. Furthermore, as they heavily depend on medical treatments, models with several characteristics but more effective may not serve suitable outside of hospital settings. [14]
[15] aimed to evaluate the efficacy of machine-learning models using qSOFA variables in predicting 3 day mortality in the Emergency Department (ED) for patients with suspected infection. The purpose of this study was to discover if these models had performed better in precision than regular qSOFA score.
Hallym University Medical Center's Smart Clinical Data Warehouse (CDW), that is based on the QlikView Elite Solution, contributed the clinical data for the purpose of the research. The information came from the medical files of patients over the age of 18 with suspected diseases from four EDs in Korea during January 2016 and December 2018. Min-Max scaling was implemented to standardise the datasets. 23,587 of the 447,926 individuals who came in during this time underwent testing for possible infections. The evaluation dataset came from Chuncheon Sacred Heart Hospital (n = 4,234), while the training-validation datasets came from Dongtan Sacred Heart Hospital, Kangnam Sacred Heart Hospital, and Hallym University Sacred Heart Hospital (n = 19,353).

Considering there were significantly fewer fatalities and ICU inpatients in the sample used for training than survivors and general-care inpatients, there was an imbalance in the data that might have had an effect on the accuracy of the model. This discrepancy had been corrected using the artificial minority over sampling approach, which created a 1:1 ratio between the XGB and LGBM models. The proportionate random forest approach was thereafter employed to fix the other problems. Three different models

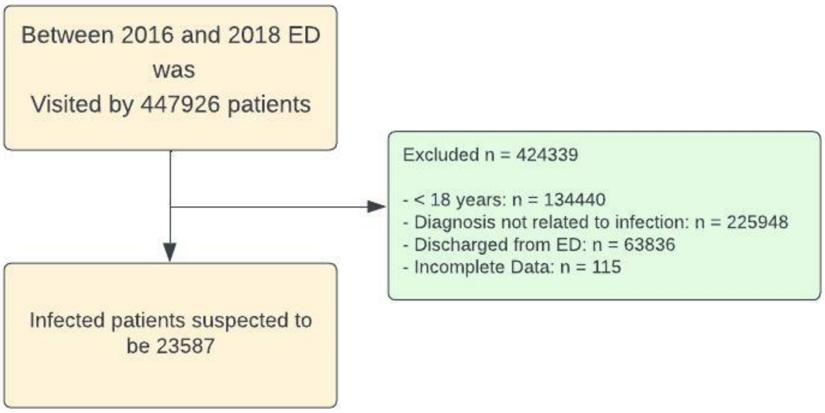

Between 2016 and 2018 ED
was
Visited by 447926 patients
Excluded n = 424339
- < 18 years: n = 134440
- Diagnosis not related to infection: n = 225948
- Discharged from ED: n = 63836
- Incomplete Data: n = 115
Infected patients suspected to be 23587

Figure 4.1: Flowchart for number of patients 'n' in Dataset [15]

were trained with simple hyperparameters and examined with a set of parameters for validation to figure out which of them worked most effectively. Grid search and the 5-fold cross-validation approach were utilised for performing hyperparameter alterations. A series of tests were carried out to confirm the model's ultimate form. Given enhanced robustness, the information has been divided into validation, training, and test sets. The data sets have been splited in the following order, taking into account a 9:1 ratio for training-validation: 74 percent for training, 8 percent for validation, and 18 percent for testing. The outcomes showed that, having an AU-ROC of 0.86, its qSOFA-based approach surpassed the regular qSOFA scoring. In-hospital mortality was projected by the model to be 0.75, while three-day ICU admission and cumulative ICU admission were anticipated to be 0.79. This framework has been developed in order to assist clinicians in emergency rooms in hospitals decide about sepsis treatment. [15]

Although the paper does not state its weaknesses some potential drawbacks are implied like the retrospective study design narrow focus on 3 day mortality, potential hospital biases and specific geographically focusing Korea. [15]

[16] main objective of this paper is to use machine learning algorithms to enhance the on time detection of sepsis. The importance for early sepsis identification is highlighted by the writers of this paper, especially in Intensive Care Unit (ICU). They address the way to approach this issue of inconsistent sample in medical data and suggest employing

spline-based interpolation techniques to impute unreported samples and normalise the timeframe of parameters.

Data coming from 40,336 patients was included in this study which included test results, vital signs, sepsis status and symptoms appearing time. The datasets times and irregular gathering of data intervals varied. This examination for missing values, preparing the data, and length distribution was the main goal of the scholars. By assigning 10 [ercent] of the data for evaluation, 10 percent for validation, and 80 percent for training, they used k-fold cross-validation. Following log-transforming all continuous variables, z-score normalisation were used. Uneven rates of sampling were fixed via individually cubic Hermite interpolation and zeroes were added to the remaining missing values after z-score normalisation. [16]

Using the 40 variables for the neural network models 18 statistical features was taken in account for feature selection. A wide range of approaches was evaluated such as Support Vector Regression/Machine(SVM), Random Forests, Hidden Markov Models, and Sparse Quantile Regression. An CNN-LSTM deep learning structure which had been before trained on the ImageNet dataset was used as well within the study together with a KNN-GAK system that combined K-nearest neighbours with Triangular Global Alignment Kernels and Dynamic Time Warping for the diagnosis of sepsis. The study determined the neural networks, Random Forests and Naive Bayes algorithms all of the 3 performed well in terms of accuracy, sensitivity, and specificity. The effectiveness for limited quantile regression to recognise sepsis and the capability to withstand overfitting were significant. The neural network LSTM-CNN suggested excellent effectiveness, attaining a challenging score of 0.076. [16]

However, this research had some weaknesses. The triangle inequality criterion wasn't met by the basic DTW distance utilised during KNN-GAK that's reduced the accuracy of the distance measurement. The memory consumption and runtime of KNN-based approaches limited their scalability, making them less useful for real-time online forecasts. Moreover, several types of classifiers revealed poor specificity (< or = 5 percent) for recognising sepsis. [16]

[17] Fourth Paper main goal had been to find out degree to which machine learning anticipates sepsis in patients who are critically ill in the Intensive Care Unit (ICU). This

is into great length regarding the physiologic information on patients treated in the ICU causing a differentiation among people who had sepsis versus those who did not. The core objective was to figure out whether or not each of the important variables may function in the initial indication of sepsis as well as assist in avoiding it. The papers methodology is split into two parts: First it groups patients through groups based on the time they were admitted to the intensive care unit(ICU) either during sepsis or not; second it seeks into the different contributions of different variables to those estimates.

The study used 43 variable Sepsis dataset from the PhysioNet/Computing in Cardiology Challenge 2019 and then used SMOTE NC to deal with the dataset s imbalance (93 percent non-sepsis records) and absence markers to manage missing data. Sepsis was predicted using machine learning models like Random Forest, AdaBoost, XGBoost, K nearest neighbour, ExtraTrees, and Logistic Regression; variable significance was evaluated by SHapley Additive exPlanation (SHAP) analysis. [17]

The primary finding of paper include 7.88 percent sepsis rate in the very first hour of intensive care, a greater incidence within men (61 percent) and increased risk in elderly individuals especially those in their 80s and 70s. The intensive care unit (ICU) stay for patients in sepsis was 2.3 days on average. The most optimal model was logistic regression which demonstrated a 66 percent sensitivity and 70 percent AU ROC at hour two and a slight decline by hour six. Heart rate, systolic blood pressure, gender, WBC, and calcium were found by SHAP analysis as being significant to Hour 2 predictions with age being more important by Hour 6. [17]

CHAPTER 5

Description of Dataset

5.1 Source of Dataset

This Dataset have been taken from PhysioNet website. It is a open access challenge for Sepsis Prediction by the name of "Early Prediction of Sepsis from Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019", published on 5th August 2019. The link for the above mentioned Data can be found here:
https://physionet.org/content/challenge-2019/1.0.0/

5.2 Dimensions of Dataset

Under the sub-section of Accessing the Data, The file being used is "training set A" which has the observations of 20,336 patients. "training set A" is a .psv file, a type of text file that uses a Pipe character (|) for a delimiter to separate the values. Thus the term "PSV" stands for "Pipe-Separated Values". This kind of format is similar to the CSV (Comma-Separated Values) but here it uses a different delimiter i.e., Pipe.

For this project "training set A" was converted to a .csv file for which the code is attached in the Appendix.

The dataset contains 790,215 rows and 41 columns. The dataset has a diverse range of unique values across different columns. For instance HR (Heart Rate) column has 333 unique values, indicating a wide range of measurements.

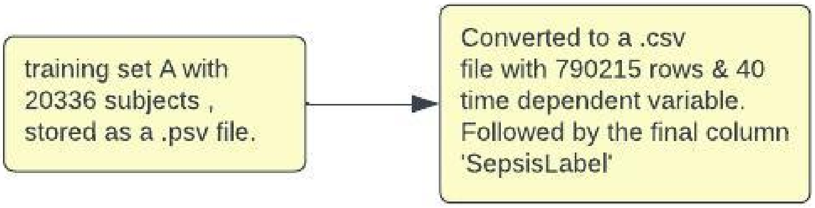

training set A with 20336 subjects , stored as a .psv file.
Converted to a .csv file with 790215 rows & 40 time dependent variable. Followed by the final column 'SepsisLabel'

Figure 5.1: Flowchart explaining the Dataset

Overview of the Challenge: The 2019 PhysioNet/Computing in Cardiology Challenge aimed to utilise physiological information of patients in Intensive Care Units (ICU) to predict sepsis 6 hours before the clinical diagnosis.
Relevance of sepsis: Around 24 billion USD (or rather, 13 percent of U.S. healthcare costs) each year sepsis becomes one of the most costly medical disease in U.S. hospitals. The vast majority of these expenses are borne by individuals who were not diagnosed having sepsis at the moment of admission. Around the globe sepsis expenses are far higher with countries that are developing especially at risk. Everything at considered sepsis represents a serious public health concern which contributes greatly to morbidity, dying and healthcare expenses. [18]
Objective of the Challenge: The challenge is to use the Sepsis criteria to the preliminary detection for sepsis. This was required over participants to generate algorithms that, utilising a utility function for evaluate predictions, predicted sepsis based on changes in the SOFA rating and signals of illness.

CHAPTER 6

Methodology

6.1 Handling Missing Values

The Dataset obtained after getting a csv file as "data csv A" having a total of 790215 rows and 41 columns. In the Dataset issue of Duplicate columns was not there because hardly there were any but many columns were having missing values. There was no such pattern being followed that missingness of one variable is depending on other thus it is a Non-Monotone pattern. [20]

Dataset show significant variation in level of missing data across different column within the dataset which could be seen in the percentage of missing values as shown in image below. With more than 98 percent of data missing columns including 'EtCO2', 'TroponinI', 'Bilirubin direct', 'Fibrinogen' and 'Bilirubin total' show high missing values. This implies that these records were not taken for majority of data. Conversly percentage of missing values in columns like 'HR', 'Resp', 'MAP', 'O2Sat' and 'SBP' is much lower in the range from 7.74 percent for 'HR' to 15.21 percent for 'SBP'.

The approach chosen to deal with the values that are missing is mostly determined upon the fact that how important is each column for the prediction of sepsis. As there are some columns with exceptionally high percentages of missing values especially those near or at 100 percent may not provide useful information. For example 'EtCO2' with 100 percent missing value is being removed.

But percentage of missing values is not the only factor that is being considered while deciding which columns to drop from Dataset. Here we will see how each variable relates to the predictor column "SepsisLabel" which will be taken into account by using heatmap for correlation matrix.
For the columns which were there in the Vital columns list and other columns that did not have more than 25 percent of missing values, Imputation was used in them. Method like front fill and back fill was used for filling the empty spaces in certain columns which are relevant for analysis.

At last this is the list of the columns being dropped: 'TroponinI', 'ID person', 'Bilirubin direct', 'Unit', 'Lactate', 'SaO2', 'AST', 'FiO2', 'Bilirubin total', 'Unnamed: 0', 'SBP', 'DBP', 'EtCO2', 'HCO3', 'pH', 'PaCO2', 'Alkalinephos', 'Calcium', 'Magnesium', 'Phosphate', 'Potassium', 'PTT', 'Fibrinogen', 'Unit1', 'Unit2', and 'category of sepsis'.

Given below is a plot which shows missing data percentage per column with Columns name on the x-axis and missing data percentage of that respective column on the y-axis. The bars displaying the missing data per column gradually get fewer towards the right after starting with the highest percentages on the left.

Green shows a smaller proportion of missing data whereas pink indicates higher percentages the colours of the bars change from solid pink to green. 'EtCO2' on the left has all of its data missing and thus is represented by the deep pink colour. 'Bilirubin direct' and 'Fibrinogen', the following two columns, seem to have a little less missing data, but they are still in the category of column with almost 100 percent missing values.

The degree of pink reduces as we shift to the right indicating a decrease in the proportion of data that is missing. The less intense pink colour of columns like 'SaO2', 'Phosphate' and 'Magnesium' shows a lesser but statistically significant percentage of missing data mostly between 80 and 90 percent.
The bars to the right changes from light pink to green demonstrating the decline in the

percentages for missing data. Particularly the green colour of columns like ’Temp’, ’SBP’ ,’DBP’, ’Resp’, ’MAP’ and ’HR’ shows significantly smaller percentages of missing data possibly between 0 and 20 percent. Because of the fact it is the only column with a fully green bar ’HospAdmTime’ has the lowest percentage of missing data.

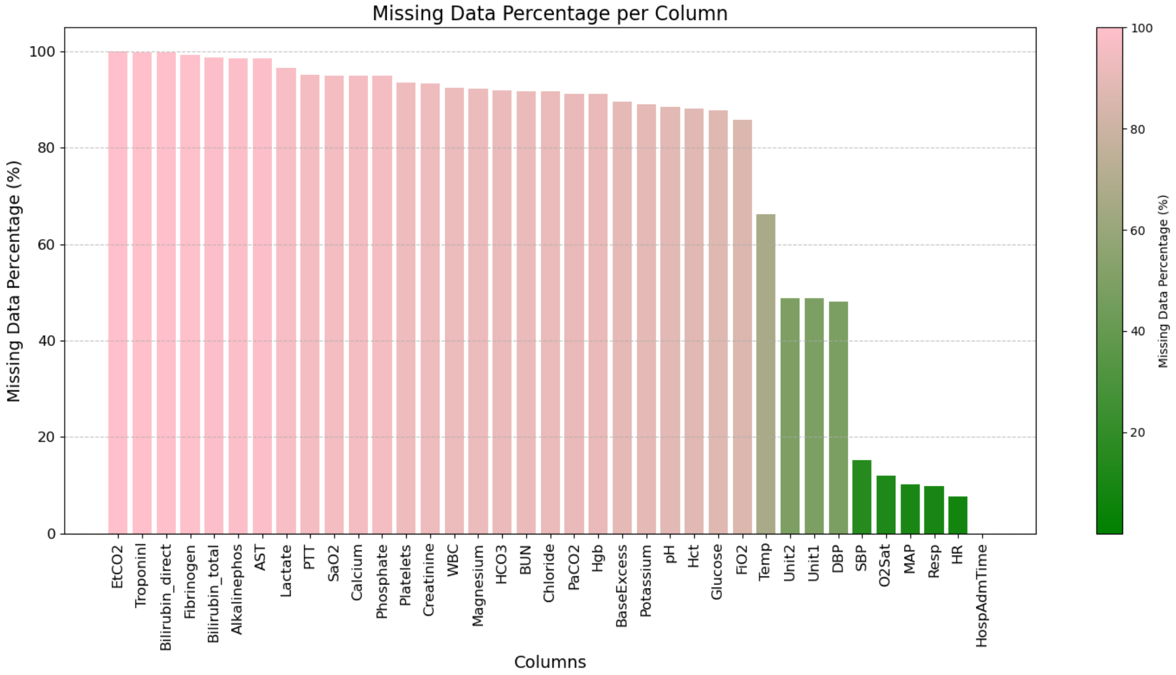


Figure 6.1: Chart Showing percentage of missing data in each column

6.2 Descriptive Statistical Analysis and Visualization

Descriptive statistical analysis is necessary for understanding the summary, data distribution and the visual representation. Exploratory data analysis (EDA) gives essential phase in the study process. With assistance of EDA we can identify intricate and hidden patterns in dataset.

6.2.1 Pie chart for Gender Distribution

The pie chart shown below by the name of 'Distribution of Gender' depicts the gender breakdown from training set A, showing a little disparity that males make up 57.8 percent and females 42.2 percent. This indicates a higher prevalence of males considering overall patients.

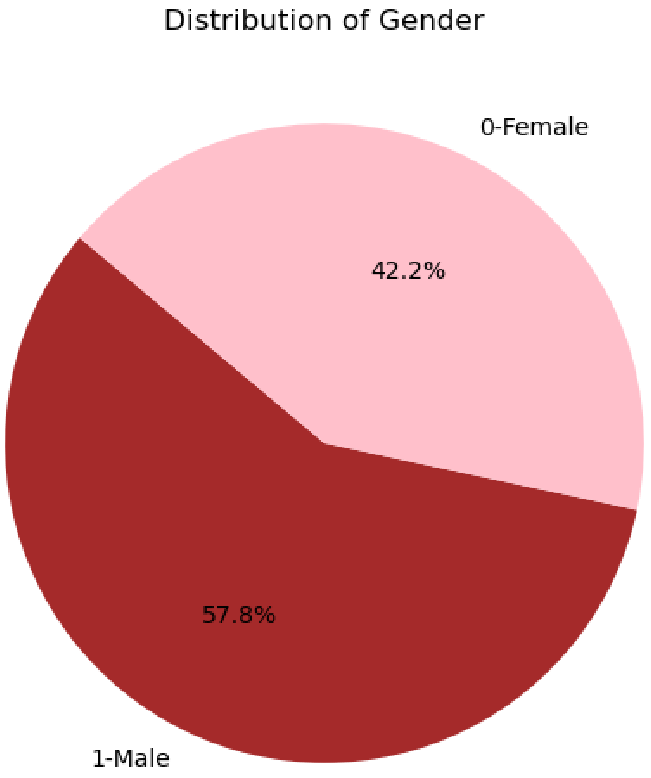


Figure 6.2: Distribution of Gender

6.2.2 Box Plots for Vital Signs

(a) Respiration Rate (Resp): The compact box for respiration rate indicates some small interquartile range (IQR) implying that respiration rates of the top 50 percent of patients are now identical. Additionally, there seems to have a trend towards lower respiration rates since the median is slightly beneath 20 breaths per minute and is near the third quartile. Although most patients had similar rates a few have substantially lower or higher rates as suggested by their whiskers that improve to greater variation than IQR. However there are plenty of anomalies beneath the lower whisker which indicate instances of very low breathing. These might be a result of various conditions.

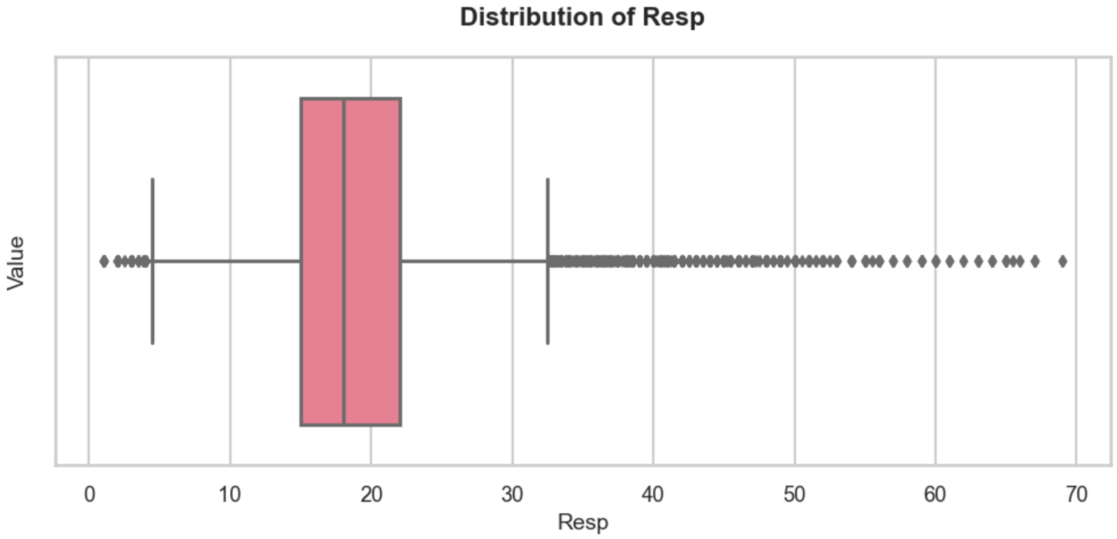


Figure 6.3: Distribution of Respiration Rate

(b) Systolic Blood Pressure (SBP): Situated in the centre of the box median systolic blood pressure measures approximately 120 mm Hg. The IQR is higher than the value of respiration rate plot. This implies that middle 50 percent of patients had a broader distribution of systolic blood pressure levels. The whiskers indicate a even wider range of values both noticeably lower and higher. There are a lot many outliers in the top whisker which can reach up to 200 mm Hg or higher. They indicate that certain individuals have very high systolic blood pressure that could indicate hypertensive episodes.

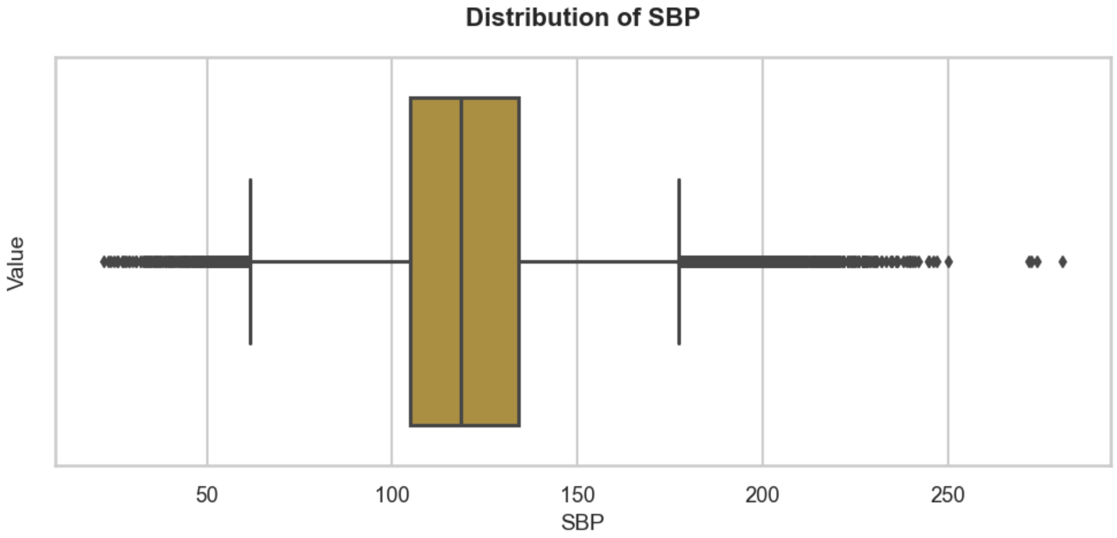


Figure 6.4: Distribution of Systolic Blood Pressure

(c) Pulse Oximetry (O2Sat): The majority of individuals had sufficient oxygenation shown by high median pulse oximetry with almost 98 percent. Considering small box it may be inferred that interquartile values with a larger saturation range skew and were tightly packed around the median. The upper whisker remains near to the box indicating no shift in higher oxygen saturation levels while the bottom whisker stretches to almost 90 percent. A few percent of people with oxygen saturation levels under the lower whiskers average are probably those who have breathing issues.

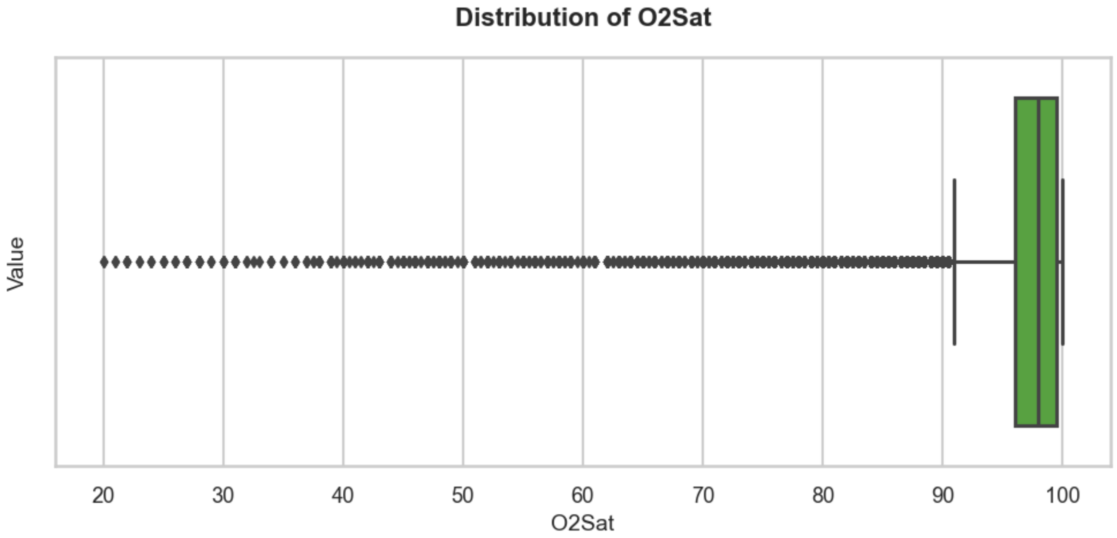


Figure 6.5: Distribution of Oxygen

(d) Age: The age distribution is symmetric having an equally aged IQR ranging between 50 to 75 years old and the median age in mid 60s. Patient group looks to have an almost equal age distribution without any noteworthy age extremes as shown by the symmetry and no outliers.

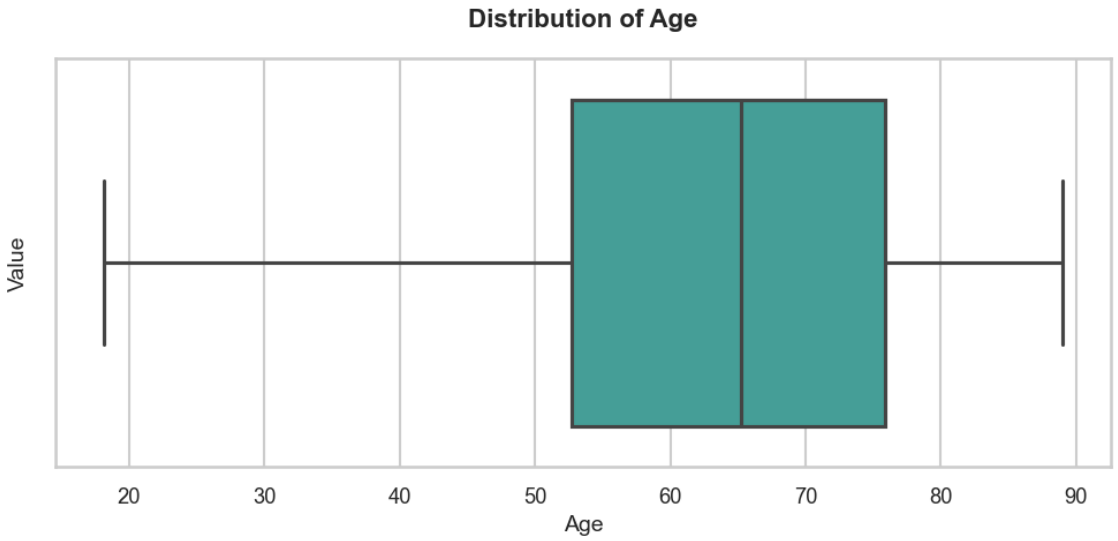


Figure 6.6: Distribution of Age

(e) Mean Arterial Pressure (MAP): A IQR which shows fluctuation yet cannot be as wide as SBP is shown in the box plot for MAP having the median value near 80 mm Hg. There are some outliers on each side showing just a small number of patients having MAP metrics exceed the usual range that may be reason for medical attention.

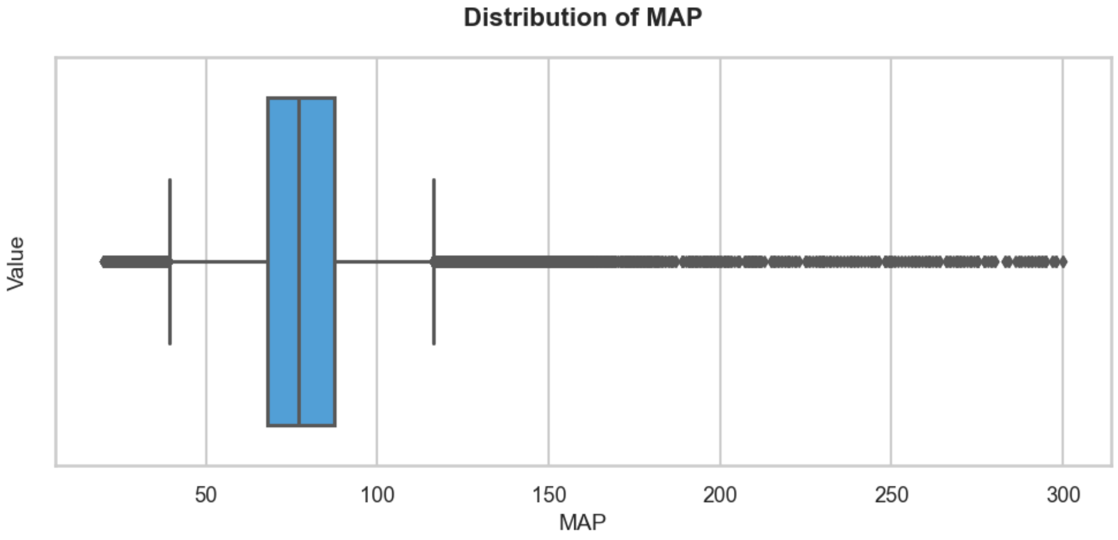


Figure 6.7: Distribution of Mean Arterial Pressure

(f) Temperature (Temp): In ideal adult the body temperature around 37Â°C is thought as median. The tight IQR shows that a majority of patients body temperatures fall in agreement with median. On the contrary the whiskers and outliers at each side of the range suggest irregular slips in hypothermic states.

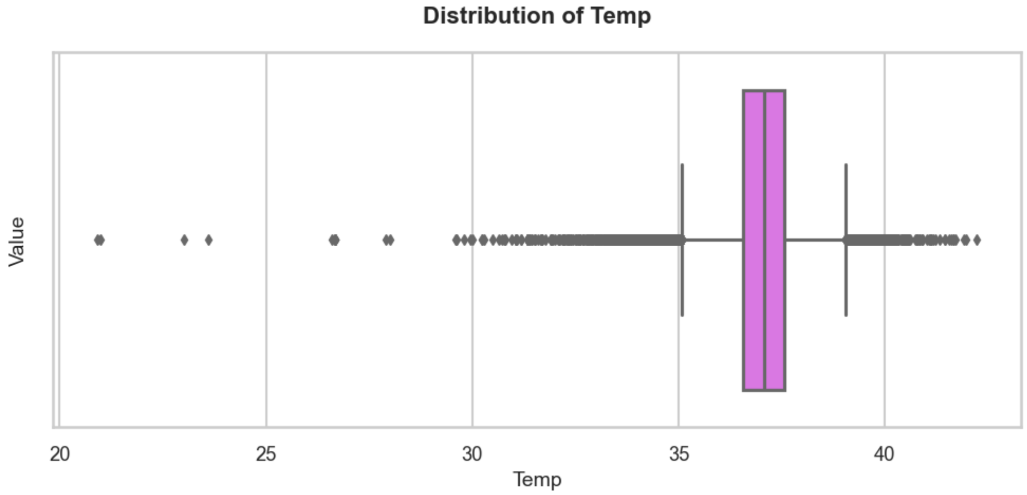


Figure 6.8: Distribution of Temperature

(g) Heart Rate (HR): IQR shows a range of cardiac rates across patients with median heart rate at slightly higher than 80 beats per minute. Only a few of patients experienced raised heart rates as reflected by several high outliers beyond the top quartile. If such elevated heart rates were associated with fundamental health concerns these may have clinical significance.

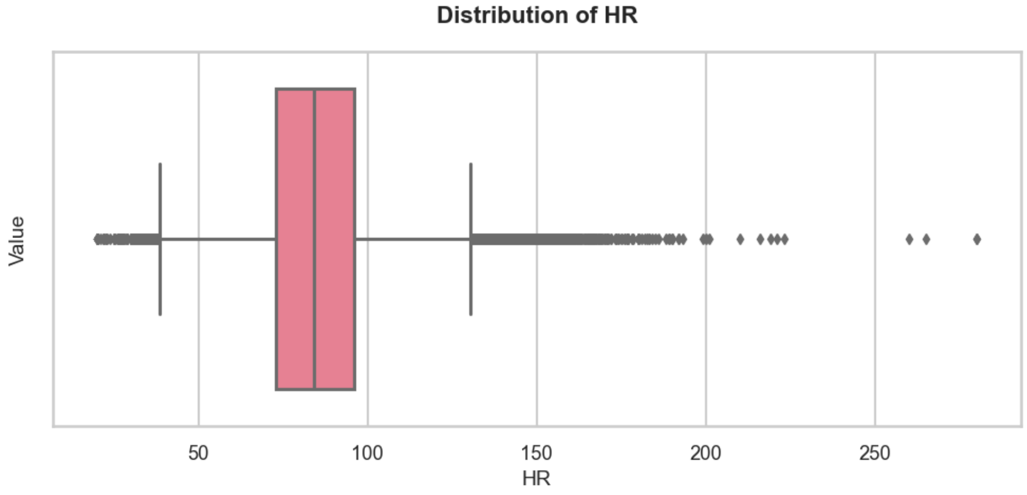


Figure 6.9: Distribution of Heart Rate

(h) Hospital Admission Time (HospAdmTime): This figure displays the median that is near 0 implying as most patients get taken to ICU soon after arriving at the hospital. The right-skewed distribution with outliers on the opposite hand suggest a small number of patients that may have faced admission delays some outliers report delay surpassing 20 hours.

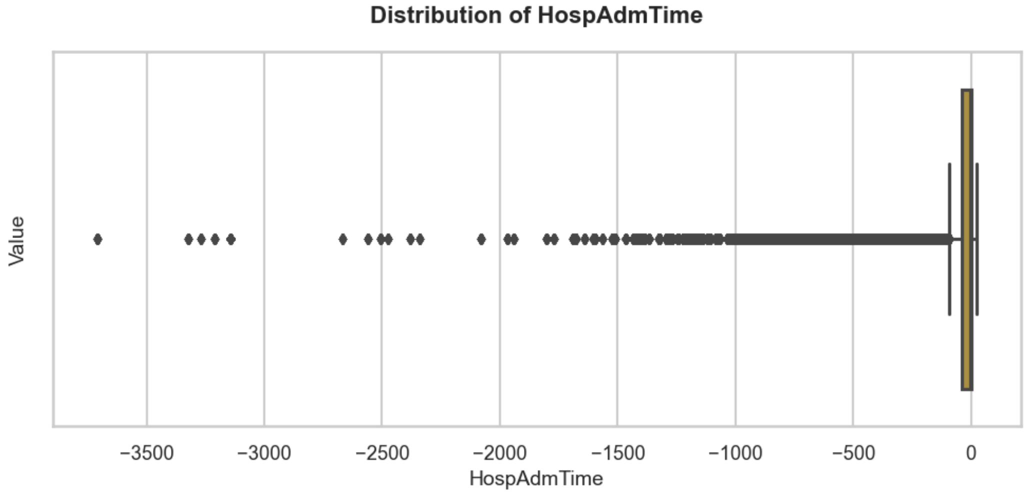


Figure 6.10: Distribution of Hospital Admission Time

(i) Diastolic Blood Pressure (DBP): At about 60 mm Hg average diastolic blood pressure lies in the normal range. As there are no outliers the box is small and the IQR indicates minimal movement amongst middle 50 percent of patients indicating a typically constant diastolic BP within the group.

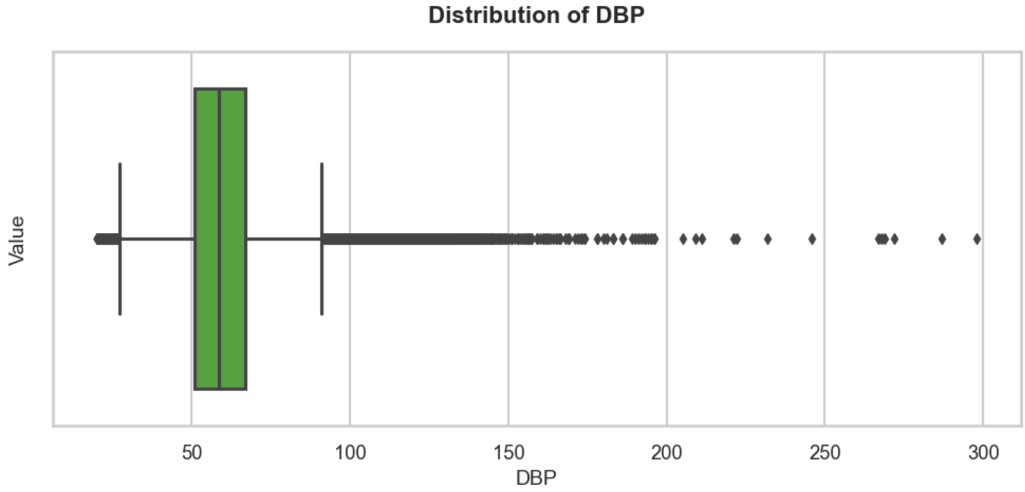


Figure 6.11: Distribution of Diastolic Blood Pressure

6.2.3 Cat plot for Sepsis Label across Age and Gender

Overall mid-60s represents median age for all groups demonstrating that presence of sepsis does not significantly alter median age. When compared to the non-sepsis groups the age distributions with interquartile ranges in the sepsis groups were more closely spread. This suggests that sepsis is more prevalent in a particular age group. While there is a concentration of instances in 60s to 70s age range sepsis may impact individuals outside of this age bracket as shown by presence of outliers across all categories. Although sepsis is particularly prevalent in a significant age bracket it is not limited to a particular age range with outliers seen in both genders and conditions.

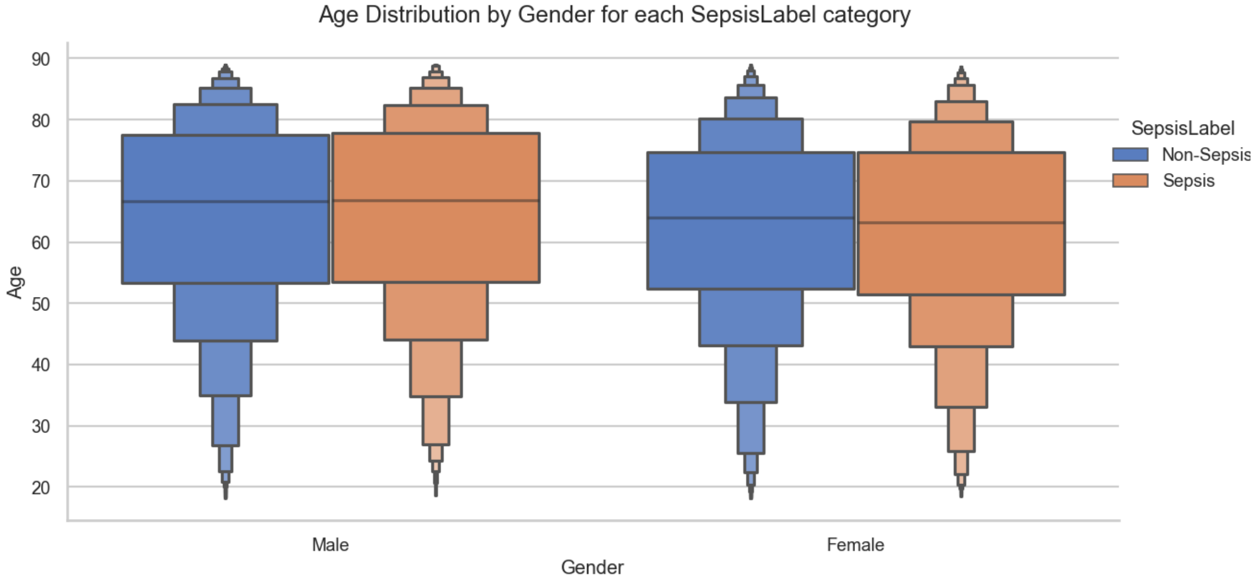
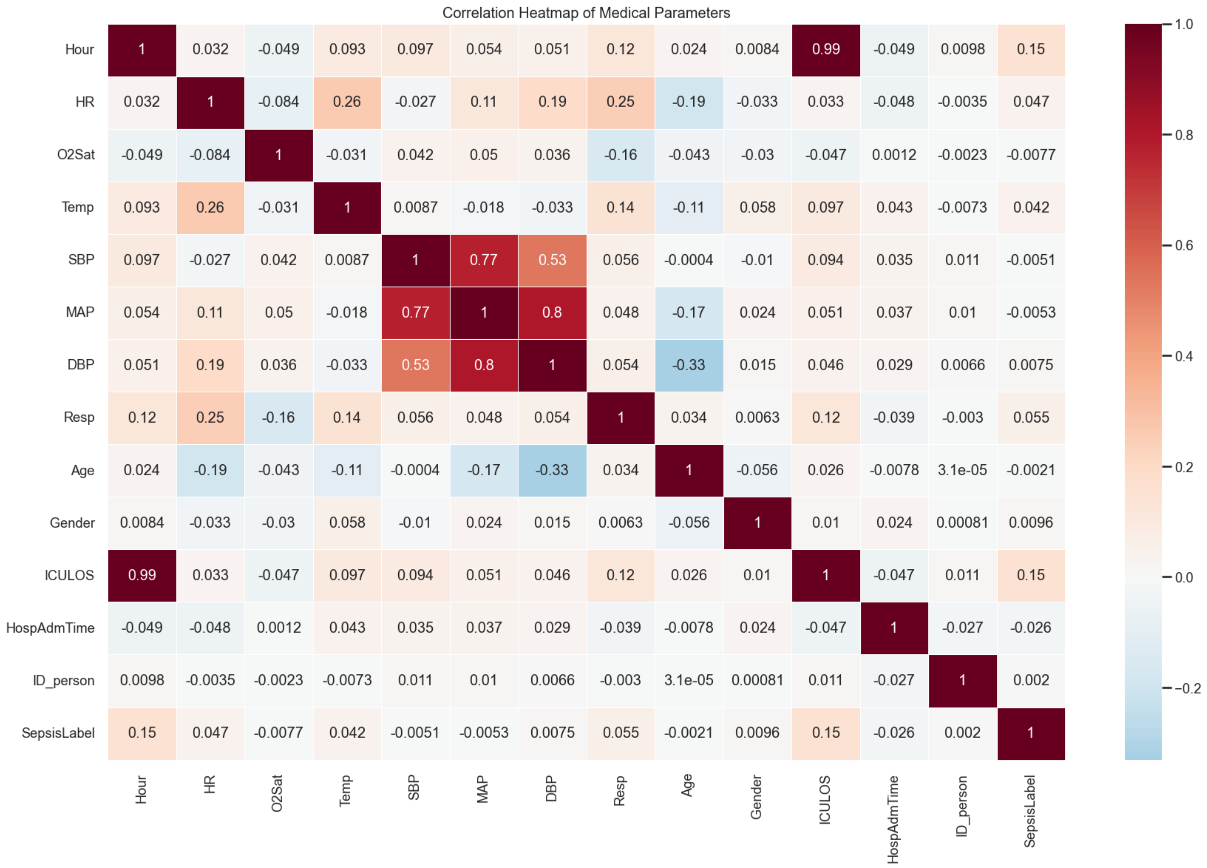


Figure 6.12: Age Distribution by Gender for each Sepsis category

6.2.4 Correlation heatmap for medical parameters



	Hour	HR	O2Sat	Temp	SBP	MAP	DBP	Resp	Age	Gender	ICULOS	HospAdmTime	ID_person	SepsisLabel
Hour	1	0.032	-0.049	0.093	0.097	0.054	0.051	0.12	0.024	0.0084	0.99	-0.049	0.0098	0.15
HR	0.032	1	-0.084	0.26	-0.027	0.11	0.19	0.25	-0.19	-0.033	0.033	-0.048	-0.0035	0.047
O2Sat	-0.049	-0.084	1	-0.031	0.042	0.05	0.036	-0.16	-0.043	-0.03	-0.047	0.0012	-0.0023	-0.0077
Temp	0.093	0.26	-0.031	1	0.0087	-0.018	-0.033	0.14	-0.11	0.058	0.097	0.043	-0.0073	0.042
SBP	0.097	-0.027	0.042	0.0087	1	0.77	0.53	0.056	-0.0004	-0.01	0.094	0.035	0.011	-0.0051
MAP	0.054	0.11	0.05	-0.018	0.77	1	0.8	0.048	-0.17	0.024	0.051	0.037	0.01	-0.0053
DBP	0.051	0.19	0.036	-0.033	0.53	0.8	1	0.054	-0.33	0.015	0.046	0.029	0.0066	0.0075
Resp	0.12	0.25	-0.16	0.14	0.056	0.048	0.054	1	0.034	0.0063	0.12	-0.039	-0.003	0.055
Age	0.024	-0.19	-0.043	-0.11	-0.0004	-0.17	-0.33	0.034	1	-0.056	0.026	-0.0078	3.1e-05	-0.0021
Gender	0.0084	-0.033	-0.03	0.058	-0.01	0.024	0.015	0.0063	-0.056	1	0.01	0.024	0.00081	0.0096
ICULOS	0.99	0.033	-0.047	0.097	0.094	0.051	0.046	0.12	0.026	0.01	1	-0.047	0.011	0.15
HospAdmTime	-0.049	-0.048	0.0012	0.043	0.035	0.037	0.029	-0.039	-0.0078	0.024	-0.047	1	-0.027	-0.026
ID_person	0.0098	-0.0035	-0.0023	-0.0073	0.011	0.01	0.0066	-0.003	3.1e-05	0.00081	0.011	-0.027	1	0.002
SepsisLabel	0.15	0.047	-0.0077	0.042	-0.0051	-0.0053	0.0075	0.055	-0.0021	0.0096	0.15	-0.026	0.002	1

Figure 6.13: Correlation Heatmap

The correlation between SepsisLabel and ICULOS is approximately 0.15. Out of all the aspects this correlation is the most significant and indicates a somewhat positive relationship. This may indicate that sepsis is more likely to be present in patients who stay longer in the ICU or that patients with sepsis are more likely to stay longer in the ICU.

Hour and SepsisLabel have a 0.15 correlation, which is same as to ICULOS's correlation. This could suggest the overall influence of hospital related factors on the patient's condition by suggesting a relationship between the risk of sepsis or the identification of sepsis and the amount of time that has passed. The heart rate and sepsis are slightly related, with a correlation of about 0.047. This means that when heart rate goes up there is a small increase in the chance of sepsis, but this connection is not very strong. This makes sense because sepsis can affect heart rate.

The oxygen level in the blood and sepsis are weakly linked in the opposite direction, with a connection of -0.077. This suggests that when sepsis occurs, oxygen levels might drop a little, but the relationship is not very strong. This is reasonable because sepsis can make it harder for the body to use oxygen well.

The temperature shows almost no correlation with SepsisLabel close to zero indicating no direct link. Similarly Systolic Blood Pressure (SBP), Mean Arterial Pressure (MAP), and Diastolic Blood Pressure (DBP) also have very weak positive correlations with SepsisLabel. The correlations are 0.094 for SBP, 0.055 for MAP, and 0.029 for DBP. These small positive correlations states that there is a very slight connection with sepsis possibly because sepsis can affects blood pressure regulation.

The respiratory rate has a weak positive correlation of 0.055 with SepsisLabel. This means that a higher respiratory rate might be slightly related to sepsis which is understandable since sepsis can causes changes in how we breathe. Age has a weak negative correlation with SepsisLabel at -0.056. This means that younger patients in this data might have a slightly higher chance of sepsis but the connection is not strong.

Gender has almost no correlation with SepsisLabel at -0.056, showing that there is no significant link between gender and sepsis in this data.

Lastly the time spent in the hospital before admission has a very weak negative correlation with SepsisLabel at -0.047 indicating there is almost no connection between how long someone is in the hospital before being admitted and getting sepsis.

In conclusion, most of the parameters show very weak correlations with SepsisLabel meaning there is no strong direct relationship. However this does not mean that these variables can not be useful in predicting sepsis when they are combined with other data and looked at with more advanced methods that can uncover some non linear connections and interactions. It is also important to keep in mind that even if there is no direct link it does not mean these variables is not clinically significant in cases of sepsis.

6.2.5 Density Plot for ICU Length of Stay

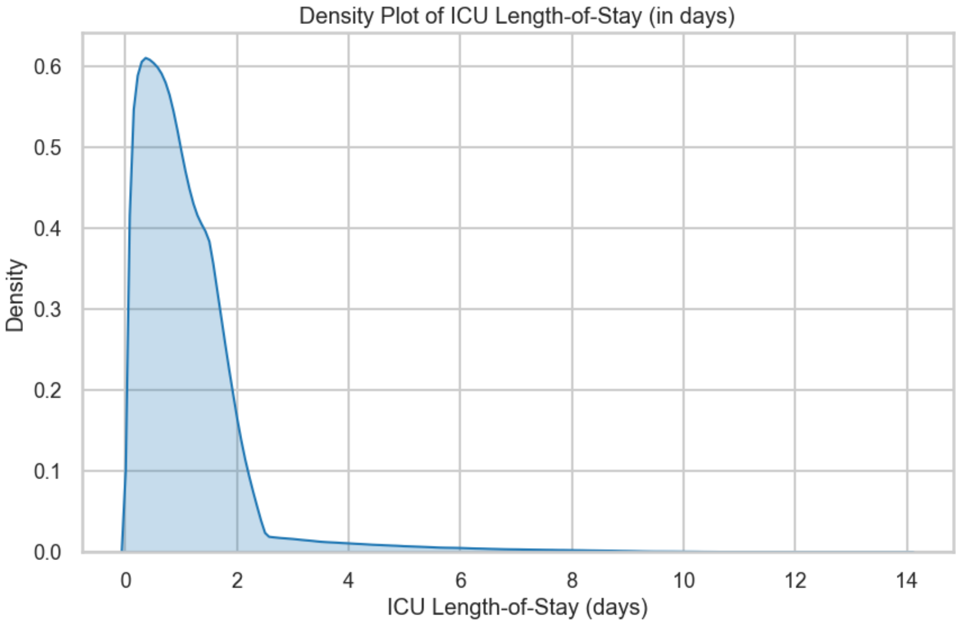


Figure 6.14: Density plot of ICU LOS

Similar to a histogram, the plot given is a Kernel Density Estimate (KDE), which is used to show the distribution of observations in a dataset. Here it displays the distribution of days spent in ICU. Here is detailed analysis: Maximum Point in the Distribution: Since the peak is close to zero, it is likely that patients spend less than a day in ICU on average. Because it is the dataset's mode, the precise density peak value is not given. Skewness: The distribution is right-skewed, meaning that as hospital stays get longer, there are fewer cases of longer stays in the ICU. The tail reaches up to approximately 14 days and extends to the right. Spread: The KDE curve's breadth provides information about the data spreadness. The curve in this instance is wide, indicating large degree of variability in the duration of ICU stays. But after around two days there is a sharp decline suggesting a sharp decline in the frequency of longer stays. Kurtosis: A leptokurtic distribution is suggested by the distribution sudden peak and severe drop-off. This indicates that there is a strong concentration of stay lengths around the mean. Potential Outliers: A long tail is shown to the right. As individual data points is not displayed by the KDE, this may indicate the existence of outliers that correspond to unusually lengthy ICU stays. Numerical Data: It is not possible to get accurate numerical values from the plot alone without the raw numerical data or any extra markings on the plot (such as quartiles or mean/median lines). But because of the plot's right skewness, may assume that the mean will probably be higher than the median.

Area Under the Curve: The probability distribution is represented by the entire area under the curve, which should equal 1. The probability that ICU stay chosen at random will be shorter than that duration indicated by the area under the curve to the left of a certain point on the x-axis.

6.2.6 Trends Over Time

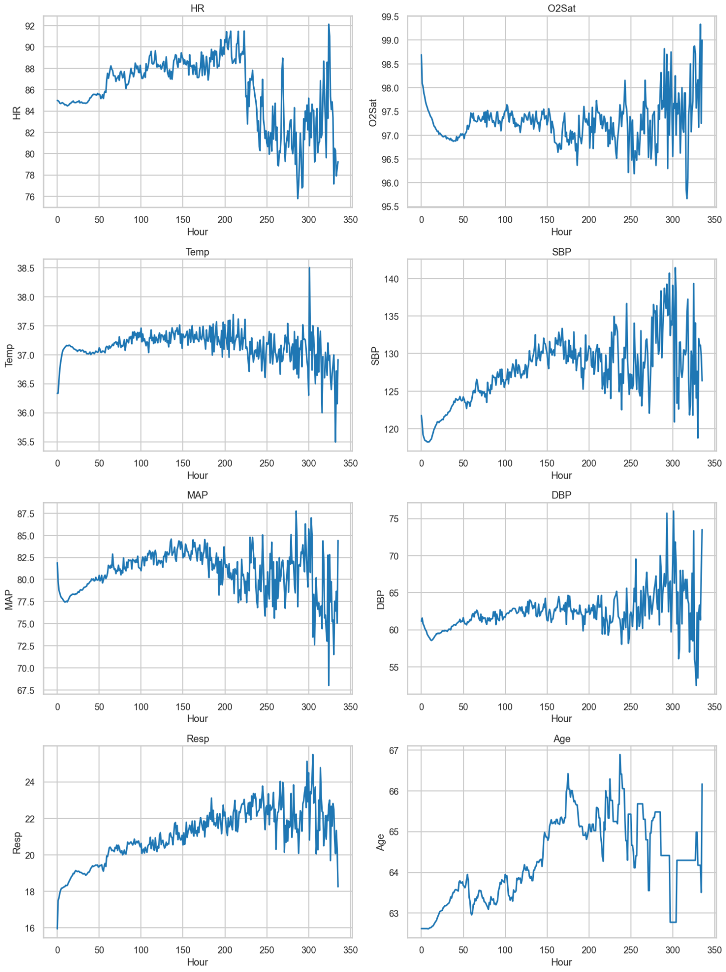


Figure 6.15: Trends in Health Parameters Over Time

The data for different health measures change in various ways over time. The Heart Rate (HR) goes up and down between about 76 bpm and 92 bpm, and gets more variable as time goes on like a big jump near hour 100 and a big drop near hour 250. Oxygen Saturation (O2Sat) levels begin near 98.5% and slowly go down, with big drops sometimes such as one around hour 200 to around 96.5%. The patient's temperature is usually in the normal range of 36.5 celsius to 37.5 celsius but sometimes it goes over 38 celsius which might mean they are having fever periods.

Systolic Blood Pressure (SBP) readings generally go up from about 120 mm Hg to almost 140 mm Hg and they change a lot which might mean there are times of high blood pressure. Mean Arterial Pressure (MAP) starts at about 75 mm Hg and goes up changing quite a bit especially later on. Diastolic Blood Pressure (DBP) starts steady near 60 mm Hg and also goes up but does not change as much as SBP getting to about 75 mm Hg at the end.

The respiration rate slowly increases from around 16 breaths per minute to more than 24 which could be because of stress breathing problems or more physical activity. Finally the age data goes up and down between 63 and 67 years which might mean it is from different patients or there is a mistake in how it was recorded since age usually does not change much in such short times.

6.2.7 Bar Chart of Sepsis Patient

The ′Patient Counts in Different Categories′ bar chart depicts whether patients are divided throughout three different categories which are : "Sepsis After Admission " "Sepsis HR Zero " and "Data of Non - Sepsis". The majority of patients 91.20 percent of the total are in the ′Data of Non-Sepsis′ category highlighting the high proportion of non-sepsis cases in the dataset. In contrast ′Sepsis After Admission′ comprises 7.80 percent of the cases suggesting a smaller but still important subset of individuals who get sepsis following hospital admission. Having only 1.00 percent of the instances fitting into ′Sepsis HR Zero′ category it is clear the sepsis patients who have Sepsis when they are being admitted to the hospital is the least.

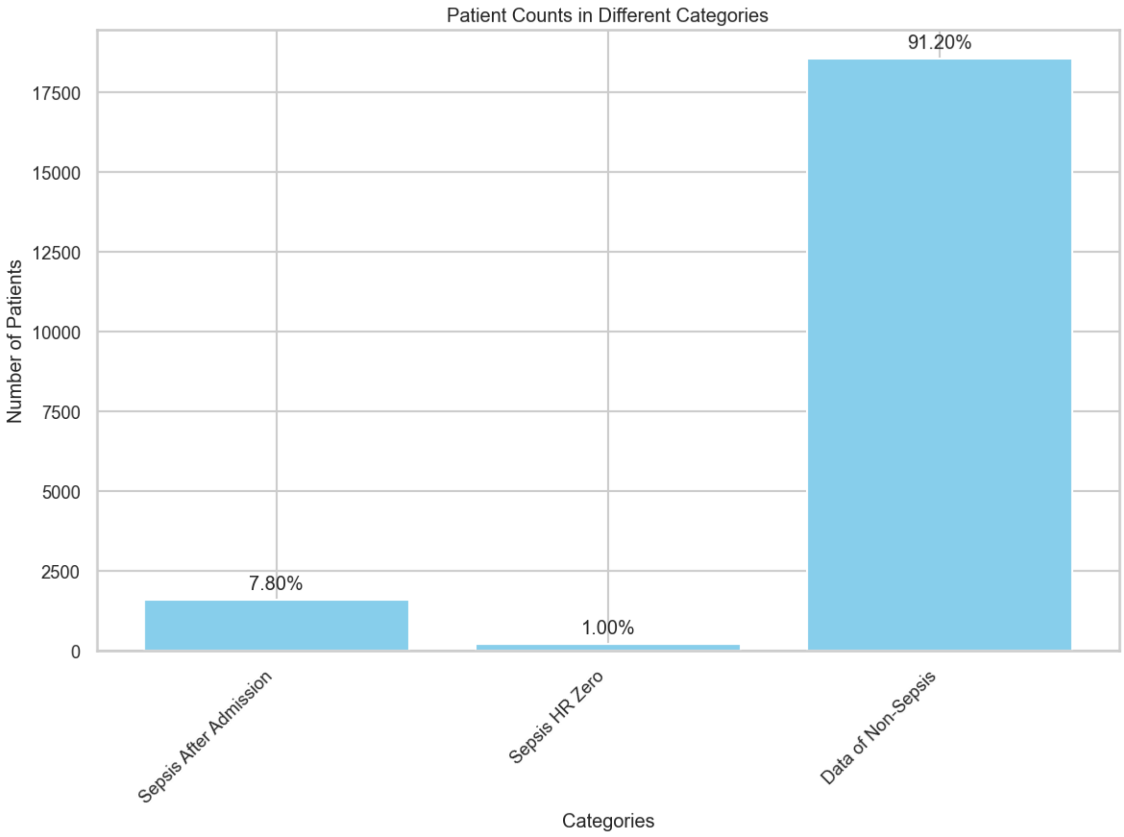


Figure 6.16: Distribution of Patients across different category

6.2.8 Donut Chart of Sepsis in different Hours

This donut chart presents the prevalence of unique cases of sepsis as time passes in 3 separate onset time intervals: from hour zero (Hour 0), shortly thereafter (Hour 1-6) and after 7 hours (Hour 7+). A significant number of the total cases 75.6 percent were discovered around Hour 7+, shown in the significant area of blue. This indicates the majority of cases of sepsis emerge long after the initial period of hospitalisation or monitoring of patients. 13.0 percent of the cases fall within green part which corresponds to Hours 1-6. It suggests that less sepsis diagnosis happen within those critical hours. The most modest percentage marked in red reflects Hour 0 when 11.3 percent of situations occurred. It shows how not every cases of sepsis are identified within the initial hour of a patients assessment.

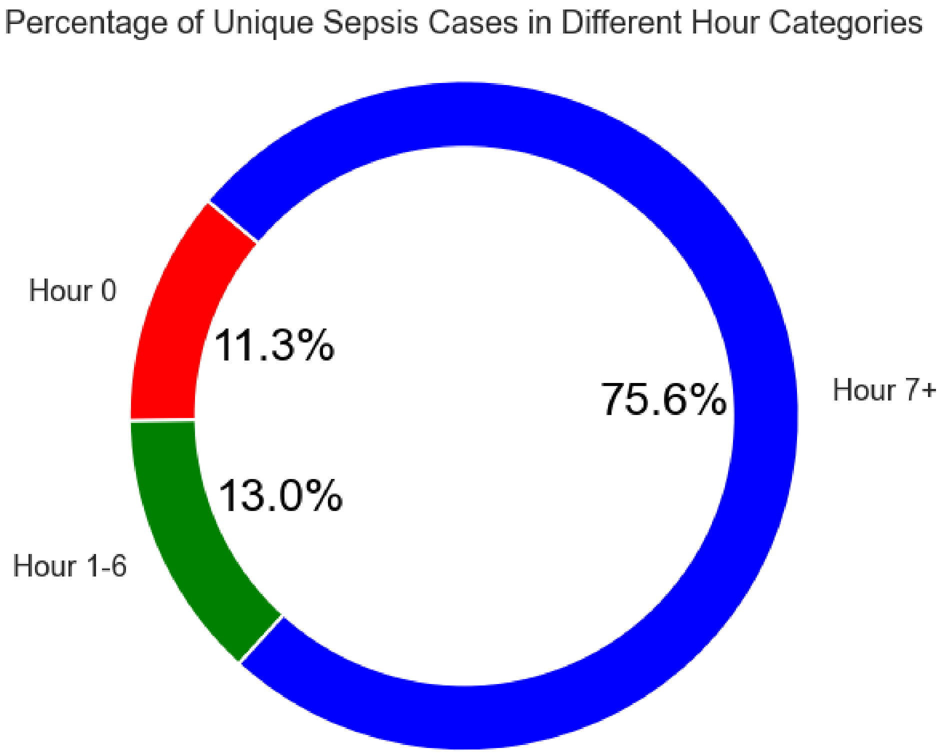


Figure 6.17: Distribution of Sepsis cases in Different Hours

6.3 Methodology Approach

6.3.1 Data Pre-processing

The process of using different statistical or machine learning techniques to put in the missing values within dataset is termed as **Data Imputation**. When working with dataset that has missing data but the parameter are relevant for classification then there is a need to use all available data to make accurate predictions or analyses. [21]

While using the Pandas Dataframe the values which were empty were imputed for each patient as they were grouped by ID person. By using both Front fill and Back fill simultaneously missing values are being filled by the nearest possible non missing value thus values missing are being filled by the details of the same patient.

One-Hot Encoding represent a machine learning data pre - processing method that converts strings of text corresponding to categorical variables as numerical format that machine learning models can interpret. It must be used as text - based categorical variables cannot be dealt with directly by many machine learning algorithm that operate only using numerical data. In one - hot encoding a categorical variables unique categories are expressed by a binary columns (0 and 1). It guarantees that each category lack any innate ordinal relationships which might lead algorithm to read any relationship wrongly. [22]

In this project One - Hot encoding is applied to Gender Column so that the machine learning algorithm thus creating two new columns with binary values i.e. one for Male and one for Female.

6.3.2 Development and Training of Model

(a) **Train-Test Split of Dataset:**

The Pre - processed Dataset is being split for training model where train size is 80 percent and test size is taken 20 percent to evaluate performance of the model to classify Sepsis Label. Because Dataset is in Time series it is splitted based upon ID person to maintain regularity within both training set and testing set.

(b) **Handling Imbalanced Dataset:**

Among the clinical data majority of patient never get positive for Sepsis thus it makes a highly Imbalanced data. For addressing this problem two approaches were adopted which are Undersampling and SMOTE technique for oversampling which will be discussed later in report.

(c) **Selection of the Model:**

As after deploying all 3 Machine learning Algorithms which is Gaussian Naive Bayes a probabilistic classification problem, Random Forest a ensemble learning algorithm for regression and classification and Decision Tree Supervised ML algorithm for classification and regression in Imbalanced Data.

(d) **Evaluation of Model:**

As this is a Classification prediction and seeing the mortality rate for Sepsis it becomes essential to analyze best performing model. For the above mentioned Machine learning algorithm they will be evaluated for their performance with Imbalanced data, Undersampled data and SMOTE method for Oversampled Data. Furthermore we will consider parameter such as Area under the ROC Curve (AU ROC) considering imbalanced Data.

6.3.3 Correlation Analysis

The statistical approach applied to figure out the direction and strength for a linear relationship among variables within the dataset is referred as **Correlation Analysis**. Through a correlation coefficient ranging from -1 to 1 it determines exactly how much two or more variables proceed together or in contrast to one another. When two variables are prone to rise together this is shown as a positive correlation and when one is on the rise as another decreases this is indicated by a negative correlation. [23]

In this Dataframe there were many Positive and Negative correlations were discovered like 'ICULOS' and 'ICULOS days' and thus those columns were removed.

6.3.4 Handling Imbalanced Dataset

Class imbalances may create serious obstacles in healthcare sector in which distribution of the desired variable **'SepsisLabel'** is essential. The breakdown of instance in this dataset varies as follows: **2.17 percent** of cases are categorised as **severe sepsis** while **non-septic** cases make up the majority of **97.83 percent**. This large disparity in the breakdown of class highlight the importance for employing particular modelling method. The dataset containing details about a substantial number of 20,336 patients provides a representative pool suitable for modelling as well as indepth analysis. The class composition of this dataset points out importance of dealing with class imbalance while creating predictive model in sepsis related results.

(a) **Undersampling:**

In order to get around issue of disparate class distributions in classification task undersampled data is essential. Dataset utilised for medical research among other real world application often reveal imbalanced class distribution with one class significantly outnumbering another. Classifier which are biassed towards the dominant class because of imbalance might do badly when it relates to prediction of minority class. Choosing representative data point from the majority and minority class whereas reducing the effect of the disparate class distribution is referred to as **Undersampling**. [24]

In this project the process starts with the split dataset into training and testing set where feature data and target variable which is SepsisLabel are separated. The key objective is to address class imbalance. To achieve this balance undersampling method calculate the size of the minority class which is positive for Sepsis. Subsequently the indice corresponding to the majority class which means patients without Sepsis are determined. The random selection of a subset of instances from the majority class matching the size of the minority class. This selection process is carried out without the repetition making sure that each chosen instance is different. By joining these randomly selected majority class instances with the minority class instance the undersampled dataset is made to train model.

(b) **Synthetic Minority Oversampling Technique (SMOTE):**

It is over -sampling method to address the class imbalance problem by creating synthetic instance in minority class instead of simply replicating existing minority samples a process that may lead to overfitting issue to enhance this SMOTE algorithm is used. The main idea behind SMOTE is to make artificial minority class example with line segments connecting minority sample with their 'k' nearest neighbor from minority class. The value of 'k' identifies total of nearest neighbors considered in the process allowing for control over the oversampling rate. Thus SMOTE expands the minority class representation in this dataset. [25]

To deal with the problem of class imbalance in a binary categorization. To start the target variable 'SepsisLabel' is separated from the characteristic in the dataset by dividing it into characteristic and binary label. After this data is divided in training and testing set followed by standard procedure. Following that a instance of SMOTE object is generated which handle the oversampling. By developing synthetic sample of minority class - Sepsis label 1 to match the majority class - Sepsis level 0 method is applied to fix the disparate class distribution. Following the fitting of SMOTE object to the training set it produce synthetic instance by looking at the distribution of sepsis cases in patients. The end product is oversampled training dataset made up of both the freshly developed synthetic sample and the initial training data.

6.3.5 Machine Learning models employed

(a) **Gaussian Naive Bayes:** Machine Learning algorithm for classification task is a probabilistic machine learning algorithm. In a effort to simplify the computation of conditional probabilities simpler it depend on the Bayes theorem and make certain presumption about data distribution. This approach work especially well in continuou feature (also called as attribute) that have a Gaussian (normal) distribution. [26]
Considering the label of the class every characteristic are taken to be conditionally independent which gives Gaussian Naive Bayes its 'naive' name. In other words provided the class label it assumes that existence or absence of a particular attribute is unaffected by of the existence or absence of other characteristics. This assumption increase the ease of calculation and is almost reasonable almost all the time. [26]
Bayes Theorem: Gaussian Naive Bayes is made upon Bayes theorem which is the fundamental concept in probability theory and statistics. Bayes theorem allows to calculate the probability of certain event occurring given the probability of another related event. It is expressed as: [27]

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{6.1}$$

P(A | B) - Probability of A given event B
P(B | A) - Probability of B given event A
P(A) - Probability of event A
P(B) - Total probability of event B [27]

Gaussian Naive Bayes (GNB) Algorithm: It is a probabilistic type classifier and it work in 2 parts:

Training Part:
Class Prior Probability - Gaussian Naive Bayes (GNB) initially computes the probability of a particular class (P(C)) according to training dataset which is the ratio of number of instance for a particular class (C) to total number of the training instance.
Mean (μ) and Variance(σ^2) for Each Feature in Each Class:

For each parameter in the Dataset, Gaussian Naive Bayes (GNB) calculates Variance and Mean for the parameter value for each class. [26]

Gaussian Distribution: The fundamental assumption of GNB is this that the continuous valued features with each class follow the Gaussian (normal) distribution. Gaussian distribution is characterized by 2 parameters: mean (μ) and variance (σ^2).
GNB assume that the feature value within each class are normally distributed.
Probability density function is computed as: [28]

$$[P(X|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(X-\mu)^2}{2\sigma^2}\right)] \tag{6.2}$$

X - Value of the feature.
μ - Mean of the distribution.
σ^2 - Variance of the distribution. [29]

Decision Rule: At last the final decision rule is going to select the class that has the highest posterior probability: [29]

Predicted Class = $\arg\max_C P(C|X)$ for all classes C

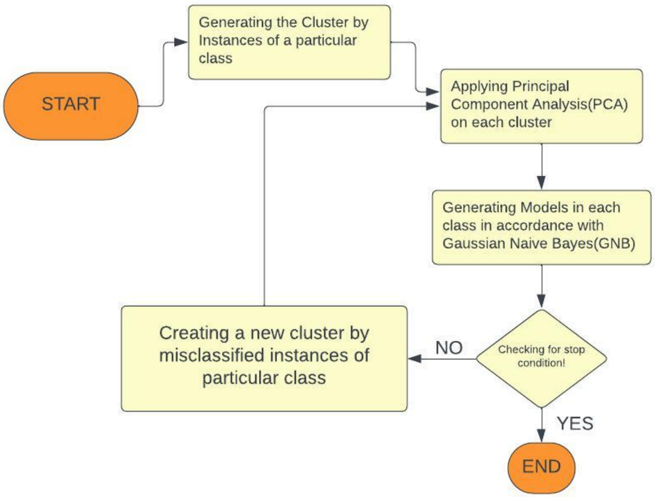

START
Generating the Cluster by Instances of a particular class
Applying Principal Component Analysis(PCA) on each cluster
Generating Models in each class in accordance with Gaussian Naive Bayes(GNB)
Checking for stop condition!
NO
Creating a new cluster by misclassified instances of particular class
YES
END

Figure 6.18: Flow chart for proposed Gaussian Naive Bayes Classifier [28]

(b) **Random Forest:**

Random Forest algorithm is a method of ensemble learning where numerous decision tree are built for classification and regression task throughout the training. So as to make sure diversity between decision tree they function by building a forest for them every one based on randomly selected portion of training. This range help in decreasing overfitting which is a prevalent issue in machine learning model. Each decision tree in Random Forest is a set of condition organised in a tree like hierarchical structure which link feature to target variable. This method create a effective, exact and general purpose model for forecast through combining the result of each individual tree and implementing voting by majority for classification while averaging for regression. [30]

Bootstrap aggregating (bagging): Represent a basic ensemble learning method. The objective is to raise accuracy and stability of machine learning algorithm specifically when faced with overfitting and high variance. Bootstrap Sampling refer to the statistical technique for resampling with a replacement. Generate multiple dataset (sample) with the initial training dataset is exactly what refer to in this setting. These sample have all the same size just like the original dataset yet they were created through replacing some from the original dataset instance at random. In simple word a case may show up multiple times in a sample or could not come up at all. [31]

Provided the training set $X = \{x_1, x_2, ..., x_n\}$ with the respective responses $Y = \{y_1, y_2, ..., y_n\}$, the Bagging algorithm repeated B times performs as follows:
For every iteration b where $b = 1, 2, ..., B$ it randomly samples with replacement n training examples with X and Y, thus making bootstrap sample X_b and Y_b. The decision tree f_b is then trained on (X_b, Y_b). This method yield B decision trees based on a unique dataset sample due to randomization with Bootstrap Sampling. [31]

Decision Tree Algorithm: A vital component of a lot of machine learning algorithm including Random Forest tend to be decision tree. They are utilised for task including classification as well as regression. For it to generate a decision tree which will resemble a tree data has to be divided according to particular criteria. **Information Gain (Entropy)** and **Gini Impurity** are two of the significant parameter deployed in the dividing process. [32]

Gini Impurity is used in the decision tree algorithm to quantify probability of erroneous categorization in the occurrence that random data point is classified depending on subset class label distribution. While at given node Gini Impurity is 0 it depict perfect purity as every component of it are of identical class. Split are chosen in a decision tree method of construction in order to reduce Gini Impurity and optimise the homogeneity of the subset in respect to the target variable. [33]

$$I_G(p) = 1 - \sum_{i=1}^{J} p_i^2 \tag{6.3}$$

where J is number of classes and p_i the proportion of element belonging to the class i in set.

Entropy is measurement of ambiguity, higher entropy value denote more disorder. Entropy evaluate the amount of impurity in data. Information Gain is measure of the way an attribute split the data; it is closely linked to entropy. So as to optimise the decrease in entropy it is calculated through the difference in entropy before and following the split. Basically a decision tree choose a division that will enhances a decrease in uncertainty about the desired variable and improving the uniformity of the resulting subset. [34]

$$H(T) = -\sum_{i=1}^{J} p_i \log_2(p_i) \tag{6.4}$$

where p_i denote the proportion of element in set T belonging to class i, and J is the total number of classe. [34]

Random Subspace method (RSM): A subset of feature are selected at random over splitting at each node on a tree raising the variety of each of decision trees inside the forest. RSM introduces randomness by feature selection compared to bootstrap sampling which does this by picking various subset in data points. Each attribute in decision tree is considered for splitting at each node. However only a random subset of feature are taken to account for each split in the Random Forest. Since different tree will make their decisions in various subset of features this approach lessens the relationship among the trees in the forest. Without substantially increasing bias the reduction in tree to tree correlation leads to a decrease in the models in general variance. [35]

$$\text{Features for splitting at each node} = \text{Choose}(m, M) \tag{6.5}$$

Here, Choose(m, M) shows process for randomly selecting m feature from total M feature available where $m < M$ to know best split at every node of decision tree. [35]

Variance and Bias: As each decision tree typically exhibit considerable **variability** Random Forest are made to fix this issue. The degree of how a machine learning model prediction vary for a particular data point is recognised as the variance. Decision tree often exhibit elevated variance since they are strongly sensitive to the details of training set which may result in overfitting for the given model. So to tackle this problem Random Forest use Random Subspace Method (RSM) and Bootstrap Sampling (BS) as the main ways to generating randomness. Random Forest successfully lower the total variance by averaged the prediction from multiple of such tree. Through combining the error and peculiaritie of each of the tree this aggregation establishe an increasingly reliable and consistent model. [36]

Bias is a term to describe the error that result from using a simplified model to estimate a real world problem. The ensemble models bias in the setting of Random Forest is almost same as that of one decision tree. This is done to ensure this can record complex pattern and nuance of the data as every tree in a forest typically grow to its maximum depth. Yet in contrast to variance, bias does not decrease lower as the total amount of trees in a forest rises. Random Forest model becomes more precise and solid if applied to unknown data as a result of its ability to reduce variation via unpredictability and averages although the bias of the model remain identical to individual tree. [36]

$$\text{Error} = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error} \tag{6.6}$$

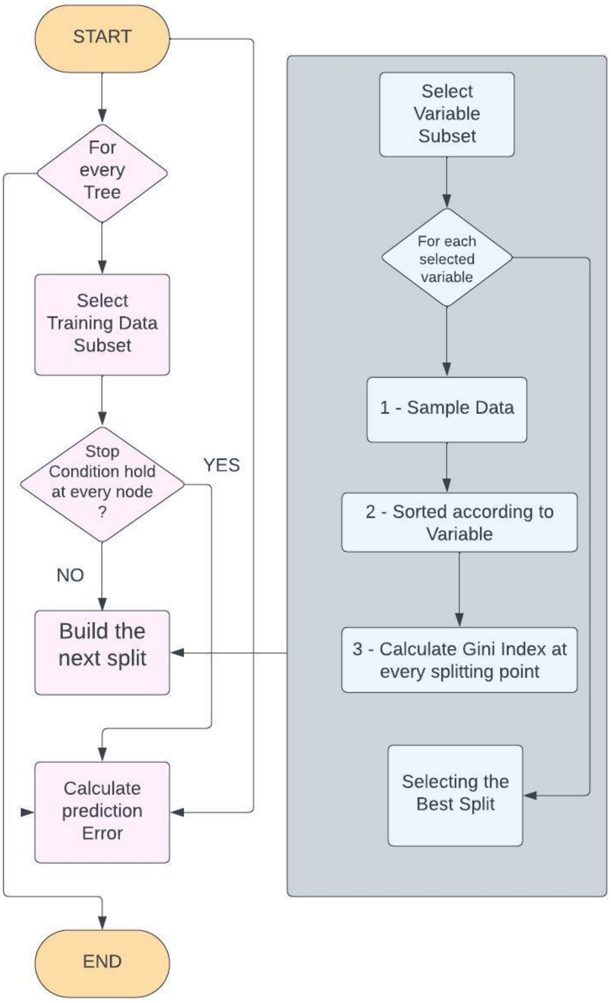

START
For every Tree
Select Training Data Subset
Stop Condition hold at every node ?
YES
NO
Build the next split
Calculate prediction Error
END
Select Variable Subset
For each selected variable
1 - Sample Data
2 - Sorted according to Variable
3 - Calculate Gini Index at every splitting point
Selecting the Best Split

Figure 6.19: Flow chart of Random Forest Algorithm [37]

(c) **Decision Tree:**

Decision trees are used extensively in projects which involve classification and regression. They possess a flowchart like order inside node standing in over test from various attribute branche over test result and leaf nodes to decision results or class label. Several characteristics test take place in the path about the root up to the leaf to generate a rule based framework. The needed phase affecting accuracy happens when the tree split the data at every node into sub nodes. In order to optimise the homogeneity of the ending node these division are based on parameters such as variance for regression and Gini Impurity or Information Gain for classification. [38]

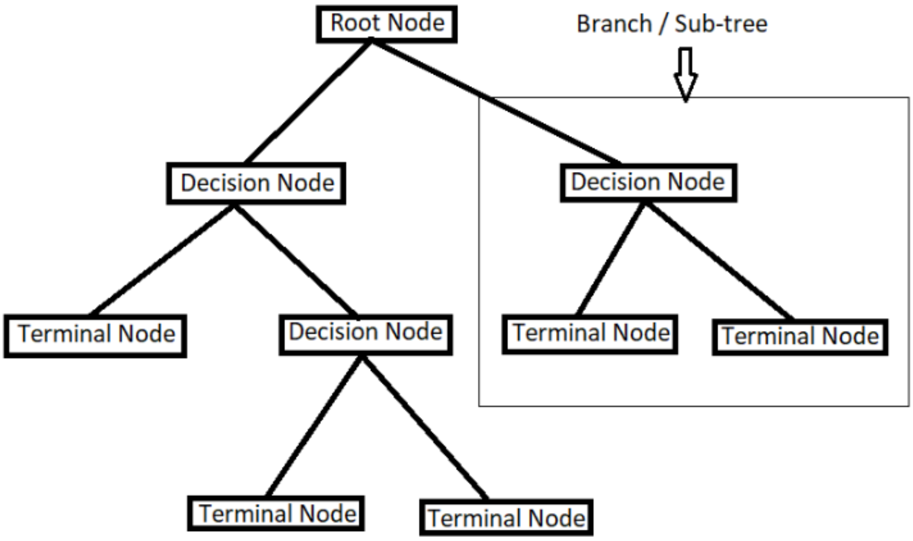


Figure 6.20: Decision Tree Flowchart [38]

Gini Impurity: A decision tree metric called **Gini Impurity** is being used to measure the probability that a component selected arbitrarily is going to be incorrectly classified if the labels are randomly assigned based on the subset label distribution. A node′s "pure" (i.e., how blended its classes have been) indicate could be inferred from the Gini score where 0 denotes perfect purity (all of the node elements belong to a one class). This indicator aids in finding the nodes where the tree will divided making it vital to the CART (Classification and Regression Trees) algorithm. The more effective separation the node attains the lower is the Gini Impurity. [39]
The Gini Impurity I_G of a set S is calculated as:

$$I_G(S) = 1 - \sum_{i=1}^{n} p_i^2 \tag{6.7}$$

where p_i is the proportion of items labeled with class i in the set. [39]

Entropy: It is then assessment of data randomness in the setting of decision trees. It helps in quantifying the datasets disorder or noise. It is harder to derive conclusions from data with greater entropy. When a division is made based on a particular attribute Decision Trees are utilised to figure out which feature provides the most benefit from information gain (reduction in entropy). [40]

The entropy $H(S)$ of a set S is given by:

$$H(S) = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{6.8}$$

where p_i is the proportion of the items labeled with class i. [40]

Information Gain: An essential concept in decision trees is Information Gain especially when applying algorithms like ID3 (Iterative Dichotomiser 3). It computes the reduction in impurity along with entropy prior to and after a dataset splited according to a feature. Essentially it help in choosing the attribute that produce the greatest information gain (or greatest reduction in uncertainty) in each level for the tree for dividing. The idea is to enhance the Information Gain in each split that help with the creation of a profitable and effective tree. [41]

For the dataset S split among attribute A in the subset $S_1, S_2, ..., S_k$ the Information Gain $IG(S, A)$ is:

$$IG(S, A) = H(S) - \sum_{i=1}^{k} \frac{|S_i|}{|S|} H(S_i) \tag{6.9}$$

Here $|S|$ is total number of samples in dataset S and $|S_i|$ represent number of samples in the subset S_i. [41]

Decision Tree Algorithm: In the effort for determining an optimal split over different class Decision Tree Algorithm divide features at each node. This method make sure that each partition that occur at every point as homogeneous as attainable which implies that division criteria must be certain. This approach in efficiently break down hard decision making procedure by breaking down complex dataset to portion that are easier to manage. [42]

The below given table shows 6 Decision Tree Algorithms with Classification and Description.

Algorithm Name	**Classification**	**Description**
CART (Classification and Regression Trees)	Uses Gini Index as a metric.	Numerical splitting allows us to build the CART based tree.
ID3 (Iterative Dichotomiser 3)	Uses Entropy function and Information gain as metrics.	The discrete values are the only trouble. Thus the continuous dataset needs to be classified inside discrete dataset.
C4.5	The improved version on ID 3	Manages continuous as well as discrete dataset. Can also handle incomplete dataset. The issue of overfiltering is fixed by the "PRUNNING" technique.
C5.0	Improved version of the C4.5	C5.0 offers a choice to allocate the case statistically among the outcomes or just estimate missing values as a function of other attributes.
CHAID (Chi-square Automatic Interaction Detector)	Predates the original ID3 implementation.	TThis sort of decision tree is used for a nominal scaled variable. The approach extracts the variable that is dependent from a dataset's categorised variables.
MARS (Multi-adaptive regression splines)	Used to find the best split.	We can use the MARS-based regression tree to get the optimal split.

Table 6.1: Summary of Decision Tree Algorithms [42]

CHAPTER 7

Discussion and Results

7.1 Comparative Analysis of Machine Learning

(a) **Results of Gaussian Naive Bayes Algorithm:** A Gaussian Naive Bayes classifiers performance in various datasets reveals the unique benefits and drawbacks. With regards to accuracy (93.89 percent) and precision (10.49 percent) Unbalanced dataset performs best suggesting that it is the best option for making accurate predictions in general and for predicting the positive classes. Contrary to this SMOTE Oversampled dataset performs better than the others with regard to of recall (99.33 percent) as well as AUC (95.79 percent)illustrating its ability to identify between classes and identify the positive instances. Despite having a moderate AUC (68.02 percent) the Undersampled dataset performs insufficiently with regard to precision (7.95 percent) and F1 Score (13.60 percent) that may indicate an elevated rate of false positives and a poor compromise between precision and recall. In short, the SMOTE Oversampled data is preferred at recognising and differentiating positive cases despite the fact the Unbalanced data is more accurate.

Data Type	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**AUC**
Unbalanced	0.9389	0.1049	0.2855	0.1534	0.6186
Undersample	0.8801	0.0795	0.4720	0.1360	0.6802
SMOTE Oversample	0.9239	0.2037	0.9933	0.3381	0.9579

Table 7.1: Comparison of Gaussian Naive Bayes Performance

(b) **Results of Random Forest Classifier:** Evaluating the efficiency of Random Forest classifier on the a range of datasets and comeimg across distinct results. Through an accuracy of 99.47 percent Unbalanced dataset results most accurate predicts the right answers. It is clear that the predictions for Sepsis negative cases are very accurate with precision achieving the highest level for Unbalanced dataset (96.44 percent). Conversely, Undersampled dataset displays a significant contrast with precision following considerably on 20.23 percent but stands out with the highest recall (97.59 percent), highlighting its ability to capture nearly all true positive cases. Regardless of becoming least precise (7.59 percent) and accurate (88.14 percent) the SMOTE Oversampled dataset continues to have an acceptable recall rate (45.33 percent). Once again the Unbalanced dataset has the finest F1 Score (84.77 percent) suggesting an appealing trade-off between recall and precision.

On the opposite hand Undersampled dataset works very well (94.86 percent) according to the AU ROC suggesting greater capacity to distinguish between classes. The SMOTE Oversampled dataset is falling notably at 67.16 percent while the Imbalanced dataset possesses an acceptable AUC of 87.78 percent. While the Undersampled data indicates a high degree of sensitivity for recognising cases that are positive despite a decrease in precision and accuracy Imbalanced dataset works firmly throughout the majority indicators particularly when accurately identifying the majority class which is negative for Sepsis. Despite with an average recall rate SMOTE Oversampled data usually ranks poorly throughout the majority measurements particularly for precision and the F1 Score.

Data Type	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**AUC**
Unbalanced	0.9947	0.9644	0.7562	0.8477	0.8778
Undersample	0.9225	0.2023	0.9759	0.3351	0.9486
SMOTE Oversample	0.8814	0.0759	0.4533	0.1301	0.6716

Table 7.2: Comparison of Random Forest Performance

(c) **Results of Decision Tree Classifier:** The performance measurements of Decision Tree Classifier indicate distinct variations among the datasets. With an accuracy of 99.05 percent Unbalanced dataset is the most noteworthy compared to SMOTE Oversampled dataset getting in second at 98.00 percent just 1 percent difference. The Undersampled dataset on the opposite hand falls behind on 82.64 percent indicating a significant drop on overall predictive credibility. SMOTE Oversampled dataset encounters a substantial drop to 49.13 percent on precision whereas Unbalanced dataset achieves 73.65 percent suggesting a high probability of accurate positive predictions. With a decline to 9.62 percent Undersampled dataset had the least precision. Recall has the superior performance for the Undersampled dataset (91.39 percent) which is better at identifying the true positives compared toother datasets (12 percent lower for the Unbalanced dataset and greater than 26 percent below the SMOTE Oversampled dataset).

Unbalanced dataset prevails with regard to of F1 Score that combines recall and precision having a score of 76.30 percent. SMOTE Oversampled and Undersampled datasets on the contrary had record scores of 56.03 percent and 17.40 percent respectively indicating a decreasing balance between the precision - recall arrangement. Unbalanced dataset offers the highest AU ROC which is 89.30 percentfollowing by the Undersampled dataset (which in turn follows slightly) and SMOTE Oversampled dataset which lags by around 5 percent. Although the Undersampled dataset has an important recall with lower accuracy and precision Unbalanced dataset continuously performs superior among accuracy,precision, F1 Score and AU ROC. While retaining an acceptable performance SMOTE Oversampled dataset scores badly with regard to accuracy and AU ROC.

Data Type	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**AUC**
Unbalanced	0.9905	0.7365	0.7915	0.7630	0.8930
Undersampling	0.8264	0.0962	0.9139	0.1740	0.8693
SMOTE Oversampling	0.9800	0.4913	0.6518	0.5603	0.8192

Table 7.3: Comparison of Decision Tree Classifier Performance

7.2 Area under the ROC Curve for different data types

In line with [43] F1 score or F measure might not accurately reflect the behaviour of the classification on the minority class due to the datasets high imbalance comprises 2.17 percent of patients classified as having severe sepsis and 97.83 percent of patients classified as having no sepsis. AUC ROC evaluates how well the model separates groups with no influence by the percentage of class distribution even though it tends to be less sensitive to the class imbalance. It provides a more comprehensive evaluation of the algorithm's efficiency by evaluating the performance across all potential classification thresholds. Therefore AUC ROC reflects the model's efficacy throughout a range of class splits it turns into an additional reliable metric on datasets which are imbalanced. On the case of imbalanced data [43] recommends using the 'β' varied F-measure i which 'β' is a coefficient that adjusts the relative significance for precision versus recall instead of the traditional F-measure.

7.2.1 AU ROC for Imbalanced Data

With the AU ROC of 0.89 Decision Tree had the finest discriminative power followed closely by Random Forest on 0.88 showing that both of them work well at distinguishing among the classes. In contrast to this Gaussian Naive Bayes does considerably bad as shown by the lower AU ROC of 0.62. The black line is the line of no discrimination which refers to a model to no capability above randomness and serves to measure curves. The efficiency of Random Forest and Decision Tree model are shown by the substantial disparity in their curves from line of no discrimination whereas Gaussian Naive Bayes models curve suggest that it had difficulty in successfully separating classes and this is especially notable with the imbalanced dataset.

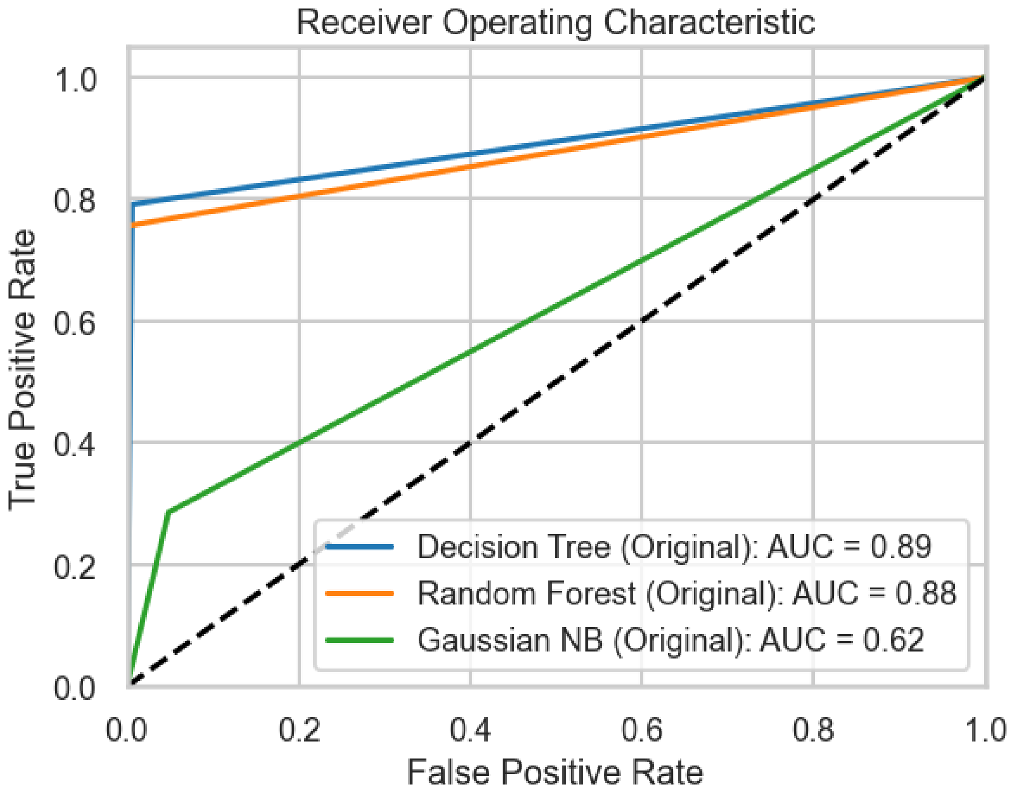


Figure 7.1: Comparing AU ROC for Imbalanced Data

7.2.2 AU ROC for Undersampled Data

Having a remarkable AU ROC of 0.87 and an acceptable true positive rate paired by a relatively low false positive rate Decision Tree is clearly able to correctly identifying minority class after under sampling. Through a notable AU ROC of 0.95 Random Forest model remains out and seems to gain most of the under sampling method. This is so because it not only retains the high true positive rate but it also minimises false positive rate. However in spite of having the capacity to recognise true positives more precisely than prior to undersampling Gaussian Naive Bayes model still likely generates more false positives compared to other models showing its inability to distinguish among classes. In contrast the models AU ROC improves somewhat to 0.68. All three curves of ML Algorithms showed the distances to the diagonal line indicate that under sampling had increased each models ability to recognise the minority class correctly whereas taking into account false positives.

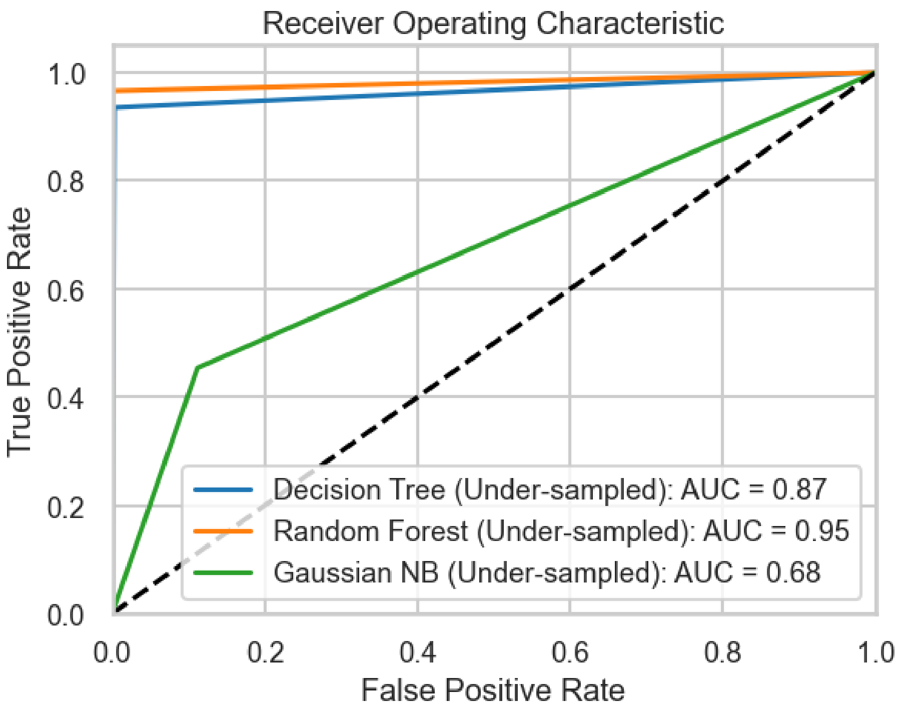


Figure 7.2: Comparing AU ROC for Imbalanced Data

7.2.3 AU ROC for SMOTE Oversampled Data

Gaussian Naive Bayes algorithm had shown a significant enhancement according to oversampled datasets ROC curve with a AU ROC increase to 0.96. This implies a substantial true positive rate along with a regulated false positive rate suggesting the oversampling technique enhanced the classifiers capacity for classification. On contrary AU ROC of Random Forest model has now decreased to 0.67 which shows a lower efficiency along with a higher false positive rate. This might be an indication of oversampling resulting in overfitting a frequent problem with this method. With an AU ROC of 0.82 Decision Trees performance is fairly stable and shows an acceptable proportion of true positive and false positive rates albeit one which is less apparent than that of the Gaussian Naive Bayes model. ROC curves indicate the way SMOTE oversampling may enhance the outcome of certain classifiers. This appears clearly when seeking at Gaussian Naive Bayes curve which becomes more steep as it come close to the upper left corner indicating either low false positive rate and high true positive rate. Conversely Random Forest models curve that approaches a diagonal line for no discrimination more closely demonstrates a relatively poorer outcome following oversampling which highlights the complex nature of the models behaviour when resampling techniques are used on the imbalanced dataset.

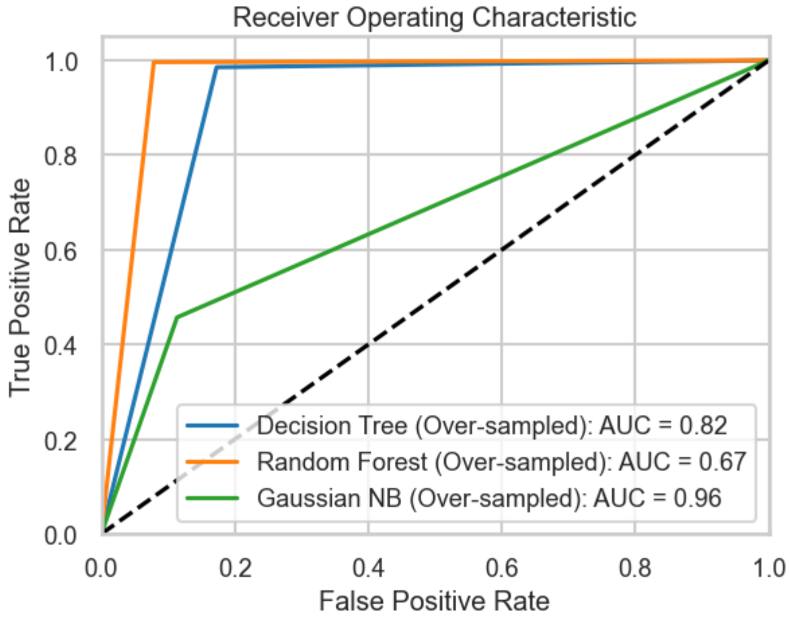


Figure 7.3: Comparing AU ROC for SMOTE Oversampled Data

CHAPTER 8

Conclusion and Future Scope

8.1 Conclusion

All the 3 Machine Learning Algorithm exhibit different benefits and drawbacks with Gaussian Naive Bayes, Random Forest and Decision Tree classifiers. Gaussian Naive Bayes Algorithm does well in oversampling while troubles on recall and AU ROC especially in undersampled data showing much larger proportion of false positives while the performance get better in SMOTE Oversampled Data. On Imbalanced dataset Random Forest Classifier does very well with regard to of precision and accuracy. Especially in negative Sepsis cases however undersampling leads to decrease in precision and F1 Score whereas oversampling produces a rise in AU ROC which may indicate overfitting. While Decision Tree Classifier differentiates among classes correctly due to the strong precision and accuracy upon unbalanced datasets consistent high AU ROC and excellent recall and precision balance throughout different methods of sampling.

In conclusion, effectiveness for measurements along with sampling strategies had considerable impact on machine learning model performance. Although undersampling enhances sensitivity for positive cases with a loss in precision, SMOTE oversampling increases Gaussian Naive Bayes classification power but might cause overfitting in the Random Forest. While undersampling increases sensitivity to positive cases by losing for precision. For unbalanced datasets, the AU ROC metric shows to be a more reliable metric because it offers a comprehensive evaluation for a models performance. Finally the classifier and resampling method must be chosen according to the specific requirements for the task; Gaussian Naive Bayes is favoured for accuracy at unbalanced datasets, Random Forest at precision and efficiency in general and Decision Trees offer stable and balanced performance throughout a variety of datasets. This emphasises the significance of choosing the correct approach for maximising the effectiveness of the model.

8.2 Future Scope

The use of machine learning (ML) for sepsis prediction holds immense potential in future for medical treatment particularly for Intensive Care Unit (ICU) setting. The development of more sophisticated ML algorithm constitutes one among the main area of focus for Sepsis prediction. Deep learning model algorithm are expected to efficiently handle the complexity of sepsis data. Prediction accuracy might significantly enhance through these model capacity to recognise intricate trend in huge dataset which traditional statistical method may overlook. Another key field for growth is the application for Real-time data analysis. Immediate risk assessment are achievable along with the help of ML model for continuous monitoring for patient vitals parameter which make timely medical intervention possible. Furthermore it is important to deal with privacy and ethical concern associated with patient information as machine learning use in healthcare grow. It is vital for developing safe systems that put the confidentiality of patient as priority. Machine learning based sepsis prediction model ensuring that they satisfy clinical requirement for them to be widely utilised. Basically the application of ML for sepsis prediction in the years to come will require global, moral, friendly and technological development that will enhance care for patient in critical healthcare scenarios.

Bibliography

[1] A. Artero, J. M. Nogueira, and R. Zaragoza. *Epidemiology of Severe Sepsis and Septic Shock*. INTECH Open Access Publisher, 2012. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Epidemiology+of+Severe+Sepsis+and+Septic+Shock&btnG=

[2] T. van der Poll, M. Shankar-Hari, and W. J. Wiersinga. *The immunology of sepsis*, 54(11):2450–2464, 2021. Elsevier. https://www.cell.com/immunity/pdf/S1074-7613(21)00449-0.pdf

[3] R. P. Wenzel, M. R. Pinsky, R. J. Ulevitch, and L. Young. *Current understanding of sepsis*, Clinical Infectious Diseases,pages 407–412, 1996. JSTOR. https://www.jstor.org/stable/4459276

[4] A. Alanazi. *Using machine learning for healthcare challenges and opportunities*, Informatics in Medicine Unlocked,30:100924, 2022. Elsevier. https://www.sciencedirect.com/science/article/pii/S2352914822000739

[5] P. Sujatha and K. Mahalakshmi. *Performance evaluation of supervised machine learning algorithms in prediction of heart disease* In 2020 IEEE International Conference for Innovation in Technology (INOCON), pages 1–7, 2020. IEEE. https://ieeexplore.ieee.org/abstract/document/9298354

[6] F. Y. Osisanwo, J. E. T. Akinsola, O. Awodele, J. O. Hinmikaiye, O. Olakanmi, J. Akinjobi, et al. *Supervised machine learning algorithms: classification and comparison* International Journal of Computer Trends and Technology (IJCTT), 48(3):128–138, 2017. Seventh Sense Research Group. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Supervised+machine+learning+algorithms%3A+classification+and+comparison&btnG=

[7] M. Shetty, S. M. Alex, M. Moni, F. Edathadathil, P. Prasanna, V. Menon, V. P. Menon, P. Athri, and G. Srinivasa. *A Machine Learning Understanding of Sepsis* In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pages 2175–2179, 2021. IEEE. https://ieeexplore.ieee.org/abstract/document/9629558

[8] I. Taneja, G. L. Damhorst, C. Lopez-Espina, S. D. Zhao, R. Zhu, S. Khan, K. White, J. Kumar, A. Vincent, L. Yeh, et al. *Diagnostic and prognostic capabilities of a biomarker and EMR-based machine learning algorithm for sepsis* Clinical and Translational Science, 14(4):1578–1589, 2021. Wiley Online Library. https://ascpt.onlinelibrary.wiley.com/doi/full/10.1111/cts.13030

[9] M. D. Font, B. Thyagarajan, and A. K. Khanna. *Sepsis and Septic Shock–Basics of diagnosis, pathophysiology and clinical decision making* Medical Clinics, 104(4):573–585, 2020. Elsevier. https://www.medical.theclinics.com/article/S0025-7125(20)30019-5/fulltext

[10] K. Tong-Minh, I. Welten, H. Endeman, T. Hagenaars, C. Ramakers, D. Gommers, E. van Gorp, and Y. van der Does. *Predicting mortality in adult patients with sepsis in the emergency department by using combinations of biomarkers and clinical scoring systems: a systematic review* BMC Emergency Medicine, 21:1–11, 2021. Springer. https://link.springer.com/article/10.1186/s12873-021-00461-z

[11] S. Lambden, P. F. Laterre, M. M. Levy, and B. Francois. *The SOFA scoreâdevelopment, utility and challenges of accurate assessment in clinical trials* In Critical Care, 23(1):1–9, 2019. BioMed Central. https://ccforum.biomedcentral.com/articles/10.1186/s13054-019-2663-7

[12] P. E. Marik and A. M. Taeb. *SIRS, qSOFA and new sepsis definition* Journal of Thoracic Disease, 9(4):943, 2017. AME Publications. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5418298/

[13] R. K. Chakraborty and B. Burns. *Systemic inflammatory response syndrome* 2019. https://europepmc.org/article/nbk/nbk547669

[14] M. J. Pettinati, G. Chen, K. S. Rajput, and N. Selvaraj. *Practical machine learning-based sepsis prediction* In 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pages 4986–4991, 2020. IEEE. https://ieeexplore.ieee.org/abstract/document/9176323

[15] Y. S. Kwon and M. S. Baek. *Development and validation of a quick sepsis-related organ failure assessment-based machine-learning model for mortality prediction in patients with suspected infection in the emergency department* Journal of Clinical Medicine, 9(3):875, 2020. MDPI
https://www.mdpi.com/2077-0383/9/3/875

[16] P.-Y. Hsu and C. Holtz. *A comparison of machine learning tools for early prediction of sepsis from ICU data* In 2019 Computing in Cardiology (CinC), pages 1, 2019. IEEE. https://ieeexplore.ieee.org/abstract/document/9005834

[17] E. Shakeri, E. A. Mohammed, Z. S. HA, and B. Far. *Exploring features contributing to the early prediction of sepsis using machine learning* In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pages 2472–2475, 2021. IEEE. https://ieeexplore.ieee.org/abstract/document/9630317

[18] C. J. Paoli, M. A. Reynolds, M. Sinha, M. Gitlin, and E. Crouser. *Epidemiology and costs of sepsis in the United Statesâan analysis based on timing of diagnosis and severity level* Critical Care Medicine, 46(12):1889, 2018. Wolters Kluwer Health.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6250243/

[19] M. Reyna, C. Josef, R. Jeter, S. Shashikumar, B. Moody, M. B. Westover, A. Sharma, S. Nemati, and G. D. Clifford. *Early Prediction of Sepsis from Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019* Published: Aug. 5, 2019. https://physionet.org/content/challenge-2019/1.0.0/

[20] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona. *A survey on missing data in machine learning* Journal of Big Data, 8(1):1–37, 2021. SpringerOpen. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00516-9

[21] J. M. Jerez, I. Molina, P. J. GarcÃa-Laencina, E. Alba, N. Ribelles, M. MartÃn, and L. Franco. *Missing data imputation using statistical and machine learning methods in a real breast cancer problem* Artificial Intelligence in Medicine, 50(2):105–115, 2010. Elsevier. https://www.sciencedirect.com/science/article/abs/pii/S0933365710000679

[22] T. Al-Shehari and R. A. Alsowail. *An insider data leakage detection using one-hot encoding, synthetic minority oversampling and machine learning techniques* Entropy, 23(10):1258, 2021. MDPI. https://www.mdpi.com/1099-4300/23/10/1258

[23] S. Kumar and I. Chong. *Correlation analysis to identify the effective data in machine learning: Prediction of depressive disorder and emotion states* International Journal of Environmental Research and Public Health, 15(12):2907, 2018. MDPI. https://www.mdpi.com/1660-4601/15/12/2907

[24] S.-J. Yen and Y.-S. Lee. *Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset* In Proceedings of the International Conference on Intelligent Computing, ICIC 2006, Kunming, China, August 16–19, pages 731–740, 2006. Springer. https://link.springer.com/chapter/10.1007/978-3-540-37256-1_89

[25] A. Gosain and S. Sardana. *Handling class imbalance problem using oversampling techniques: A review* In Proceedings of the 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pages 79–85, 2017. IEEE. https://ieeexplore.ieee.org/abstract/document/8125820

[26] H. Kamel, D. Abdulah, and J. M. Al-Tuwaijari. *Cancer classification using Gaussian Naive Bayes algorithm* In Proceedings of the 2019 International Engineering Conference (IEC), pages 165–170, 2019. IEEE. https://ieeexplore.ieee.org/abstract/document/8950650

[27] G. Tzanos, C. Kachris, and D. Soudris. *Hardware acceleration on Gaussian Naive Bayes machine learning algorithm* In Proceedings of the 2019 8th International Conference on Modern Circuits and Systems Technologies (MOCAST), pages 1–5, 2019. IEEE. https://ieeexplore.ieee.org/document/8741875

[28] A. H. Jahromi and M. Taheri. *A non-parametric mixture of Gaussian Naive Bayes classifiers based on local independent features* In Proceedings of the 2017 Artificial Intelligence and Signal Processing Conference (AISP), pages 209–212, 2017. IEEE. https://ieeexplore.ieee.org/abstract/document/8324083

[29] D. Berrar. *Bayes theorem and naive Bayes classifier* In Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, volume 403, pages 412, 2018. Elsevier Science Publisher Amsterdam, The Netherlands. https://books.google.co.uk/books?hl=en&lr=&id=rs51DwAAQBAJ&oi=fnd&pg=PA403&dq=Bayes%E2%80%99+Theorem+and+Naive+Bayes+Classifier&ots=q-UU7aTnQS&sig=oB6eM8h4wAY4fUnaLJFA5bPSgtw&redir_esc=y#v=onepage&q=Bayes%E2%80%99%20Theorem%20and%20Naive%20Bayes%20Classifier&f=false

[30] I. Reis, D. Baron, and S. Shahaf. *Probabilistic random forest: A machine learning algorithm for noisy data sets* The Astronomical Journal, 157(1):16, 2018. IOP Publishing. https://iopscience.iop.org/article/10.3847/1538-3881/aaf101/meta

[31] T.-H. Lee, A. Ullah, and R. Wang. *Bootstrap aggregating and random forest In Macroeconomic Forecasting in the Era of Big Data: Theory and Practice* pages 389–429, 2020. Springer. https://link.springer.com/chapter/10.1007/978-3-030-31150-6_13

[32] N. M. Abdulkareem and A. M. Abdulazeez. *Machine learning classification based on Random Forest algorithm: A review*. International Journal of Science and Business, 5(2):128–142, 2021. IJSAB International. https://ideas.repec.org/a/aif/journl/v5y2021i2p128-142.html

[33] L. Jiang, B. Zhang, Q. Ni, X. Sun, and P. Dong. *Prediction of SNP sequences via Gini impurity based gradient boosting method* IEEE Access, 7:12647–12657, 2019. IEEE. https://ieeexplore.ieee.org/abstract/document/8615995

[34] R. A. Disha and S. Waheed. *Performance analysis of machine learning models for intrusion detection system using Gini Impurity-based Weighted Random Forest (GIWRF)*

feature selection technique , 5(1):1, 2022. Springer. https://link.springer.com/article/10.1186/s42400-021-00103-8

[35] Q. Wang, F. Wang, Z. Li, P. Jiang, F. Ren, and F. Nie. *Efficient random subspace decision forests with a simple probability dimensionality setting scheme* Information Sciences, 638:118993, 2023. Elsevier. https://www.sciencedirect.com/science/article/abs/pii/S0020025523005741

[36] P. Ramchandani. *Random Forests and the Bias-Variance Tradeoff* Towards Data Science, Oct 10, 2018. https://towardsdatascience.com/random-forests-and-the-bias-variance-tradeoff-3b77fee339b4

[37] N. Kunhare, R. Tiwari, and J. Dhar. *Particle swarm optimization and feature selection for intrusion detection system* Sadhana, 45:1–14, 2020. Springer. https://link.springer.com/article/10.1007/s12046-020-1308-5

[38] H. Sharma, S. Kumar, et al. *A survey on decision tree algorithms of classification in data mining* International Journal of Science and Research (IJSR), 5(4):2094–2097, 2016. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=A+survey+on+decision+tree+algorithms+of+classification+in+data+mining&btnG=

[39] Y. Yuan, L. Wu, and X. Zhang. *Gini-impurity index analysis* IEEE Transactions on Information Forensics and Security, 16:3154–3169, 2021. IEEE. https://ieeexplore.ieee.org/abstract/document/9420091

[40] Q. R. Wang and C. Y. Suen. *Analysis and design of a decision tree based on entropy reduction and its application to large character set recognition* IEEE Transactions on Pattern Analysis and Machine Intelligence, (4):406–417, 1984. IEEE. https://ieeexplore.ieee.org/abstract/document/4767546

[41] P. Guleria, N. Thakur, and M. Sood. *Predicting student performance using decision tree classifiers and information gain* In Proceedings of the 2014 International Conference on Parallel, Distributed and Grid Computing, pages 126–129, 2014. IEEE. https://ieeexplore.ieee.org/abstract/document/7030728

[42] H. H. Patel and P. Prajapati. *Study and analysis of decision tree based classification algorithms* International Journal of Computer Sciences and Engineering, 6(10):74–78, 2018. ISROSET: International Scientific Research Organization for Science, Engineering and Technology. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Study+and+analysis+of+decision+tree+based+classification+algorithms&btnG=

[43] M. Bekkar, H. K. Djemaa, and T. A. Alitouche. *Evaluation measures for models assessment over imbalanced data sets.* J Inf Eng Appl, 3(10), 2013. https://eva.fing.edu.uy/pluginfile.php/69453/mod_resource/content/1/7633-10048-1-PB.pdf