# Machine Learning

# Group Project Report

# Sepsis Prediction from Clinical Data

Artificial Intelligence and Data Science Batch 2022-23
Jio Institute
Quarter-2

---

**Group Members:**
1. Anupriya Dhiman(23PGAI0022)
2. Pranit Malik(23PGAI0043)
3. Sagar S(23PGAI0076)
4. Twinkle Chavda(23PGAI0025)

# Table of Contents

## Introduction

Sepsis is a possibly perilous condition that happens as a reaction to body's reaction to contamination. At the point when the reaction for the body's reaction turns out badly, it could cause tissue harm, organ disappointment or might actually cause passing. As per a new report in the US, almost 1.8 million individuals are inclined to sepsis and in excess of 280,000 individuals bite the dust from the condition every year. A well known country's wellbeing safeguarding organization, CDC or Community for Infectious prevention and Counteraction expresses that north of 33% individuals in U.S. medical clinics kick the bucket from Sepsis every year. Early discovery of sepsis is vital for further developing sepsis results. Every hour of deferred treatment can increment death rate by 4-8%. Responding to questions, for example, what the genuine reason for the condition is, is there a method for anticipating the condition early, way before it has been imagined during the clinical preliminaries are of most extreme significance in this test. Thus, thinking of a decent model that could possibly settle this issue becomes inevitable and significant.

## Objective

This project's objective is to detect sepsis early utilizing frequently accessible clinical data. While late or missed forecasts are potentially life-threatening, early prediction in particular has the potential to save lives. The dataset seems, by all accounts, to be to some degree complex, hence prior to utilizing the strategy, we will initially manage missingness and awkwardness in the assortment as well as attempt to generate new features.

## Dataset

As given in the details of the Project. There are two datasets. One is for Training and Testing and the other one is for validation purpose. The training set is in-turn split so we can perform some validation before testing it on the test set.

## Features

There so many features in the dataset. All the features are as stated below:

['HR','O2Sat', 'Temp', 'SBP', 'MAP', 'DBP', 'Resp', 'EtCO2','BaseExcess', 'HCO3', 'FiO2', 'pH', 'PaCO2', 'SaO2', 'AST', 'BUN','Alkalinephos', 'Calcium', 'Chloride', 'Creatinine', 'Bilirubin_direct','Glucose', 'Lactate', 'Magnesium', 'Phosphate', 'Potassium','Bilirubin_total', 'TroponinI', 'Hct', 'Hgb', 'PTT', 'WBC','Fibrinogen', 'Platelets', 'Age', 'Gender', 'Unit1', 'Unit2','HospAdmTime', 'ICULOS', 'SepsisLabel']

## Data Imputation

We discovered that numerous variables had a large number of missing values while determining the data distribution. In major features, more than ten laboratory values were missing values. Because we couldn't agree on the proper mean of the dataset, data imputation on them was quite difficult. Instead of trying to organize with newer features and then infer the missing ones as a new category, it would be preferable to develop newer features.

## Feature Engineering

The complete set of patient features included was grouped into three categories:

1. Physiological data (e.g., heart rate, temperature, etc.)
2. Laboratory test results (e.g., white blood count, glucose, hematocrit, hemoglobin, creatinine, bicarbonate, PH, and arterial blood gases)
3. Demographics/score (age, HospAdmTime, ICULOS etc.

## New Features

We developed two engineered features to improve predictive performance:

1. Shock index
2. The product of age and systolic blood pressure (ASBP)
3. SIRS score (temperature, HR, Respiratory rate, WBC)
4. Median HR for all individual patients.

## Models Implemented

We implemented various machine learning models to predict sepsis. We got different F1 Score value and Accuracy (in%) for each model, which is mentioned in the tabulated format herein further.

## F1 SCORE FOR THE MODEL

| SR NO | MODEL | IMPUTATION OF VALUES | | | | | |
|---|---|---|---|---|---|---|---|
| | | MEAN | | MEDIAN | | BACKWARD AND FOREWARD FILLING | |
| | SEPSIS LABEL→ | 0 | 1 | 0 | 1 | 0 | 1 |
| 1 | K NEAREST NEIGHBOUR CLASSIFIER | 0.71 | 0.53 | 0.69 | 0.54 | 0.71 | 0.52 |
| 2 | NAÏVE BAYES | 0.70 | 0.54 | 0.69 | 0.54 | 0.71 | 0.52 |
| 3 | XG BOOST | 0.73 | 0.61 | 0.69 | 0.22 | 0.97 | 0.97 |
| 4 | LOGISTIC REGRESSION | 0.71 | 0.63 | 0.71 | 0.63 | 0.70 | 0.66 |
| 5 | RANDOM FOREST | 0.73 | 0.58 | 0.67 | 0.00 | 1 | 1 |
| 6 | SGD CLASSIFIER | 0.70 | 0.54 | 0.69 | 0.54 | 0.71 | 0.52 |

## ACCURACY FOR THE MODEL

| SR NO | MODEL | IMPUTATION OF VALUES | | |
|---|---|---|---|---|
| | | MEAN | MEDIAN | BACKWARD AND FOREWARD FILLING |
| 1 | K NEAREST NEIGHBOUR CLASSIFIER | 63 | 63 | 63 |
| 2 | NAÏVE BIAS | 63 | 63 | 63 |
| 3 | XG BOOST | 68 | 55 | 97 |
| 4 | LOGISTIC REGRESSION | 67 | 67 | 68 |
| 5 | RANDOM FOREST | 67 | 50 | 99 |
| 6 | SGD CLASSIFIER | 63 | 63 | 63 |

## Conclusion

On perusal of the above points and the code, we derived to the conclusion that XG Boost Model got the highest accuracy given dataset. However, there was very high data imbalance. To balance the same, we did the up-sampling and down-sampling to get better prediction.