DAB302 - Health Care Analytics: Final Project Report: Section: 2 - Group: 3

Prediction of Sepsis from Clinical Data

Team Members: Rakeshwar Rao Baggu (0775058)

Riddhi Deshpande (0775846)

Sarabjeet Singh Virk (0775231)

Abstract

Sepsis is a leading cause of death in Hospitals. Early detection and identification of sepsis, which is crucial for lowering mortality, is difficult since many of its signs and symptoms are like those of other, less serious illnesses. We build a Machine Learning model that predicts if a patient in Intensive Care Unit can eventually end up with sepsis infection providing health care practitioners much more info so that they can prioritize and customize the medical treatment according to the need of the patient. We put this method to the test with independent clinical data and found that it has a high

Contents

Introduction:

Sepsis is a leading cause of death in United States hospitals. Sepsis is a condition that occurs when the body's response to infection causes tissue damage, organ failure, or death. Internationally, an estimated 30 million people develop sepsis, and 6 million people die from sepsis each year; an estimated 4.2 million newborns and children are affected (WHO). Early detection and antibiotic treatment of sepsis are critical for improving sepsis outcomes, where each hour of delayed treatment has been associated with roughly an 4-8% increase in mortality.

Sepsis is a life-threatening host response to infection associated with high mortality, morbidity, and health costs. Its management is highly time-sensitive since each hour of delayed treatment increases mortality due to irreversible organ damage. Meanwhile, despite decades of clinical research robust biomarkers for sepsis are missing. Therefore, detecting sepsis early by utilizing the affluence of high-resolution intensive care records has become a challenging machine learning problem.

Doctors in the Medical Intensive care unit find it challenging to treat patients at the right time even before developing Sepsis. In this project, using Data Analysis and Machine learning we will try to predict a patient/s who can potentially develop Sepsis assessing the Vital signs of the patient, this immensely can help doctors in better treating the patients. This would be an important contribution to the society and as well as to the medical care units/hospitals as we can predict the sepsis outcome way before and it can really help save many lives.

Related Work:

Our project relates to an earlier research work that has been completed by “Wenqian Shen”, and “Guanjun Wang” where two integrated tree algorithms are considered, **XGBoost** and **LightGBM**.

The metrics precision, recall, F1-score, Kappa coefficient, and Matthew’s coefficient were used to evaluate the prediction performance of the algorithm. The feature importance score and SHAP value are chosen to explain the model.

For the feature importance score, both XGBoost and LightGBM algorithms can be used to get the output feature importance, which can intuitively reflect the importance of each feature in the data set through the score. The calculation formula of feature importance is shown below, and the importance of feature in the entire model is

$$\widehat{I}_j^2 = \frac{1}{M}\sum_{m=1}^{M}\widehat{I}_j^2(T_m)$$

Where M is the number of trees in the model, and T_m represents the m^{th} tree.

The Feature importance on a single tree is determined by the below formula.

$$\widehat{I_j^2}(T) = \sum_{t=1}^{L-1} \widehat{I_t^2} I(v_t = j)$$

Another research work by **JMIR publications** uses minimal electronic patient data as input to the Machine learning model and predict the Sepsis in patients admitted in the Medical Intensive Care Unit.
They have used '**Insight**' a machine learning classification system developed by a company named '**Dascena'** that uses multivariable combinations of obtained patient data. It uses patient vital signs data like peripheral capillary oxygen saturation, Glasgow Coma Score, and age to predict Sepsis using the retrospective Multiparameter Intelligent Monitoring in Intensive Care. Patients aged 15 years or older are considered for this analysis.
They have compared the classification performance of InSight Vs quick sequential organ failure assessment (qSOFA), modified early warning score (MEWS), systemic inflammatory response syndrome (SIRS), simplified acute physiology score (SAPS) II, and sequential organ failure assessment (SOFA) to determine whether patients will develop sepsis at a fixed period before onset.

The result of this study was that In a test dataset with 11.3% sepsis prevalence, InSight produced superior classification performance compared with the alternative scores as measured by area under the receiver operating characteristic curves (AUROC). The Machine learning model was 88% accurate in predicting the sepsis outcome.

Dataset:

The dataset we have used in our project is sourced from Physio Net and it contains 60,000 patients' data from three separate hospital systems and it contains 43 Feature variables.

Below mentioned are the list of feature variables from the data set:

Vital signs (columns 1-8)	
HR	Heart rate (beats per minute)
O2Sat	Pulse oximetry (%)
Temp	Temperature (Deg C)
SBP	Systolic BP (mm Hg)
MAP	Mean arterial pressure (mm Hg)
DBP	Diastolic BP (mm Hg)
Resp	Respiration rate (breaths per minute)
EtCO2	End tidal carbon dioxide (mm Hg)
Laboratory values (columns 9-34)	
BaseExcess	Measure of excess bicarbonate (mmol/L)
HCO3	Bicarbonate (mmol/L)
FiO2	Fraction of inspired oxygen (%)
pH	N/A
PaCO2	Partial pressure of carbon dioxide from arterial blood (mm Hg)
SaO2	Oxygen saturation from arterial blood (%)
AST	Aspartate transaminase (IU/L)
BUN	Blood urea nitrogen (mg/dL)
Alkalinephos	Alkaline phosphatase (IU/L)
Calcium	(mg/dL)
Chloride	(mmol/L)
Creatinine	(mg/dL)
Bilirubin_direct	Bilirubin direct (mg/dL)
Glucose	Serum glucose (mg/dL)
Lactate	Lactic acid (mg/dL)
Magnesium	(mmol/dL)
Phosphate	(mg/dL)
Potassium	(mmol/L)
Bilirubin_total	Total bilirubin (mg/dL)
TroponinI	Troponin I (ng/mL)
Hct	Hematocrit (%)
Hgb	Hemoglobin (g/dL)
PTT	partial thromboplastin time (seconds)
WBC	Leukocyte count (count*10^3/µL)
Fibrinogen	(mg/dL)
Platelets	(count*10^3/µL)
Demographics (columns 35-40)	
Age	Years (100 for patients 90 or above)
Gender	Female (0) or Male (1)
Unit1	Administrative identifier for ICU unit (MICU)
Unit2	Administrative identifier for ICU unit (SICU)
HospAdmTime	Hours between hospital admit and ICU admit
ICULOS	ICU length-of-stay (hours since ICU admit)

Project Architecture / Work Flow:

- **Extraction**: Extracting the dataset from the Physio Net Repository.
- **Transformation**: The data set has missing values and hence we replaced all the missing values with the mean of the respective feature variables. Further, transformed the feature variable Gender from binary values of '0' and '1' to categorical values as Males (1) and Females (0).
- **Load**: Loaded the data into the python for analysis.

- **Analyse**: Performed the Exploratory Data Analysis, perform Co-relation analysis and pick the co-related feature variable for further analysis.

- **Visualise**: Perform Visualisations to derive key Insights.

- **Build ML Models**:

  Step :1 - Use *DABL* to build and assess various ML models and pick the best performing ML model. In our case, it is a Decision Trees ML Model.

  Step: 2 - Built a Decision tree baseline model with a base line accuracy of 96%.

  Step: 3 - Tuned the ML Model to an accuracy of 98%.

  Step : 4 – Evaluated and Predicted the Sepsis Outcome successfully.

Methods:

In our project we have chosen **Decision Tree classifier** to build the machine learning model. Picking Decision Tree Classifier model for our project was done based on analysing other machine learning models like **Dummy Classifier**, **Gaussian Naïve Bayes**, **Multinomial Naïve Bayes**, and **Decision Tree Classifier** using **Data Analysis Baseline Library (DABL) :** A python library that can be used to automate tasks. In our case, we have used DABL to develop 5 Machine learning models without a needing to rewrite the code, DABL evaluates and assesses the best performing model using various ML Model hyper parameters that suits our data set and will provide the results with the best model.

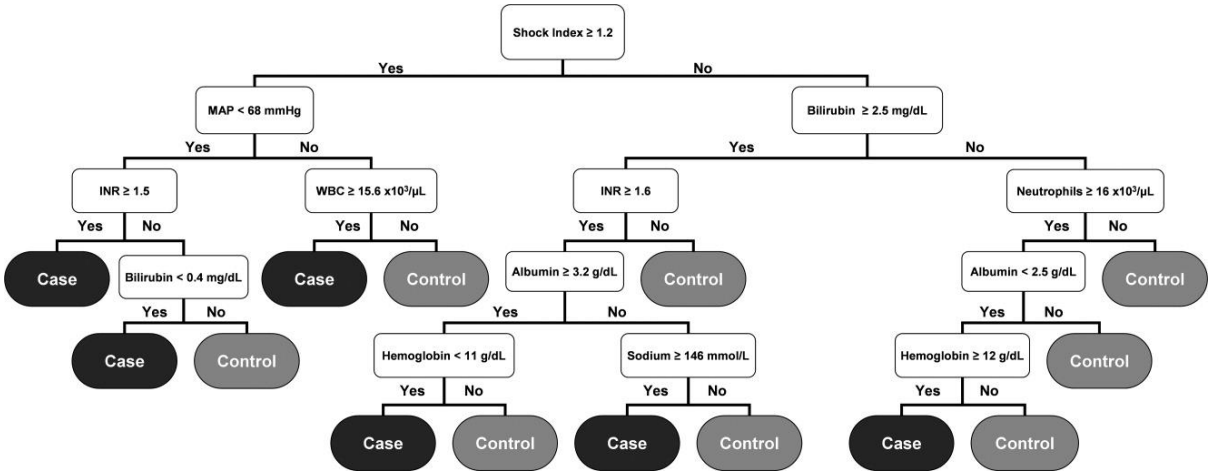
The results of our analysis by comparing various machine learning model are as below:

```
DummyClassifier()
accuracy: 0.981 average_precision: 0.019 roc_auc: 0.500 recall_macro: 0
.500 f1_macro: 0.495
=== new best DummyClassifier() (using recall_macro):
accuracy: 0.981 average_precision: 0.019 roc_auc: 0.500 recall_macro: 0
.500 f1_macro: 0.495

GaussianNB()
accuracy: 0.966 average_precision: 0.032 roc_auc: 0.614 recall_macro: 0
.516 f1_macro: 0.517
=== new best GaussianNB() (using recall_macro):
accuracy: 0.966 average_precision: 0.032 roc_auc: 0.614 recall_macro: 0
.516 f1_macro: 0.517

MultinomialNB()
```

```
accuracy: 0.981 average_precision: 0.030 roc_auc: 0.597 recall_macro: 0
.500 f1_macro: 0.495
DecisionTreeClassifier(class_weight='balanced', max_depth=1)
accuracy: 0.799 average_precision: 0.023 roc_auc: 0.558 recall_macro: 0
.558 f1_macro: 0.471
=== new best DecisionTreeClassifier(class_weight='balanced', max_depth=
1) (using recall_macro):
accuracy: 0.799 average_precision: 0.023 roc_auc: 0.558 recall_macro: 0
.558 f1_macro: 0.471

DecisionTreeClassifier(class_weight='balanced', max_depth=5)
accuracy: 0.696 average_precision: 0.033 roc_auc: 0.619 recall_macro: 0
.595 f1_macro: 0.438
=== new best DecisionTreeClassifier(class_weight='balanced', max_depth=
5) (using recall_macro):
accuracy: 0.696 average_precision: 0.033 roc_auc: 0.619 recall_macro: 0
.595 f1_macro: 0.438

DecisionTreeClassifier(class_weight='balanced', min_impurity_decrease=0
.01)
accuracy: 0.831 average_precision: 0.022 roc_auc: 0.545 recall_macro: 0
.545 f1_macro: 0.474
LogisticRegression(C=0.1, class_weight='balanced', max_iter=1000)
accuracy: 0.648 average_precision: 0.033 roc_auc: 0.615 recall_macro: 0
.585 f1_macro: 0.418
LogisticRegression(class_weight='balanced', max_iter=1000)
accuracy: 0.648 average_precision: 0.033 roc_auc: 0.615 recall_macro: 0
.585 f1_macro: 0.418

Best model:
DecisionTreeClassifier(class_weight='balanced', max_depth=5)
Best Scores:
accuracy: 0.696 average_precision: 0.033 roc_auc: 0.619 recall_macro: 0
.595 f1_macro: 0.438
Accuracy score 0.6943714597693306
```

A **decision tree** is a tree-like model that serves as a decision-making aid, visually exhibiting decisions as well as their potential outcomes, consequences, and costs. The "branches" of the decision tree can then be easily evaluated and analysed to determine the best course of action.

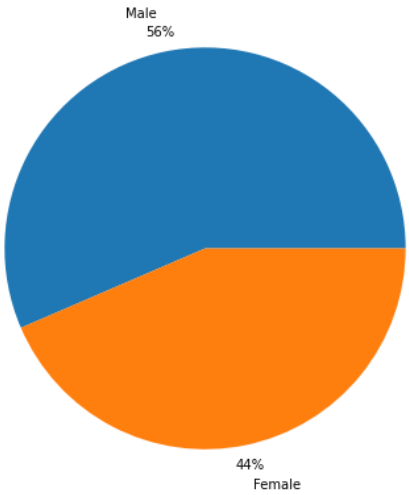
The below picture is an example which briefs decision trees work flow in patient who is diagnosed with septic shock.



Courtesy: https://www.journalofhospitalmedicine.com/jhospmed/article/127109/early-prediction-septic-shock

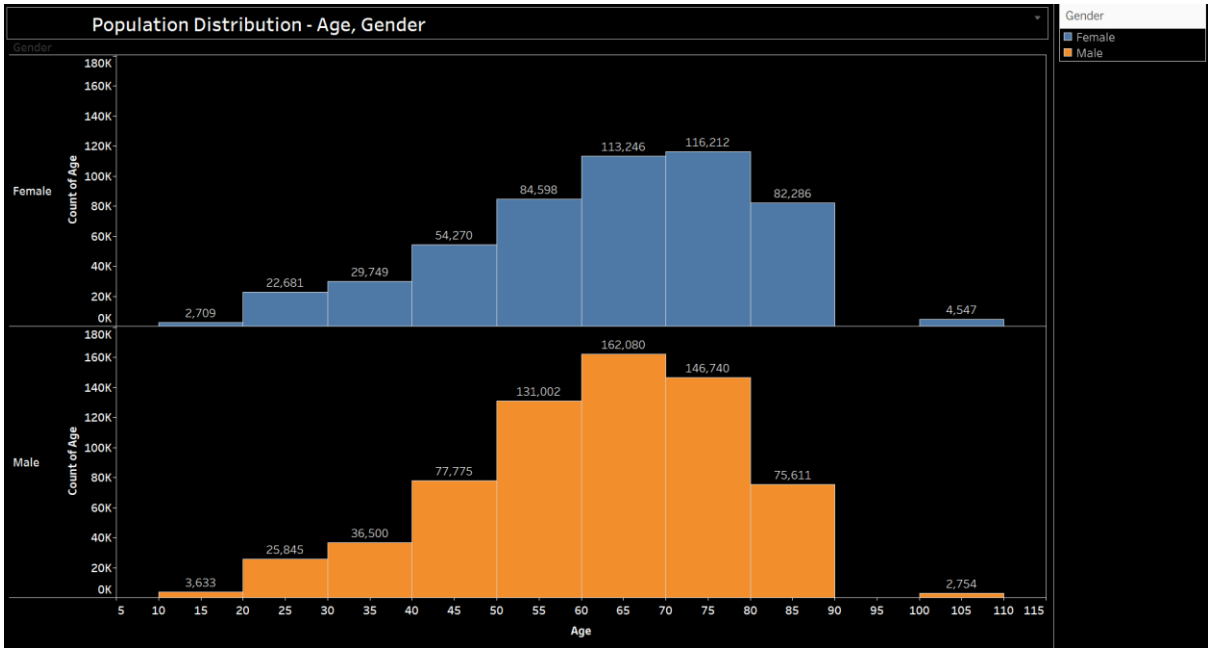
Results:

Population Distribution:



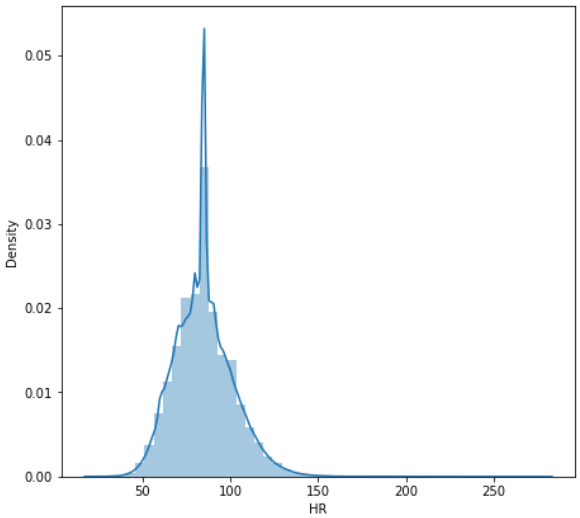
From the above visual, we can see the distribution of the Population (in percentage). We can observe that 56% of the patient population is Male while 44% of the patients are Females.

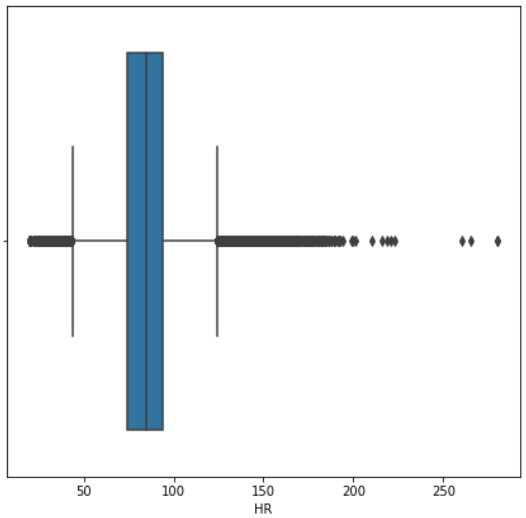
Population Distribution by Age & Gender:



The above Visualization shows the distribution by age and gender. We can see that male patient population is more compared to females and there are a greater number of patient population between the age range 60-70 years for and Males and 70-80 years old patients for females.

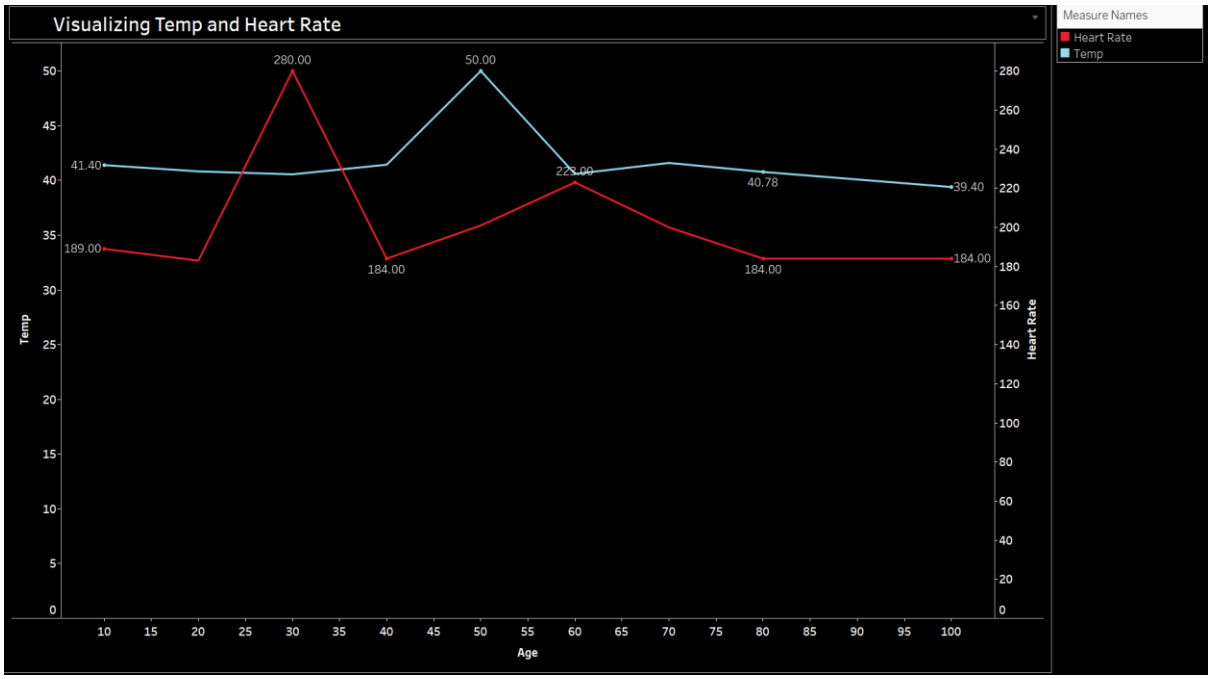
Heart Rate:

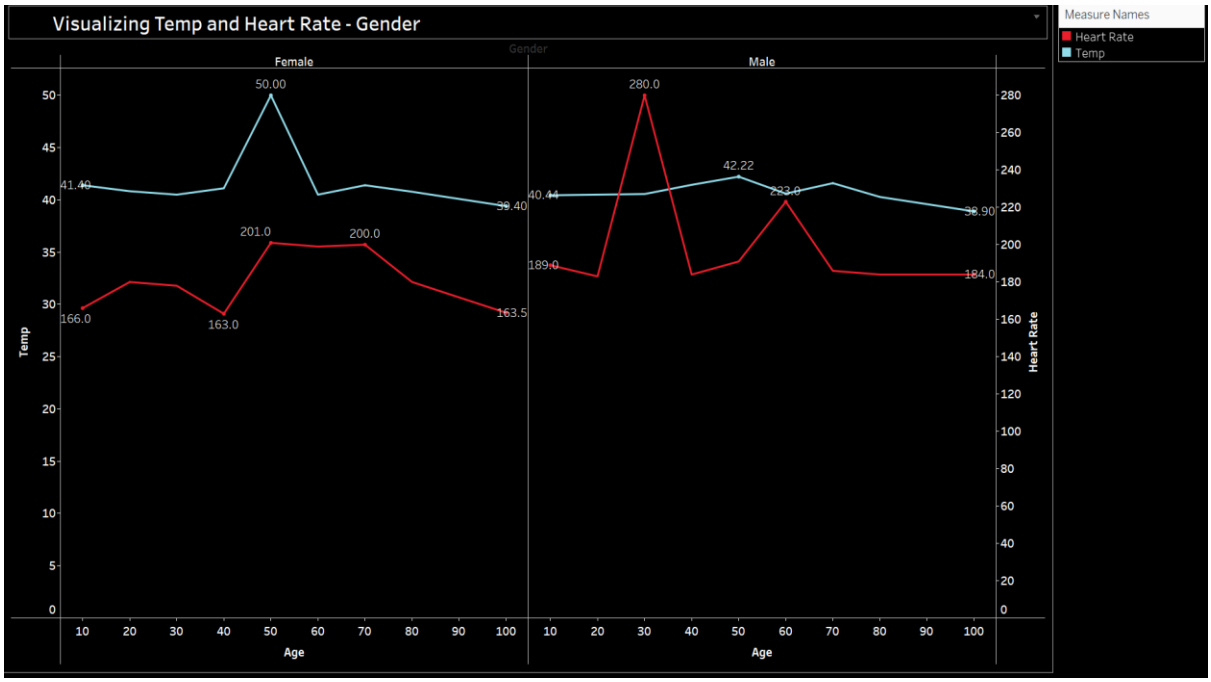




The above density plot shows the heart rate of the patients. There are some visible outliers from the box plot. The outliers indicate patients with high heart rate. This indicates those patients might be having a potential chance of developing sepsis.

Temperature and Heart Rate:

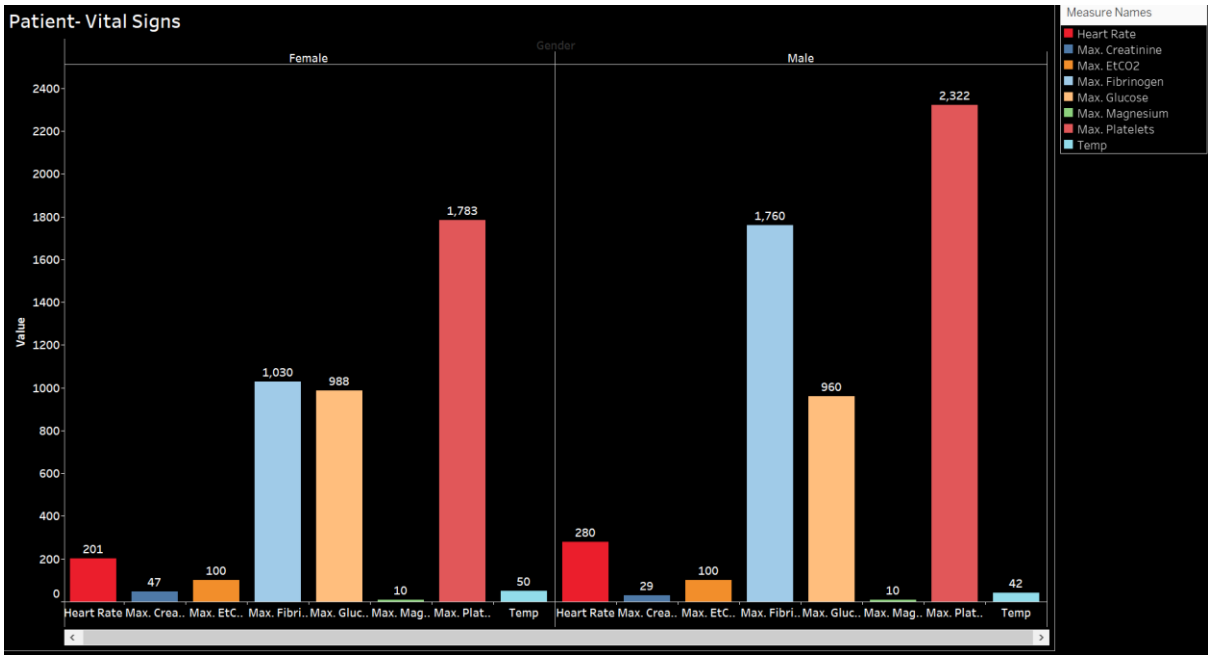




From the above visuals, we can see that the patient vital signs such as Heart rate and Temperature spike at the age 30 years and 50 years approximately.

Vital signs - Review

This visual provides us with the analysis of few more vital signs of the patients- Gender wise and we can notice that the while Ph levels appear stable at approximately a value of 7.5, Glucose levels, Platelets and O2Sat seem to be declining for the patients ageing above 70 years approximately.



This is a consolidated view of all the key vital signs parameters of the patients and we can notice that the blood platelets, Fibrinogen and Heart rate are High for Males while Females have High levels of Creatinine and Glucose levels.

Comparing the metrics of Various Machine Learning Models:

Machine Learning Model/s:

Split Ratio - 80:20

Machine Learning Models	Metrics				
ML Model Type	**Accuracy**	**Recall**	**Precision**	**F1 Score**	**Roc_auc**
DummyClassifier	0.981	0.5	0.019	0.495	0.5
GaussianNB	0.966	0.516	0.032	0.517	0.614
MultinomialNB	0.981	0.5	0.03	0.495	0.597
DecisionTreeClassifier-Max_depth : 5	0.696	0.595	0.033	0.438	0.619
DecisionTreeClassifier-min_impurity_decrease=0.01	0.831	0.545	0.022	0.474	0.545
Logistic Regression, C=0.1, max_iter=1000	0.648	0.585	0.033	0.418	0.615
Best Model :Decision Tree Classifier	0.696	0.595	0.033	0.438	0.619

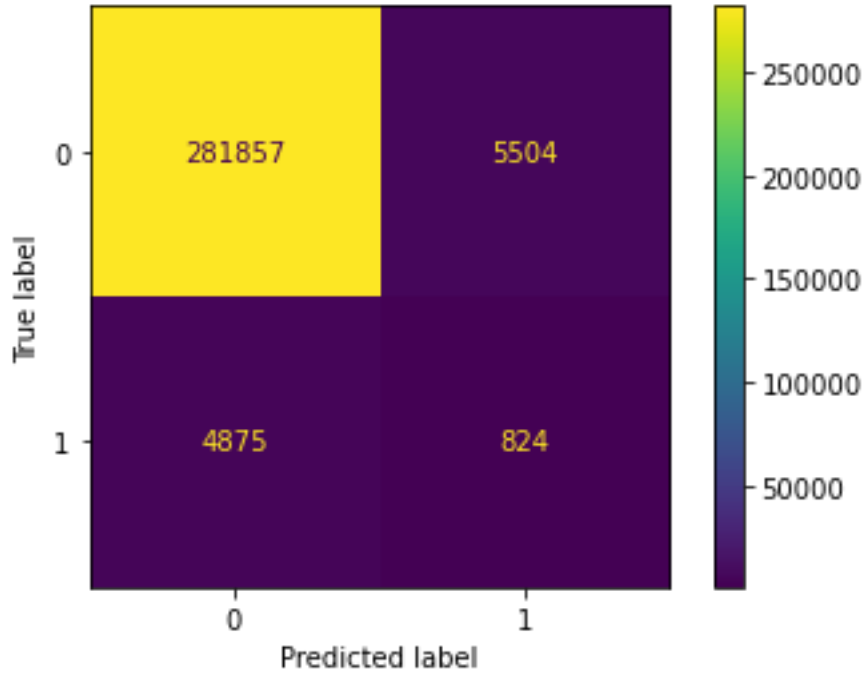
From the above table, we can see a consolidated view of the various machine learning models.

Data set split Ratio : We have split the dataset into 80:20(Train/Test) ratio depending on the Pareto's principle where it is believed that 20% of the events cause 80% of the issues.

Discussion:

After assessing the best ML model for our analysis, we have further develoeped and created a baseline model with an **Accuracy**: 0.964584044223026 ,**Precision_score**: 0.1302149178255373 , **Recall score**: 0.14458676960870329 , **F1 score**: 0.1370250270225326.

Evaluating the results with a confsion matrix:



Tuning the Model:

We tried to tune the base line model with tuning the parameter of the model with the clause : **criterion="entropy", max_depth=3** through which we have further optimised our model to a improved accuray of **98%.**

Prediction Results:

After tuning the model, we tried to test the model performance to see if it can predict the results appropriately.

The Machine Learning model accuratly predicts the sepsis outcome when we pass the patient's vital signs parameters like **'Resp'**, **'Glucose'**,**'Creatinine'**,**'Temp'**,**'HR'**,**'O2Sat'**,**'Age'**, **'EtCO2'**, **'Fibrinogen'**, '**O2Sat**'as Input to the ML Model,.

Conclusion:

With our project analysis, we can conclude that the machine learning model that we have developed has shown good predictive ability and it could accurately predict the sepsis outcome. However external validation studies with real time data would be required to validate the ML Model in a different set of population to enable it to be used for treatment practice.

Challenges:

- Finding the appropriate Dataset.
- Too many missing values from the feature variables.

Future Scope of Work:

Our project future scope of work would include performing prospective clinical trials to validate and use ML Model predictions in a real-time clinical setting.Further, develop an application interface using Flask /AWS Simplify integrating ML to predict sepsis outcome which can be used in Real-time clinical setting.

Contributions:

Name	Contributions
Rakeshwar Rao Baggu	Acquire Dataset, Building ML Models, Tune the Model, Project proposal, Preparing Final Report
Riddhi Deshpande	EDA, Creating Visualisations, Project MOM, Evaluate the various ML models.
Sarabhjeet Singh Virk	Developing Visualisations, Project Documentation, Project MOM

References:

1. Seymour CW, Liu VX, Iwashyna TJ, et al. Assessment of clinical criteria for sepsis: For the third international consensus definitions for sepsis and septic shock (sepsis-3). JAMA 2016; 315:762–774 - PMC - PubMed
2. Singer M, Deutschman CS, Seymour CW, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). JAMA 2016; 315:801–810 - PMC - PubMed
3. Shankar-Hari M, Phillips GS, Levy ML, et al. Developing a new definition and assessing new clinical criteria for septic shock: For the third international consensus definitions for sepsis and septic shock (sepsis-3). JAMA 2016; 315:775–787 - PMC - PubMed
4. Centers for Disease Control and Prevention: Sepsis. Cited 23 August 2016. https://www.cdc.gov/sepsis/datareports/index.html,. Accessed February 1 2019.
5. World Health Organization: Sepsis. Cited 19 April 2018s. Available at: https://www.who.int/news-room/fact-sheets/detail/sepsis. Accessed February 1 2019.
6. https://ieeexplore.ieee.org/abstract/document/9005808
7. Early Prediction of Sepsis in the ICU Using Machine Learning: A Systematic Review https://www.frontiersin.org/articles/10.3389/fmed.2021.607952/full
8. Predicting Sepsis based on ML Algorithms https://www.hindawi.com/journals/cin/2021/6522633/
9. DABL https://amueller.github.io/dabl/dev/

10. Insight – Machine Learning model to predict Sepsis https://www.dascena.com/insight
11. Prediction of sepsis patients using machine learning approach: A meta-analysis (https://www.sciencedirect.com/science/article/abs/pii/S016926071831602X#:~:text=The%20machine%20learning%20models%20showed,to%20onset%20was%200.89%20(Fig.)
12. Early Prediction of Sepsis in the ICU Using Machine Learning: A Systematic Review (https://www.frontiersin.org/articles/10.3389/fmed.2021.607952/full)
13. A Machine Learning Algorithm to Predict Severe Sepsis and Septic Shock: Development, Implementation, and Impact on Clinical Practice*(https://journals.lww.com/ccmjournal/Abstract/2019/11000/A_Machine_Learning_Algorithm_to_Predict_Severe.2.aspx)
14. Advances in sepsis diagnosis and management: a paradigm shift towards nanotechnology(https://jbiomedsci.biomedcentral.com/articles/10.1186/s12929-020-00702-6)
15. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU(https://bmjopen.bmj.com/content/8/1/e017833)
16. Predicting in-hospital mortality in ICU patients with sepsis using gradient boosting decision tree(https://pubmed.ncbi.nlm.nih.gov/34106618/)
17. use of decision trees for ICU outcome prediction in sepsis patients treated with statins (https://ieeexplore.ieee.org/document/5949439)
18. Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU(https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-01271-2)
19. Predicting Sepsis: A Comparison of Analytical Approaches (https://www.researchgate.net/publication/220896600_Predicting_Sepsis_A_Comparison_of_Analytical_Approaches)
20. A Case Study in Predicting the Likelihood of Sepsis (https://www.scitepress.org/Papers/2009/15541/15541.pdf)

Appendices

We have included the code below on how about we went on tuning the model as well as about the Prediction Results. From the below, we can notice that the Tuning the model gives us with a much better accuracy of 98%.

Tuning the model.

```
# Create Decision Tree classifer object
clf = DecisionTreeClassifier(criterion="entropy", max_depth=3)

# Train Decision Tree Classifer
clf = clf.fit(X_train,y_train)

#Predict the response for test dataset
y_pred = clf.predict(X_test)

# Model Accuracy, how often is the classifier correct?
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.9805534702791238
```

Prediction

Let's try to pass some feature variable and try to predict if a patient has a chance of develeoping Sepsis.

```
y_pred = classifier.predict(([[18.000000,136.121045,1.469606,37.300000,88.000000,100.000000,63.00,33.046486]]))

y_pred = y_pred.astype(int).tolist()


print("Prediction Complete : The outcome of the prediction is:" ,(y_pred))

if y_pred>[0]:
    print("Prediction Complete : This patient might possibly be a Sepsis patient")
else:
    print("This patient is not a Sepsis patient.")
```

```
Prediction Complete : The outcome of the prediction is: [0]
This patient is not a Sepsis patient.
```

From the above, we have passed the feature variables and we can see the model predicting outcome as 0, meaning not a Sepsis patient.

```
y_pred = classifier.predict(([[24.000000,136.121045,1.469606,36.995514,78.0,93.0,78.22,33.046486]]))

y_pred = y_pred.astype(int).tolist()

print("Prediction Complete : The outcome of the prediction is:" ,(y_pred))


if y_pred>[0]:
    print("This patient might possibly be a Sepsis patient.")
else:
    print("This patient is not a Sepsis patient.")
```

```
Prediction Complete : The outcome of the prediction is: [1]
This patient might possibly be a Sepsis patient.
```