**A Project Report On**

**PREDICTION OF SEPSIS USING ENSEMBLE TECHNIQUES**

***Mini project submitted in partial fulfillment of the requirements for the***

***award of the degree of***

**BACHELOR OF TECHNOLOGY**

**IN**

**INFORMATION TECHNOLOGY**

**(2021-2025)**

**BY**

**VUTUKURU V N S NIKITH 21241A1264**

**MOHAMMAD SAMEER 21241A1240**

**P.YASHWANTH REDDY 21241A1251**

***Under the Esteemed Guidance***

***of***

**DR. N.V. GANAPATHI RAJU**

**Professor**



**DEPARTMENT OF INFORMATION TECHNOLOGY**

**GOKARAJU RANGARAJU INSTITUTE OF ENGINEERING AND TECHNOLOGY**

**(AUTONOMOUS)**

**HYDERABAD**

**2023-24**



**CERTIFICATE**

This is to certify that it is a bonafide record of Mini Project work entitled **“PREDICTION OF SEPSIS USING MACHINE LEARNING”** done by **VUTUKURU V N S NIKITH (21241A1264), MOHAMMAD SAMEER (21241A1240), P.YASHWANTH REDDY (21241A1251)** of **B.Tech** in the Department of Information of Technology, **Gokaraju Rangaraju Institute of Engineering and Technology** during the period 2021-2025 in the partial fulfillment of the requirements for the award of degree of **BACHELOR OF TECHNOLOGY IN INFORMATION TECHNOLOGY** from GRIET, Hyderabad.

**Dr. N. V. Ganapathi Raju Dr. Y J Nagendra Kumar**

Professor Head of the Department

(Internal Guide)

**(Project External)**

**ACKNOWLEDGEMENT**

We take the immense pleasure in expressing gratitude to our Internal guide**, Dr. N.V.Ganapathi Raju, Professor, Dept of IT**, GRIET. We express our sincere thanks for his encouragement, suggestions and support, which provided the impetus and paved the way for the successful completion of the project work.

We wish to express our gratitude to **Dr. Y J Nagendra Kumar,** HOD IT,our Project Coordinators **Dr. R.V.S.S.S.Nagini** and **Mr. P.K. Abhilash** for their constant support during the project.

We express our sincere thanks to **Dr. Jandhyala N Murthy,** Director, GRIET**,** and **Dr. J. Praveen,** Principal, GRIET**,** for providing us the conductive environment for carrying through our academic schedules and project with ease.

We also take this opportunity to convey our sincere thanks to the teaching and non-teaching staff of GRIET College, Hyderabad.

**Email**: v.nikith@gmail.com **Email:** Sameer002md@gmail.com

**Contact No**: 8106584646 **Contact No**: 8096120786



**Email**: yashwanthreddy1102003@gmail.com

**Contact No.** 9494399065

**DECLARATION**

This is to certify that the mini-project entitled “**PREDICTION OF SEPSIS USING MACHINE LEARNING”** is a bonafide work done by us in partial fulfillment of the requirements for the award of the degree **BACHELOR OF TECHNOLOGY IN INFORMATION TECHNOLOGY** from Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad.

We also declare that this project is a result of our own effort and has not been copied or imitated from any source. Citations from any websites, books and paper publications are mentioned in the Bibliography.

This work was not submitted earlier at any other University or Institute for the award of any degree.

**VUTUKURU V N S NIKITH 21241A1264**

**MOHAMMAD SAMEER 21241A1240**

**P.YASHWANTH REDDY 21241A1251**

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
|  | **Name** | **Page no** |
|  | **Certificates** | Ii |
|  | **Contents** | V |
|  | **Abstract** | vii |
| **1** | **INTRODUCTION** | 1 |
| 1.1 | Introduction to project | 1 |
| 1.2 | Existing System | 4 |
| 1.3 | Proposed System | 5 |
| **2** | **REQUIREMENT ENGINEERING** | 6 |
| 2.1 | Hardware Requirements | 6 |
| 2.2 | Software Requirements | 6 |
| **3** | **LITERATURE SURVEY** | 7 |
| **4** | **TECHNOLOGY** | 9 |
| **5** | **DESIGN REQUIREMENT ENGINEERING** | 12 |
| 5.1 | UML Diagrams | 12 |
| 5.2 | Use-Case Diagram | 13 |
| 5.3 | Class Diagram | 14 |
| 5.4 | Activity Diagram | 15 |
| 5.5 | Sequence Diagram | 16 |
| 5.6 | System Architecture | 17 |
| **6** | **IMPLEMENTATION** | 18 |
| **7** | **SOFTWARE TESTING** | 30 |
| 7.1 | Unit Testing | 30 |
| 7.2 | Integration Testing | 30 |
| 7.3 | Performance Testing | 30 |
| 7.4 | Model Testing | 31 |
| 7.5 | Validation Testing | 31 |
| **8** | **RESULTS** | 32 |
| **9** | **CONCLUSION AND FUTURE ENHANCEMENTS** | 36 |
| **10** | **BIBLIOGRAPHY** | 38 |

**11 LIST OF DIAGRAMS**

|  |  |  |  |
| --- | --- | --- | --- |
| **S No** | **Figure Name** | | **Page no** |
| **1** | Use Case Diagram | | 13 |
| **2** | Class Diagram | | 14 |
| **3** | Activity Diagram | | 15 |
| 4 | Architecture | | 16 |
|  |  | |  |
|  | |
|  | |
|  | |

**ABSTRACT**

Sepsis is a dangerous and potentially fatal illness that develops when an infection spreads throughout the body and sets off a broad inflammatory reaction. Its mortality rate is extremely high, particularly for intensive care unit patients. Sepsis must be identified and treated as soon as possible. Numerous data sources, including electronic health records (EHRs), vital signs, laboratory test results, and demographic data, can be used to train machine learning models for sepsis detection. The analysis of exploratory data was done with various preprocessing methods. The Ensemble techniques are used to categorize the disease, and comparisons with other classifiers, such as bagging classifier and random forest classifier, are also carried out.

The early prediction of sepsis is potentially life-saving, and we challenge participants to predict sepsis 6 hours before the clinical prediction of sepsis. Conversely, the late prediction of sepsis is potentially life-threatening, and predicting sepsis in non-sepsis patients (or predicting sepsis very early in sepsis patients) consumes limited hospital resources. For the challenge, we designed a utility function that rewards early predictions and penalizes late predictions as well as false alarms.More than 60,000 ICU patients were found, and for every hour a patient spent in the ICU, they could extract up to 40 clinical variables. For the onset of sepsis, they used the Sepsis-3 clinical criteria.

Sepsis is a potentially harmful condition that results in tissue damage, organ failure, or even death of the client. The main goal is to identify sepsis as soon as a patient arrives at the emergency room for care.

**Keywords**: generalizability, sequential prediction tasks, open-source-algorithm.

**Domain**: Machine learning

1. **INTRODUCTION**

**1.1 Introduction to Project**

Consequences of an unbalanced body response with regard to such toxins, sepsis may include a series of other organ systems. Sepsis is a disease that every person may get because of infections. People with chronic diseases like cancer and diabetes, renal, lung and kidney diseases, and pregnant women mostly likely acquire it due to their weak immune systems. This disease also affects newborns who are less than a year old. This disease is a severe threat to public health caused by its high fatality and morbidity. The outcome can be improved with early detection and antibiotic therapy. Community acquired causes of sepsis are pneumonia, infections of the stomach, kidney infections, and infections of the blood. Sepsis presents with a high or low fever, particularly in hypothermia (very low body temperature), hyperventilation (quick breathing), and extreme pain or discomfort. That solution involves intravenous antibiotic medication to help fight the infection, along with vasoactive to elevate the blood pressure in people who have low blood pressure, and intravenous antibiotics to resist infection form the measures to combat sepsis.

Sepsis is a life-threatening condition wherein the body's response to an infection results in tissue damage, organ failure, or even death. In the U.S. alone, about 1.7 million people develop sepsis, and 270,000 die from it annually; one-third of all people who die in U.S. hospitals have sepsis. Globally, it is estimated that 30 million people develop sepsis and 6 million die from this condition annually. In the United States alone, it costs American hospitals are higher than for any other health condition at $24 billion annually, 13% of U.S. health care costs, and the vast majority is due to patients that develop sepsis in the course of an admitted or emergency department visit. The developing world bears additional costs of management and suffers increased risks from poor outcomes. Overall, sepsis represents an important public health condition accounting for important morbidity, mortality and health care costs.

Treatment delays can result from the syndromic nature of sepsis, which makes it difficult to reliably and early identify. Two recent studies emphasize the significance of early detection and treatment of sepsis, suggesting an increase in the adjusted death rate of sepsis patients who had their antibiotic treatment delayed. This effect is even more pronounced in septic shock patients, where hourly delays were linked to an hourly mortality increase ranging from 3.6 to 9.9%. The clinical criteria for diagnosing and treating sepsis have been proposed by professional critical care societies, but the basic requirement for accurate and timely sepsis identification is still unmet.

First, solutions to the early detection of sepsis often address slightly diﬀerent problems, and they are usually developed and validated in diﬀerent patient populations with diﬀerent clinical variables and labels resulting from the application of diﬀerent clinical deﬁnitions of sepsis. For the Challenge, we oﬀered a common problem statement in terms of the same clinical variables and sepsis criteria and shared data from two diﬀerent hospital systems, sequestering from a third hospital system. Algorithms that overﬁt on the shared databases tended to perform poorly on the hidden database, especially if they learned data collection behaviours speciﬁc to a given hospital system. Also, we ran algorithms only once on the entire hidden dataset to prevent sequential training on the hidden data, and they compared algorithms to identify teams attempting to game the rules and get more "bites of cherry" than other teams.

Great there Different studies often employ different metrics and such metrics do need not necessarily reflect the clinical utility of sepsis detection and treatment. Traditional such scoring metrics make no explicit reward for early detect or penalize false alarms or overtreatment. For the Challenge, they developed a new measure, which overcomes these problems and which may have general applicability for the prediction of rare events in time series data.

Finally, the complexity of such algorithms is all but impossible to sufficient describe in a research article. For the Challenge, they marketed and facilitated the open sourcing of Algorithms to guarantee that all the minute implementation details are shared and can be reproduced.

**Random Forest Classifier**

Random Forest is a very strong and versatile algorithm in machine learning, applied in predicting the onset of sepsis in a patient based on their clinical data. Random Forest does this by creating many decision trees during training and returning an average of the classes, in the case of classification, or the mean prediction, in the case of regression, of individual trees. This ensemble method ensures improvement in predictive accuracy, controls overfitting, and has been relatively suitable for highly complex medical datasets where the interactions among several clinical features are intricate and nonlinear.

By leveraging Random Forest, the model can effectively handle large datasets with many features, identify the most important predictors of sepsis, and deliver reliable predictions even when some data points are missing or noisy. This makes it an ideal choice for developing a predictive tool that can assist healthcare professionals in early detection and timely intervention for sepsis, ultimately aiming to improve patient outcomes and reduce mortality rates.



*[Fig-1] Random Forest Algorithm*

**1.2 Existing System**

Sepsis is an organ failure that can be fatal and is caused by an improperly managed host response to infection. To lower the high rates of sepsis and septic shock morbidity and death, it is imperative to enhance the early detection of suspected sepsis in prehospital and emergency situations. The current methodology makes use of ANN models such as Radial Basis Function Neural Networks (RBFNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), etc. They are a type of recurrent neural network that is capable of identifying order dependence in sequence prediction scenarios.Because the model cannot handle the long sequences, training it is very difficult in this case. They are built with the capacity to recognize the sequential characteristics of data and then apply patterns to predict what will happen in the future. Accuracy is impacted by the model's slow computational capability.

**Limitations:**

 **Handling Long Sequences**:

* Struggle with vanishing and exploding gradients in long sequences.

 **Slow Computational Speed**:

* Lengthy training and inference times impede timely predictions.

 **Gradient Issues**:

* Suffer from vanishing and exploding gradients, leading to unstable training.

 **Limited Accuracy**:

* Achieve only 82-87% accuracy, which may be insufficient for clinical use.

 **Training Complexity**:

* Require significant resources and time for training and optimization.

**1.3 Proposed System**

The primary objective of this work is to design and develop a Random Forest Classifier-based method for early detection of sepsis. The three primary steps in the recommended methodology are pre-processing, feature value determination, and data classification. The pre-processing step of resampling the data will come first. Yeo Johnson and log plots are used to calculate the feature importance. Results for the sepsis detection are provided by the recommended classifier, which goes by the name of Random Forest Classifier. The block diagram of the modules in the proposed system, which illustrates our suggested methodology. The recommended approach will be implemented using Python. To illustrate the effectiveness of the method, the system is evaluated in terms of Accuracy and Log Loss.This study will compare the measured performance of the proposed method with that of the earlier research.

The resampling step in pre-processing helps to address imbalanced data, ensuring that the model learns equally from both sepsis and non-sepsis cases, thereby improving its detection capabilities.

The method is implemented in Python, which has a rich ecosystem of libraries for data analysis and machine learning, making it accessible and straightforward to implement.

Random Forest can handle large datasets and a high number of features efficiently, making the approach scalable and versatile for various clinical settings and diverse datasets.

**2. REQUIREMENT ENGINEERING**

**2.1 Hardware Requirements**

* Processor – i5 and above (64-bit OS).
* Memory – 4GB RAM (Higher specs are recommended for high performance)

* Input devices – Keyboard, Mouse

**2.2 Software Requirements**

* Anaconda Navigator/Jupiter Notebook
* Python
* Python Libraries

1. pandas

2.numpy

3.matplotlib

4.scikit-learn

5.CSV

6.scipy

1. **LITERATURE SURVEY**

For the purpose of this study, vital sign data were continuously collected in an automatic fashion. It is less time consuming and prone to data entry mistake compared to research where data on vital signs were obtained manually. A Naive Bayes algorithm in a maximum a posteriori framework was used for sepsis prediction up to 24 hrs before clinical suspicion of sepsis. However, this procedure confines the number of positives, but also alters the distribution of cases in the various sub-groups due to its therapeutic implications.[1]

Deep neural networks to learn representations for early sepsis detection. The aim of this study was to compare the performance and feasibility of the new deep learning methodology with the regression approach based on conventional temporal feature extraction. The accuracy, performance, and feature extraction capacity of DNNs are enhanced through comparison with hand-crafted. Possessing the desired variable defined, adequate data available, and adequate explanatory capacity.[2]

A computational framework for the diagnosis of sepsis. In Performance evaluation, the Insight algorithm bends sensitivity and specificity three hours before a prolonged SIRS event for sepsis prediction. This is done by analysis of the correlation of nine conventionally used vital sign parameters. Sensitivity of 90%, Specificity of 81%, and fast prediction are the advantages of this model. The sensitive criteria are for SIRS in sepsis, but another limitation is the high percentage of false positive results associated with the SIRS criteria.[3]

A Novel Approach to Estimating Sepsis Using Deep Learning . Seven values from each of the six data quantification channels were selected out of the 34 features in the machine learning implementation. Decision-tree based SVM classifiers are used to create physiochemical prediction models. Using the ACNN classifier along with two additional intelligent classifiers, the proposed plan for the validation procedure of LSTM-RNN and SVM classifiers is validated. Its extreme precision demonstrates that the model can predict shock, severe sepsis, and their aftermath with high accuracy. Insensitive; unsuitable for databases holding a wide range of data.[4]

The application and effects of a machine learning algorithm for predicting sepsis and septic shock on clinical practice . To develop a machine learning algorithm, apply it, and evaluate its impact on patient outcomes and clinical practice in order to predict severe sepsis and septic shock. Benefit: The accuracy has increased. One drawback is the need for a significant volume of superbly annotated medical data.[5]

Using machine learning algorithms to forecast sepsis patients' in-hospital mortality in the intensive care unit . They employed the traditional logistic regression (LR) technique, random forest, gradient boosting machine, and the least absolute shrinkage and selection operator (LASSO) to build prediction models. Improved resource management is beneficial. Cons include limited availability in low-resource environments and ethical concerns about data privacy.[6]

Review of a machine learning algorithm for using six vital signs to predict sepsis in advance . This paper presents the validation and performance comparison of a gradient-boosted ensemble machine learning tool for sepsis detection and prediction. Compared to traditional methods, faster and more accurate detection enables continuous patient monitoring. Large volumes of annotated data are needed to mitigate the risk of bias in the models and the data.[7]

A Systematic Review for the Prediction of Sepsis in the ICU .Three processes were employed in the feature selection process: examining existing sepsis screening models, speaking with regional subject matter experts, and utilizing the gradient boosting supervised machine learning algorithm. Alert rate, receiver operating characteristic curve area under, sensitivity, specificity, and precision were among the important metrics. The model can precisely predict the onset of shock, severe sepsis, and sequelae due to its high specificity. Insensitive; not suitable for all databases in the event of a mixed data set.[8]

Sepsis is an infection-related illness that causes organ failure and has a very high death rate. With the goal of predicting early sepsis six hours in advance, two processing methods—the mean processing method and the feature generation method—are constructed using the machine learning algorithms XGBoost and LightGBM. Various features, such as statistical strength features, window features, and medical features, are combined to create the feature generation methods.Both the XGBoost and LightGBM algorithms have good prediction performance, but LightGBM performs better overall, especially on multidimensional data, and has a faster running speed (AUC of 0.979 in the feature generation method). WBC, platelets, and PTT are the main risk factors for early sepsis prediction.[9]

Predicting Sepsis in ICU Patients Using Machine Learning Algorithms: A Comparative Study,The

models such as support vector machines, random forests, and neural networks have been compared with each other in relation to accuracy and utility in clinical practice. What was underlined is that ensemble methods and deep learning approaches have better performance in predicting sepsis, reflecting their potential to improve early detection and thus enhance patient outcome in intensive care units.[10]

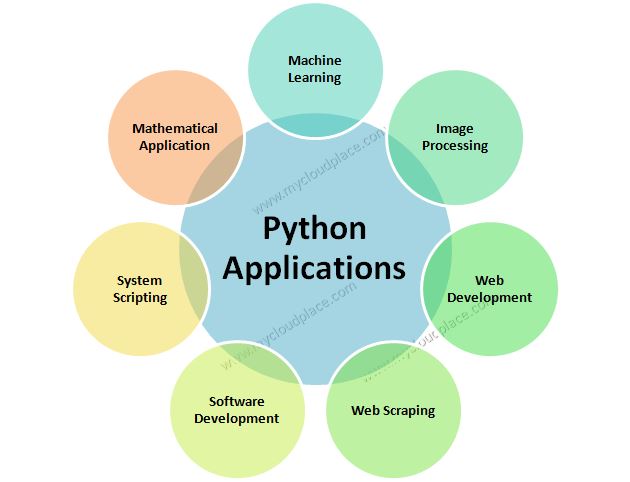
**4. TECHNOLOGY**

**4.1 ABOUT PYTHON**

Python’s environment has evolved over time, and it is becoming increasingly capable of statistical analysis. It strikes a good balance between scale and class. Python plays a premium on efficiency and readability. It has a design that emphasizes program readability and a simple syntax that is beginner-friendly and also lets programmers express the codes in fewer lines notably using with indentation. The speciality of this excessive degree language are functions of dynamic system, automatic-memory management system.

**4.2 APPLICATIONS OF PYTHON**

Python is used in many application domains. It makes its presence in every emerging field. It is the fastest-growing programming language and may be used to create any type of application.



*[Fig – 2] Applications of python*

It is used in various fields:

* Web Applications. We can use Python to develop web applications. ...
* Desktop GUI Applications. ...
* Console-based Application. ...
* Software Development. ...
* Scientific and Numeric. ...
* Business Applications. ...
* Audio or Video-based Applications. ...
* 3D CAD Applications.

**4.3 PYTHON IS WIDELY USED IN DATA SCIENCE**

We use python data science in flexible and open source language.It gives functionality which deal with mathematics and scientific function.Most probably we use because of simple syntax and it gives huge libraries.It also consumes less time to code.

The major python libraries used in Data Science are as follows

**4.3.1 PANDAS**

A library in python that's used for statistics analyzing, cleaning, exploring and manipulating. generally, dataset incorporates many beneficial and useless statistics. Pandas cause them to readable and relevant.

**4.3.2 NUMPY**

A library in python that's used for statistics reading, cleaning, exploring and manipulating. commonly dataset carries many useful and vain data. Pandas make them readable and relevant.

**4.3.3 MATPLOTLIB**

A library is used for plotting graphs in python. It built on NumPy arrays. We can plot any graph from basic plot types to bar graph, histogram, scatter and many more.

**4.3.4 SCIKIT-LEARN**

In Python, a library for device gaining knowledge of. device gaining knowledge of and statistical modelling, along with class, regression, and clustering, are completed with Scikit-learn gear.

**4.3.5 CSV**

it is a kind of file that stores tabular records, like a spreadsheet or a database. There are one or extra fields in each entry, that are separated by commas. We use the csv built in module to work with CSV files.

**4.4 Dataset Description**

A dataset has been taken from the Kaggle in order to train our model. It consists of 60000 patients information. These patients have medical information of different tests. These dataset is extracted from kaggle website.

We challenge participants to predict sepsis six hours before the clinical prediction of sepsis because early detection of sepsis may save lives. On the other hand, late detection of sepsis may be fatal, and early detection of sepsis in patients who are already in sepsis or in non-sepsis patients uses up scarce hospital resources. We created a utility function for the task that penalizes late predictions and false alarms while rewarding early predictions.

Input size: 153.82MB

Why kaggle dataset…?

1. The dataset can be trained easily as it is small in size.
2. Data is properly labelled.
3. The dataset is available for free.

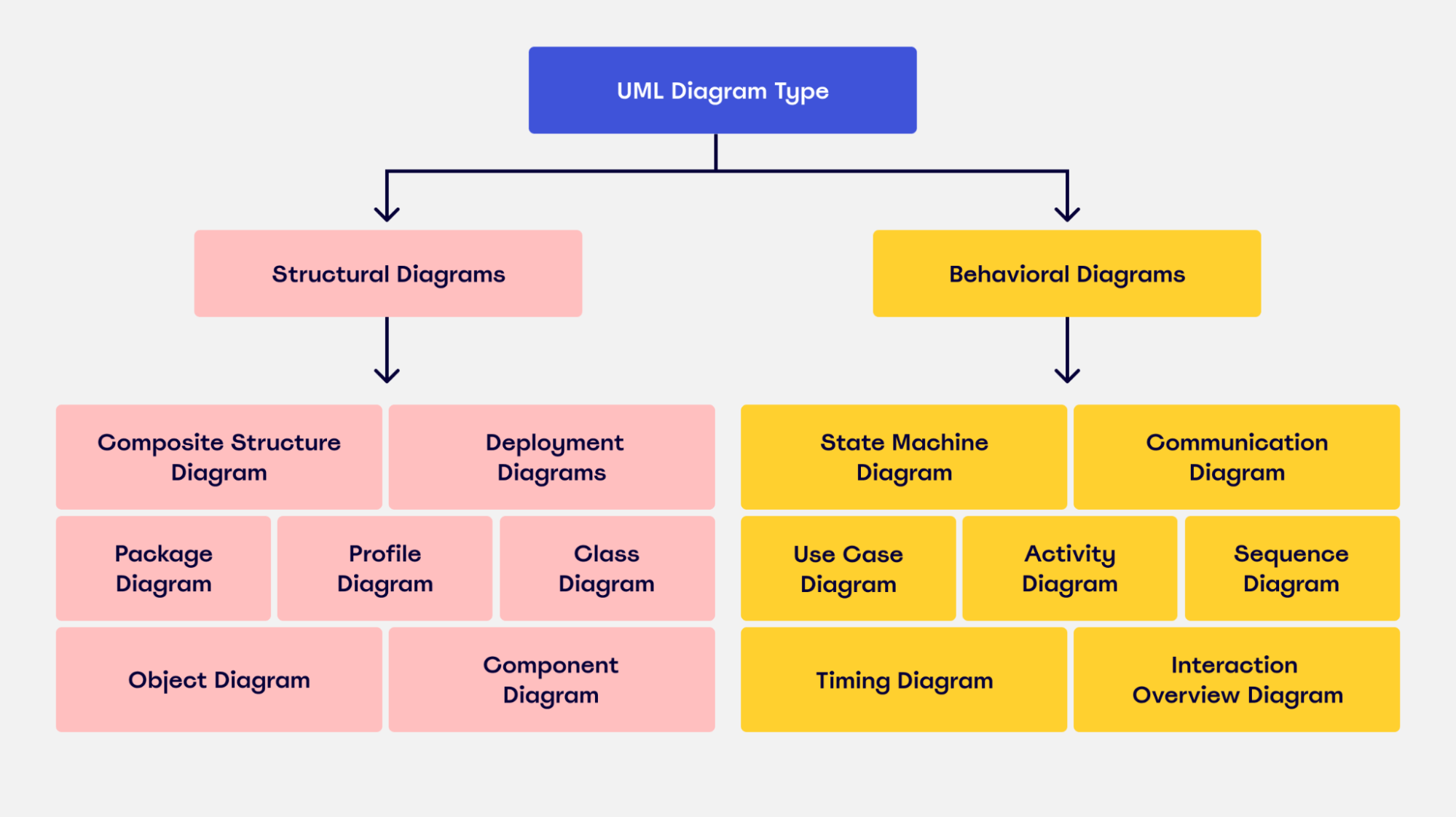
**5.DESIGN REQUIREMENT ENGINEERING**

**Concept of uml:**

The aim of those diagrams, which are based on UML, is to visually represent the machine as well as its primary players, roles, moves, objects, or training, with the intention of better understanding, manipulating, preserving, or filing statistics about the machine.

**UML DIAGRAMS:**

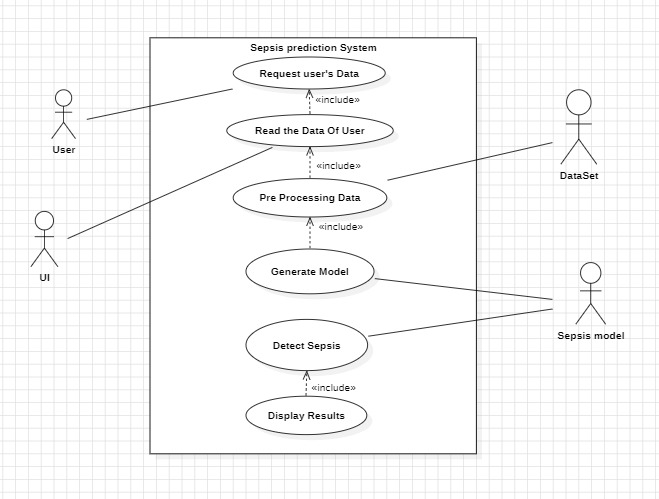
The Unified Modelling Language (UML) is a language used to create models for various purposes. Its main objective is to provide a standard way of representing the structure of a system visually, much like blueprints used in other fields of engineering. In the case of complex applications, the involvement of multiple teams necessitates clear and straightforward communication between them. Businesspeople may not understand code, and this is where UML comes in. It helps to communicate the essential requirements, features, and procedures of the system to non-programmers. By visualizing processes, user interactions, and the static structure of the system, teams can save time in the long run.

****

*[Fig-3] Concepts of uml*

**5.1 Use case Diagram:**

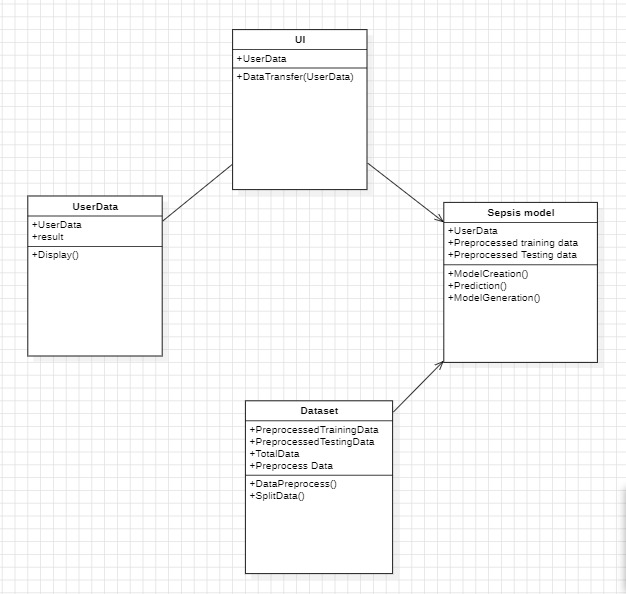
A use case diagram is a type of behavioral diagram that is a graphical explanation of the functionalities offered by the system in relation to the participants, their goals, and any dependencies between these cases.



*[Fig - 4] Use Case Diagram*

**5.2 Class Diagram:**

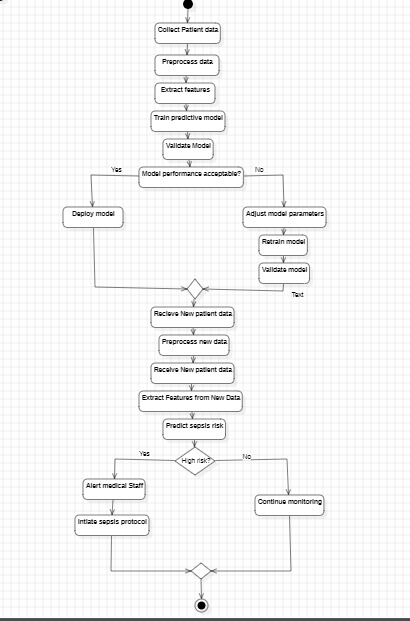
A class diagram is a static type of structural diagram that still depicts the format of a machine by means of illustrating the hyperlinks among the machine's lessons, attributes, operations, and instructions.



*[Fig – 5] Class Diagram*

**5.3 Activity diagram:**

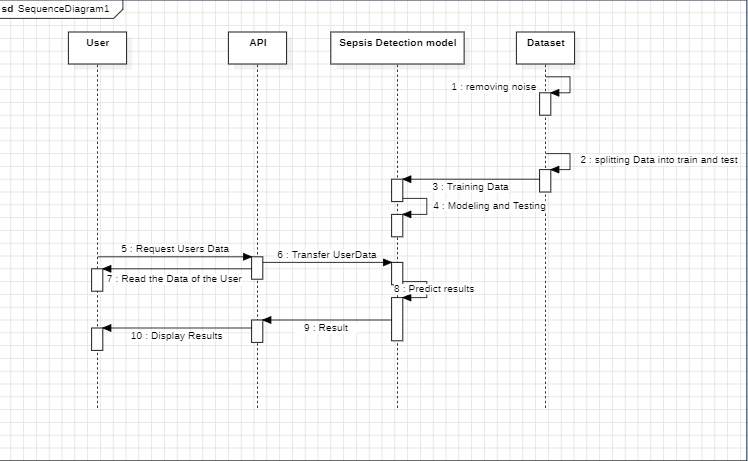
This diagram is a more complex version of a flow chart that depicts the flow of information from one activity to the next. It describes the coordination of activities in order to offer a service at various levels of abstraction.

****

*[Fig – 6] Activity Diagram*

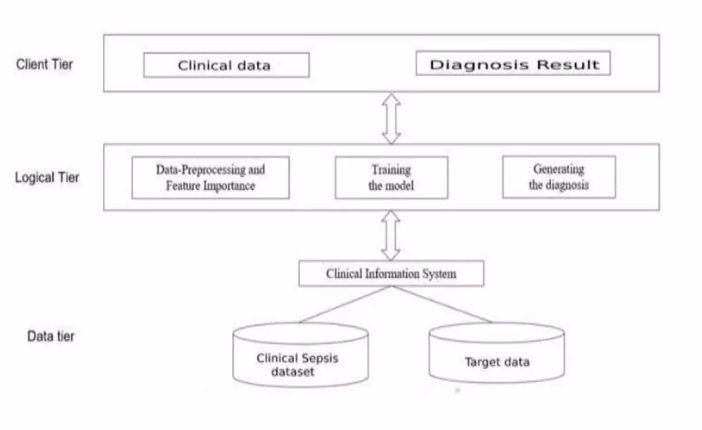
**5.4 Sequence Diagram:**

In the Unified Modeling Language (UML), a sequence diagram shows the order in which objects interact in a particular use case scenario. It displays the relationships between objects and their order of operation. The main purpose of sequence diagrams is to plan and record the system's logical flow. They draw attention to the ways in which different entities interact over time.



*[Fig – 7] Sequence Diagram*

**5.5 Architecture**



*[Fig – 8] System Architecture*

The model is structured into three tiers: client, logical, and data. In the client tier, clinical data is inputted and diagnosis results are outputted. The logical tier handles data preprocessing, feature importance identification, model training, and diagnosis generation. The data tier consists of a clinical sepsis dataset for model training and target data for making predictions.

Machine learning-based sepsis prediction systems analyze patient data to provide early warnings of potential deterioration, enabling timely interventions that can save lives. These systems leverage advanced algorithms, continuous data input, and real-time monitoring to enhance clinical decision-making.

**5.5.1 Algorithm**

Step 1 : Collect data from various sources such as electronic health records (EHRs), vital signs,

laboratory test results, and demographic information.

Step 2 : Clean the data by handling missing values, removing duplicates, and normalizing the data.

Resample the data to address any imbalances between sepsis and non-sepsis cases.

Step 3 : Identify and extract relevant features that could help in predicting sepsis, such as heart rate, temperature, blood pressure, etc.

Apply feature engineering techniques like Yeo-Johnson transformation to improve the model's performance.

Step 4 : Split the dataset into training and testing sets to evaluate the performance of the model.

Step 5 : Train an ensemble machine learning model such as a Random Forest Classifier using the training dataset.

Optimize the model's parameters using cross-validation techniques to improve its accuracy.

Step 6 : Evaluate the trained model on the testing dataset using metrics like accuracy, precision, recall, and F1-score.

Compare the performance of the Random Forest Classifier with other classifiers like bagging classifier and other ensemble methods.

Step 7 : Perform unit testing, integration testing, performance testing, and validation testing to ensure the model's reliability and robustness.

Step 8 : Deploy the trained and validated model in a clinical setting to predict sepsis in real-time and aid healthcare professionals in early diagnosis and intervention.

**6. IMPLEMENTATION**

**Sample Data for the Problem Statement:**

This project shows how to detect this illness earlier and more accurately using a range of classifiers, with an emphasis on stacking classifiers to generate reliable prediction models and lessen the requirement for laborious lab testing. We intend to enhance the application in the future by customizing it for each customer, integrating this model into a hospital website, and helping the medical staff recognize any early symptoms of illness.

**Model Construction**

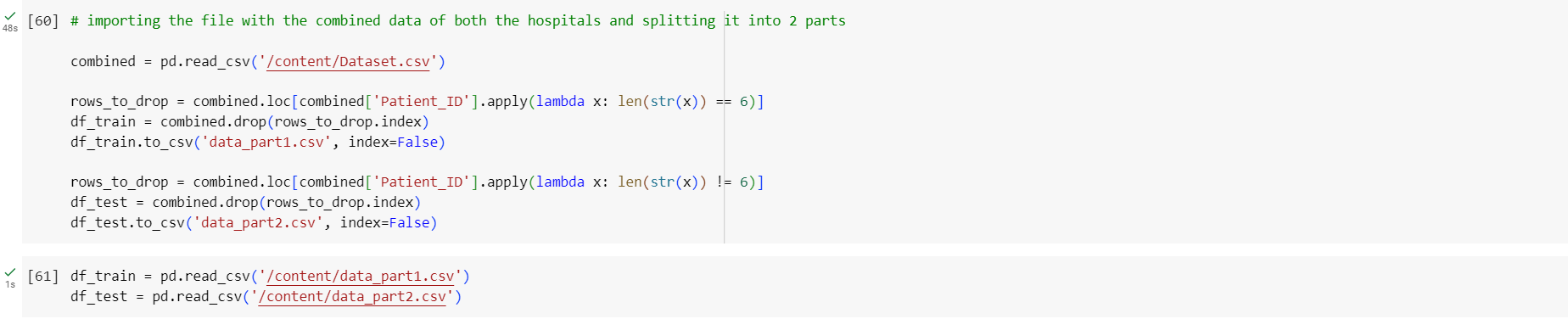
**Importing Modules** **and downloaded Packages**

To utilize a module's features, you need to first import the module with an import declaration. The import key-word is accompanied by the module's call in an import declaration. In a Python file, this will be declared on the pinnacle of the code, beneath any shebang strains or standard remarks.

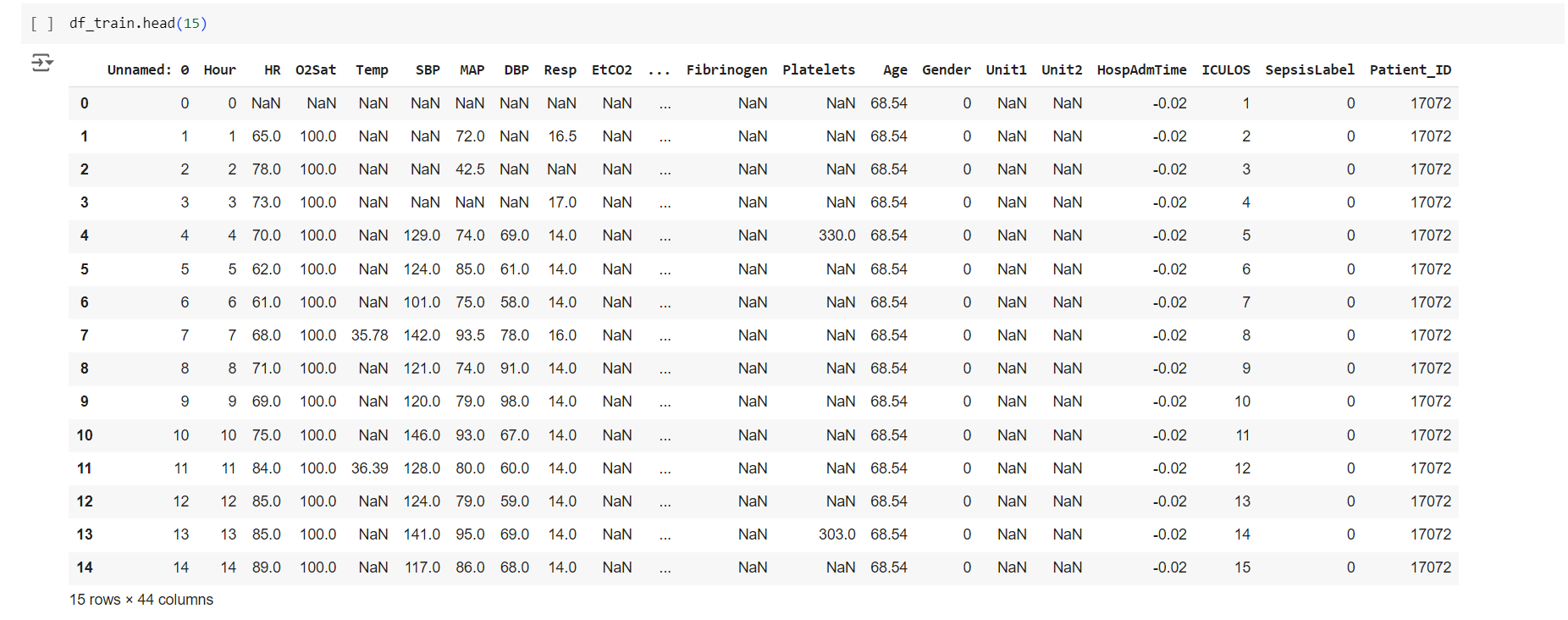


*[Fig – 9] Importing Packages*

**Importing and splitting the data into training and testing**

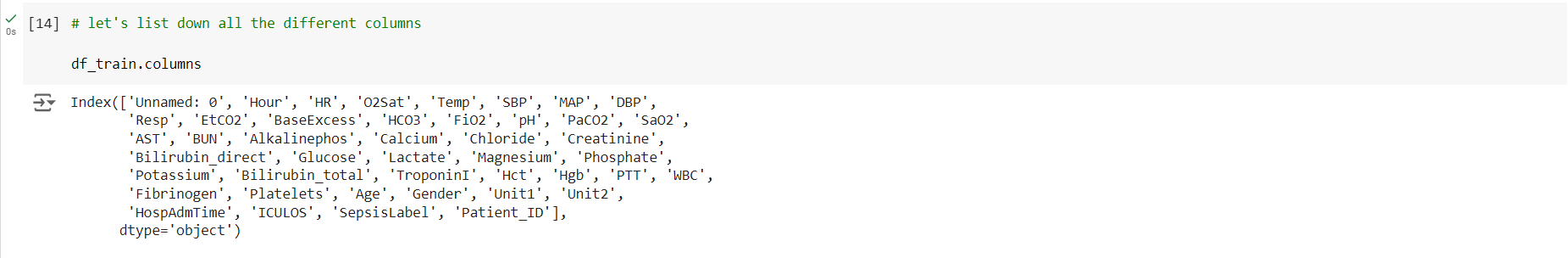
****

*[Fig -10] Importing and splitting the data into training and testing*



*[Fig-11] Display data*

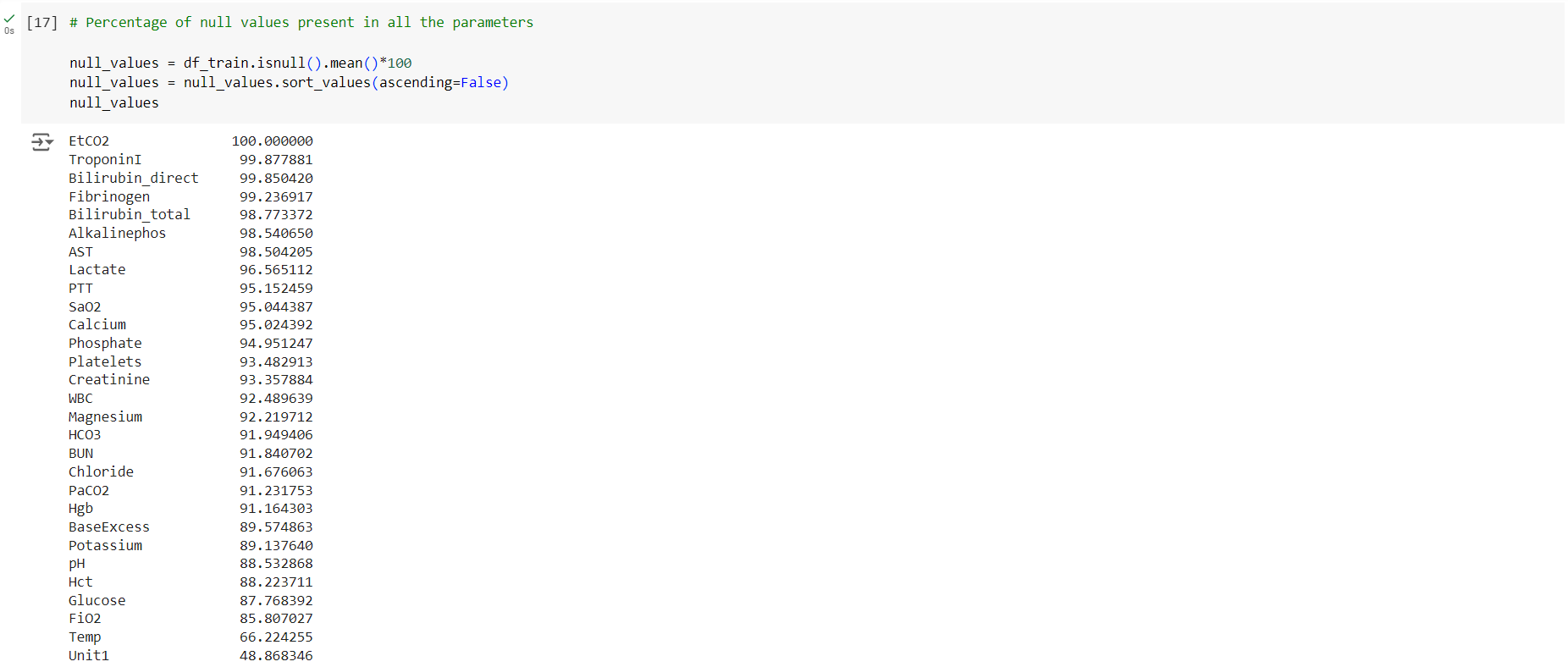
**Exploratory data analysis**

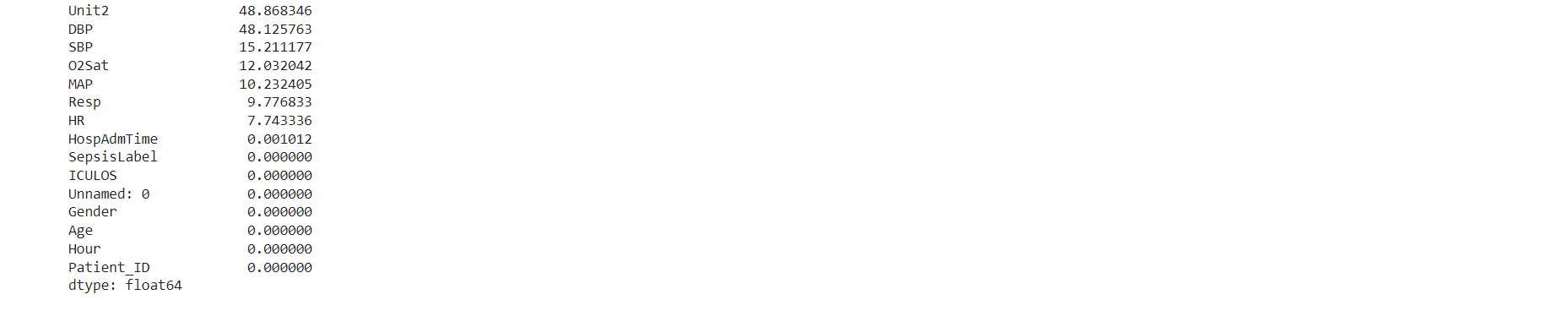
****

*[Fig-12] listing down all the different columns*

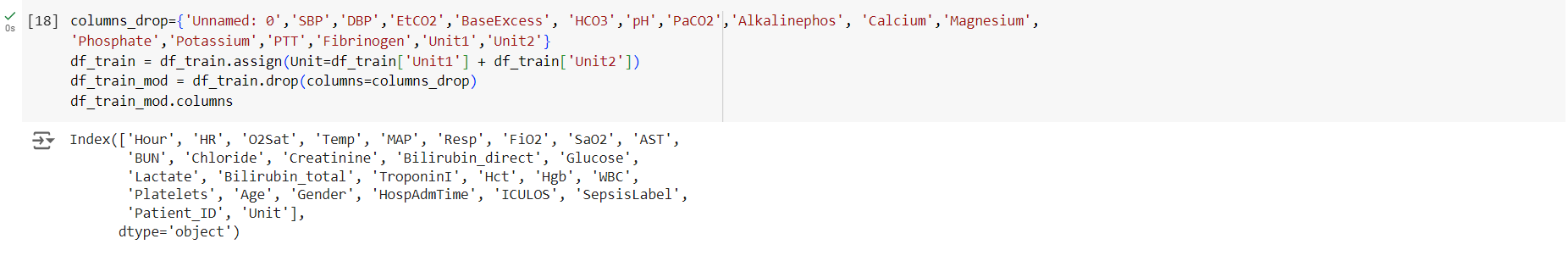
****

*[Fig-13] Function to draw correlation heat map*

****

****

*[Fig-14]Percentage of null values present in all the parameters*

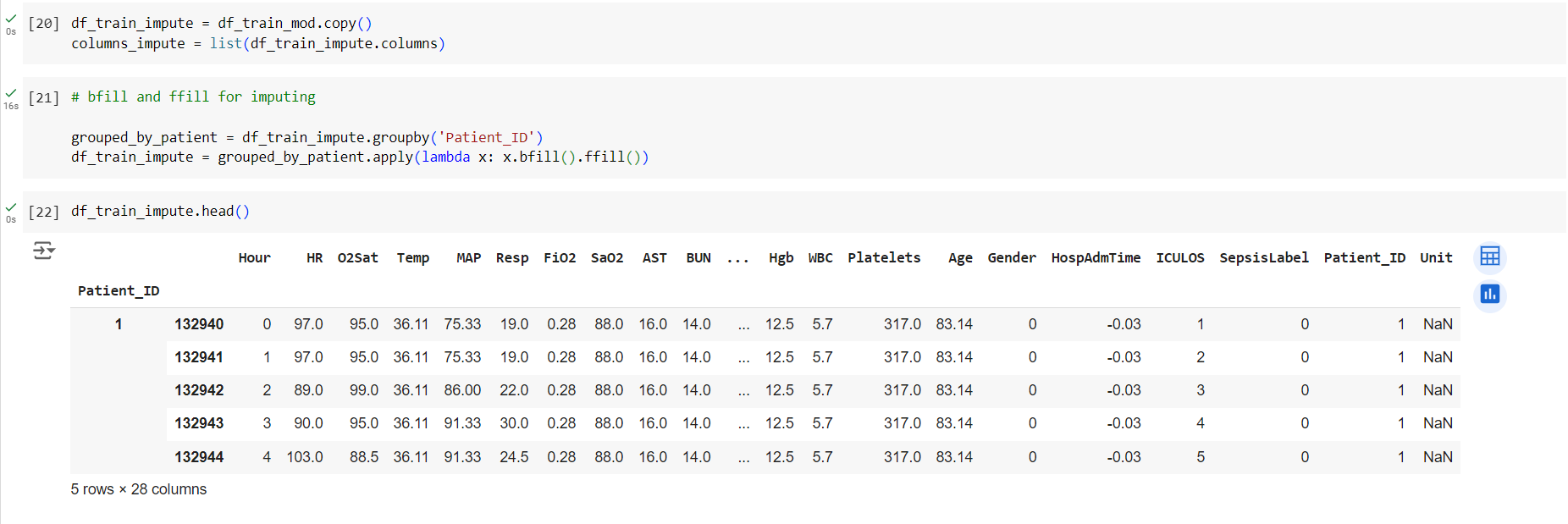
****

*[Fig-15] features have been removed based on the number of null values and redundancy*

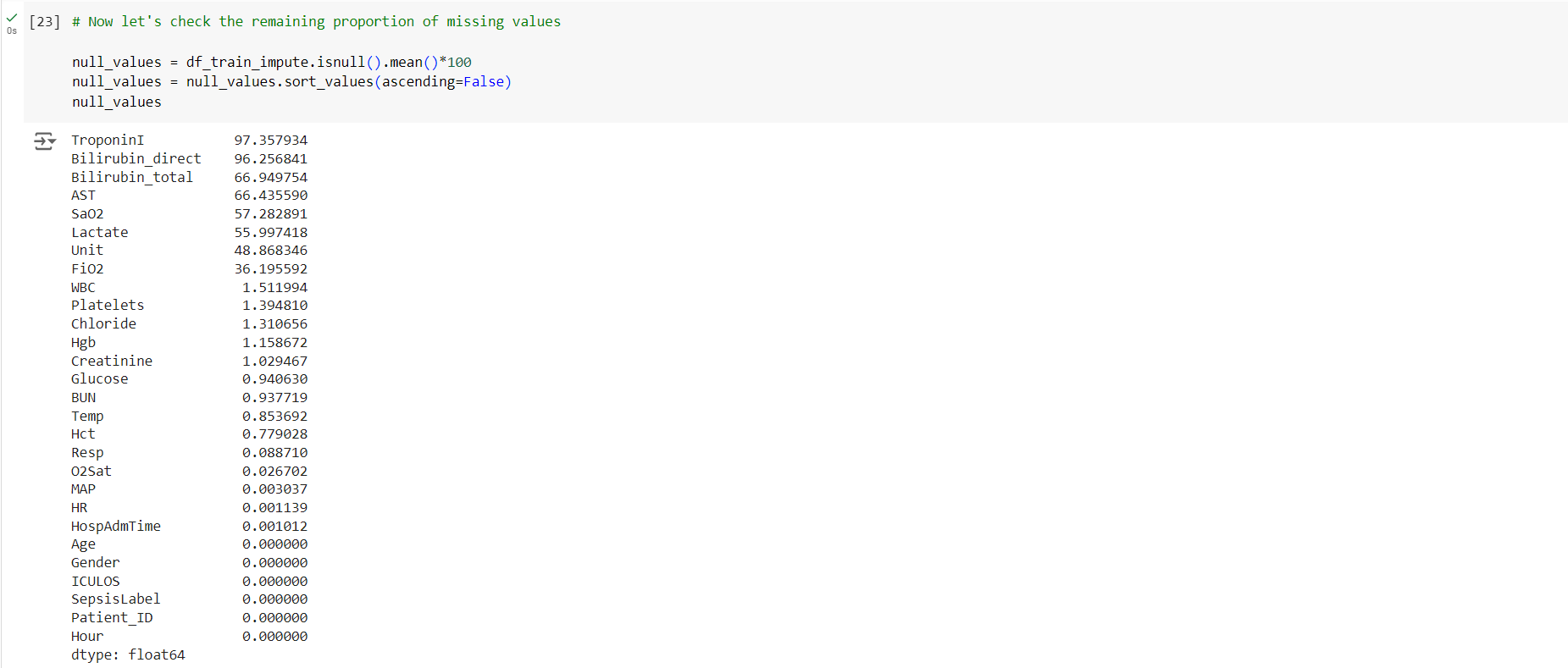
****

*[Fig-16] correlation matrix to check if there is high correlation between the remaining features*

**Imputation**

****

*[Fig – 17] Imputation*

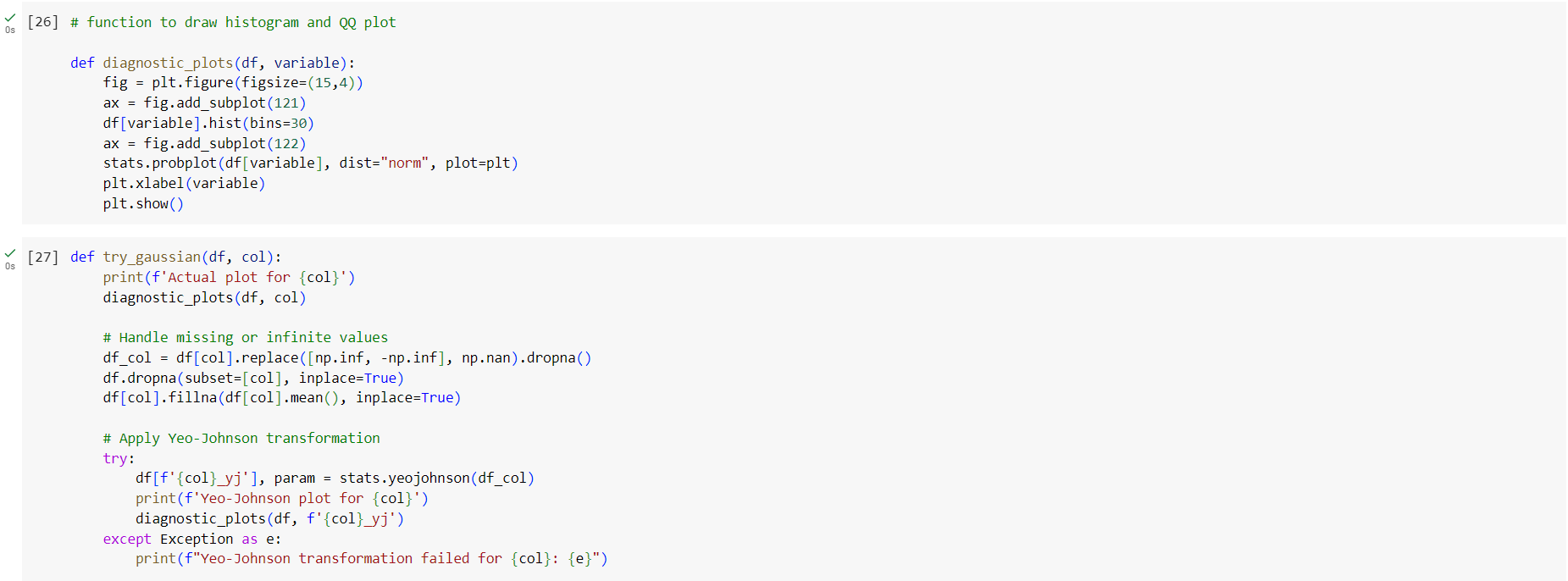
****

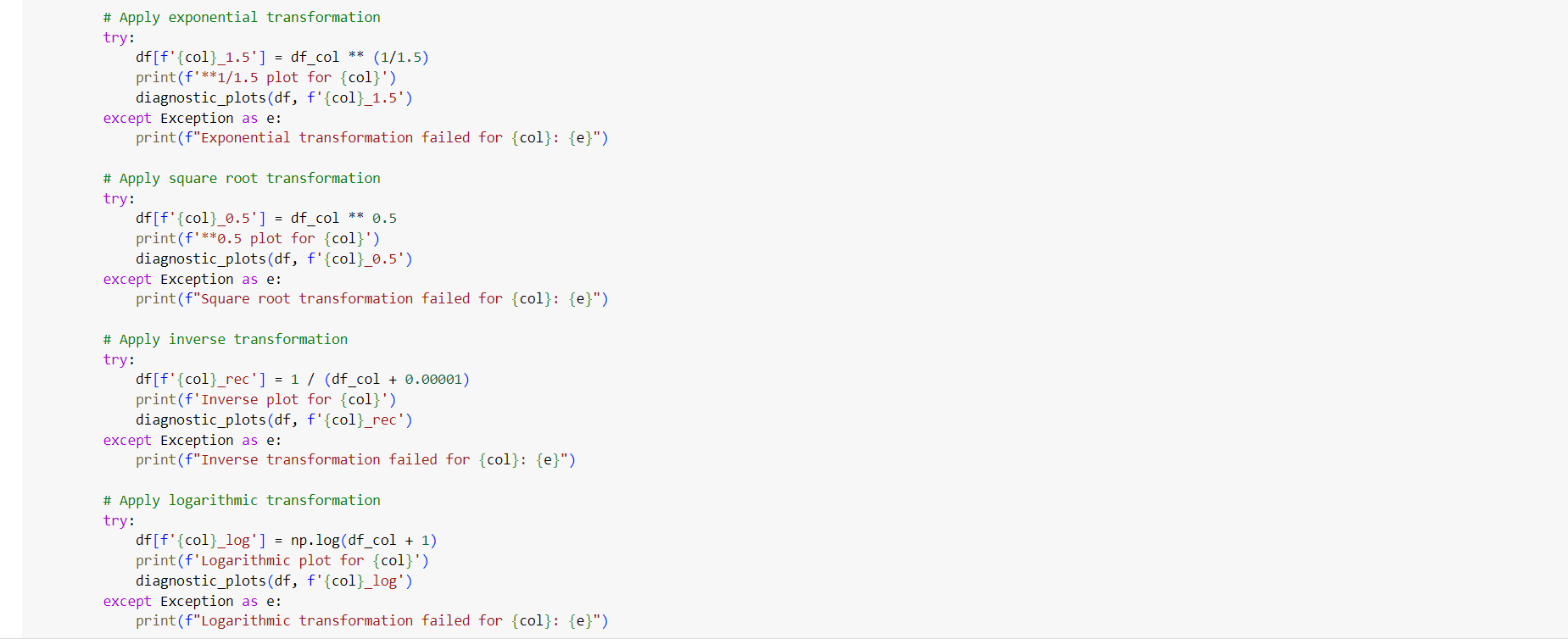
*[Fig – 18] checking the remaining proportion of missing values*

****

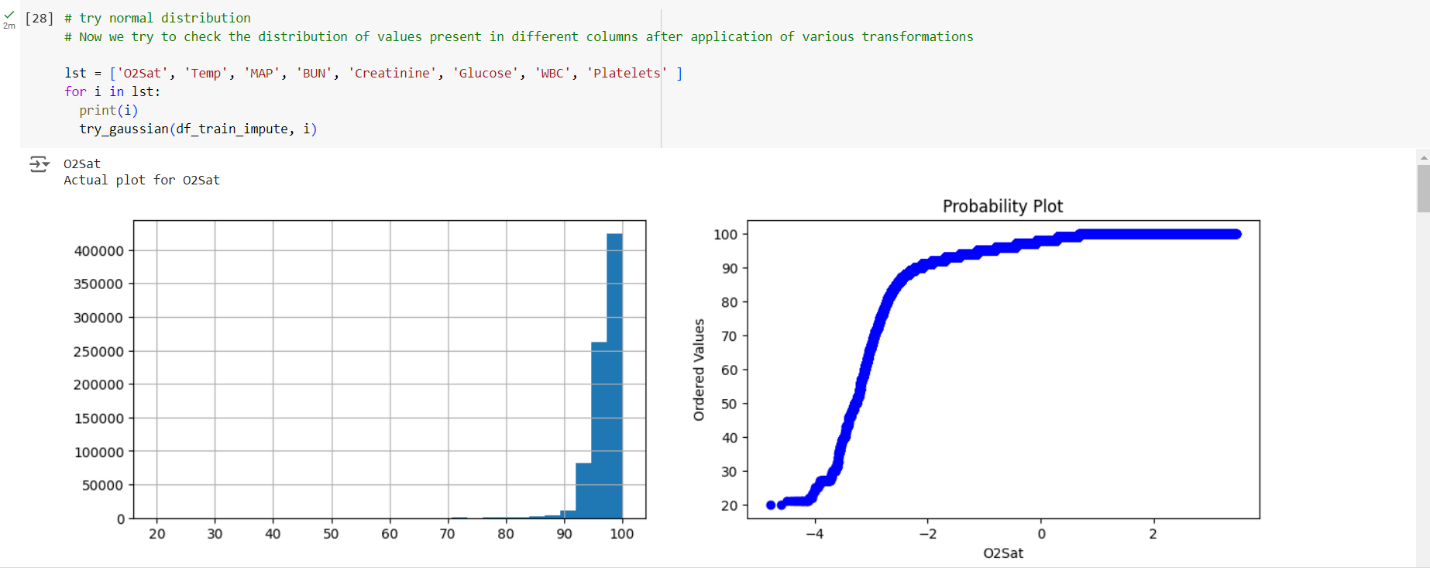
*[Fig-19] dropping all the columns with null values more than 25% and patient\_id*

**Gaussian Transformation and Standard Normalization**

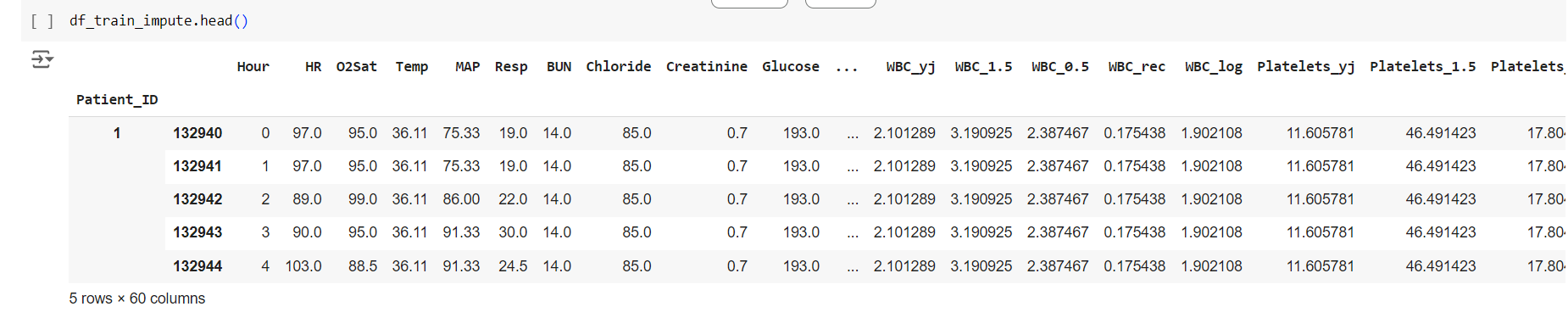
****

****

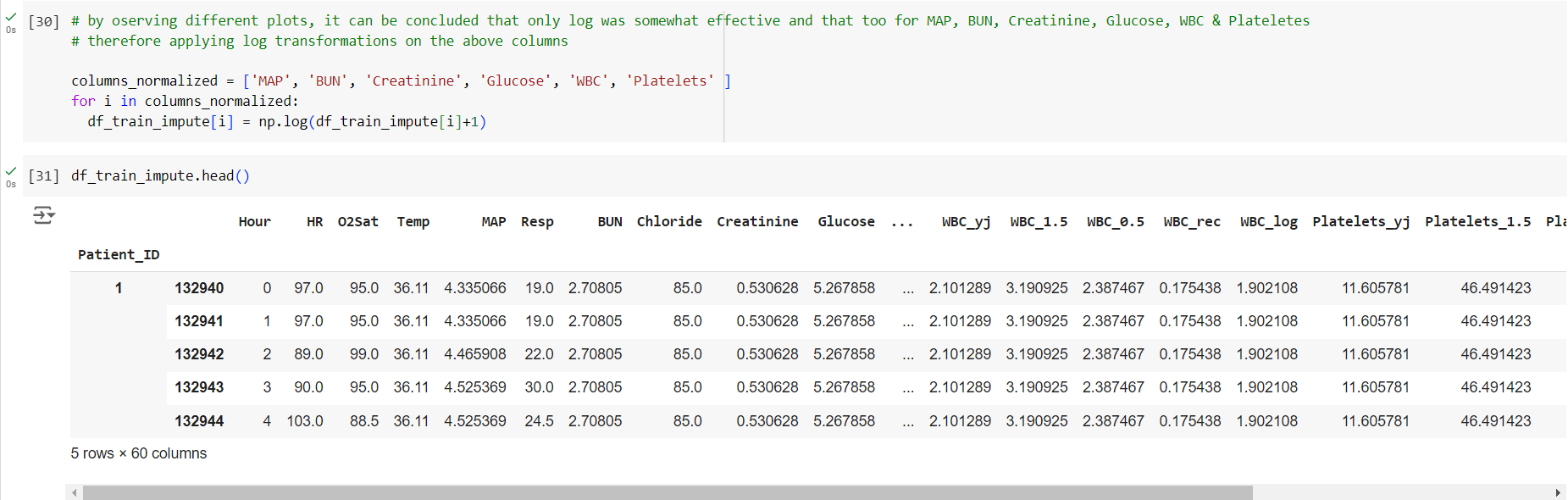
*[Fig – 16] Gaussian Transformation and Standard Normalization*

****

*[Fig – 17] Normal distribution*



*[Fig-18] Display Data*

****

*[Fig-19] Applying log transformation on the above colums*

****

*[Fig-20]-Standard Normalization*

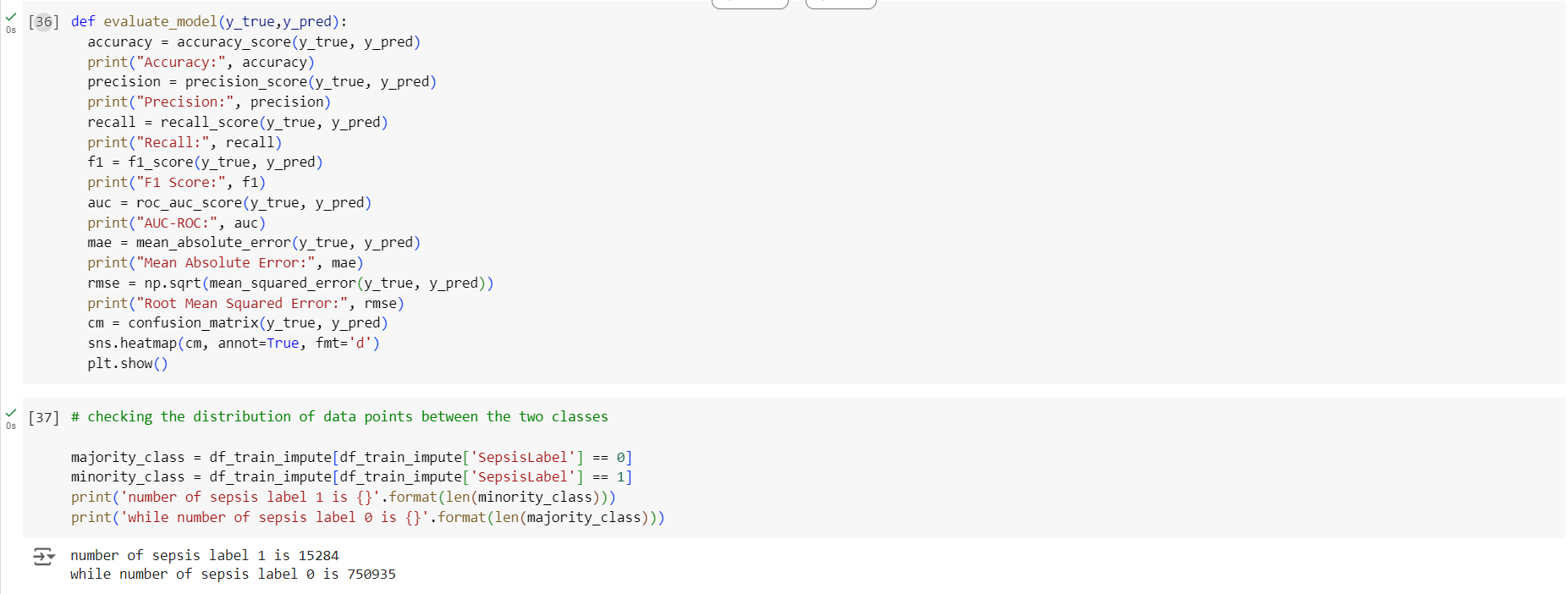




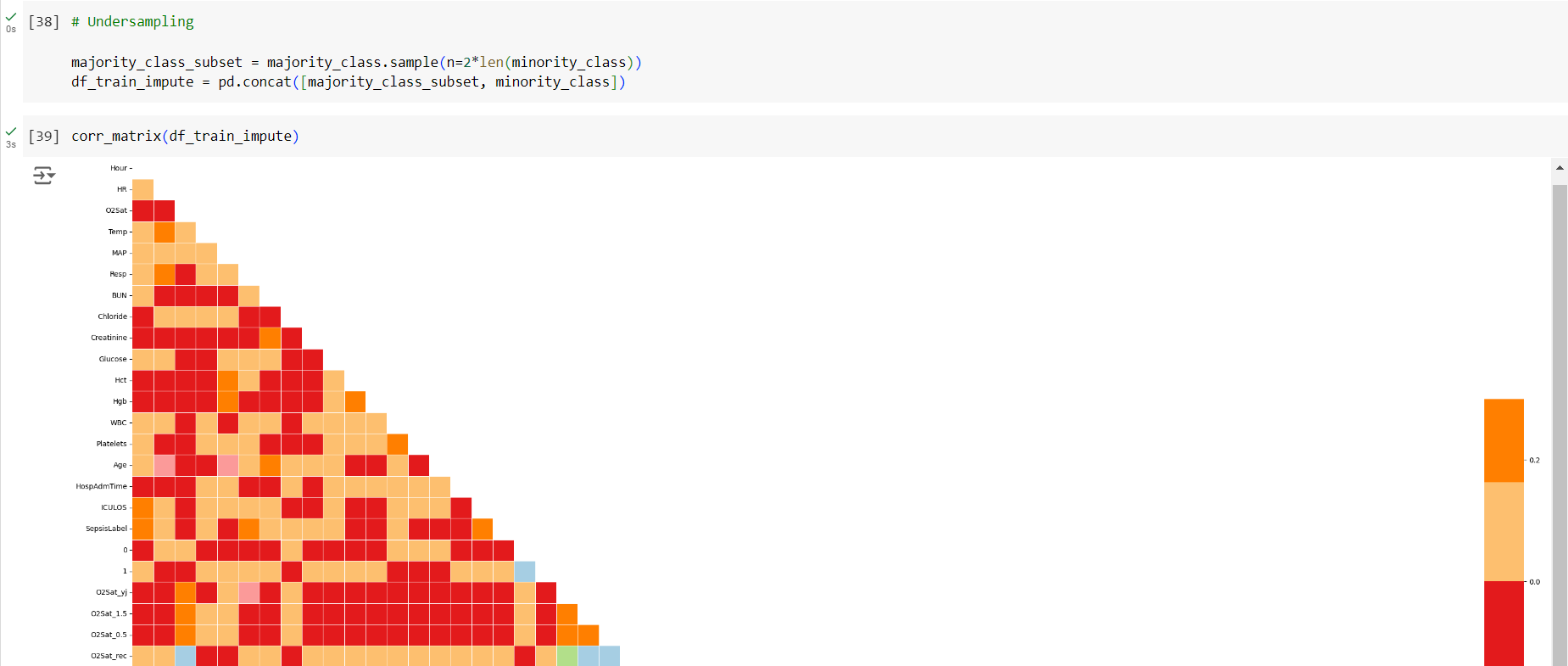
**Model Building**

****

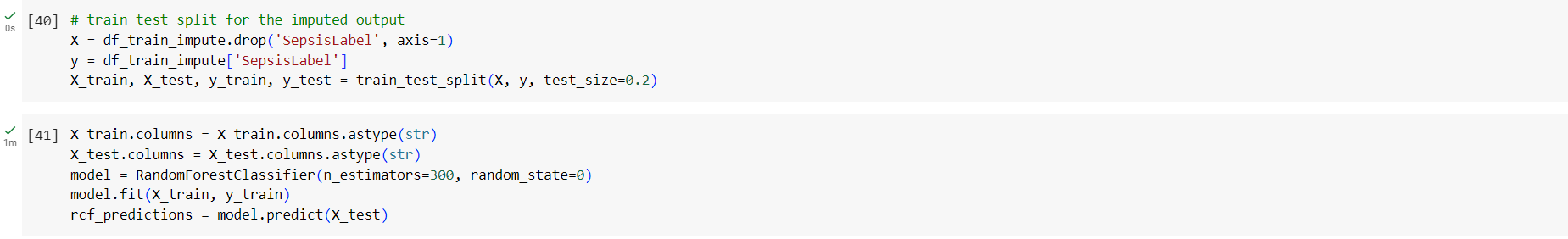
*[Fig – 21] format for the input of the model*

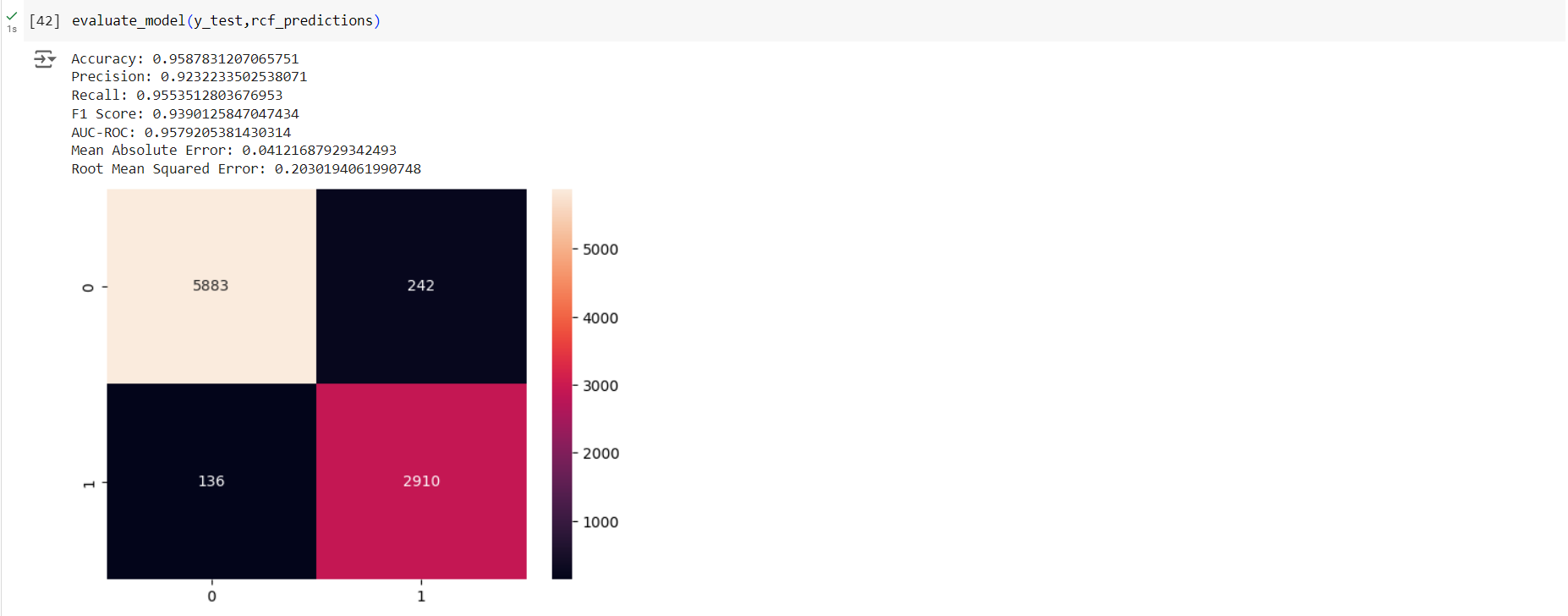


*[Fig – 22] fucntion calculates different evaluation parameters of a model and checking the distribution of data points between the two classes*

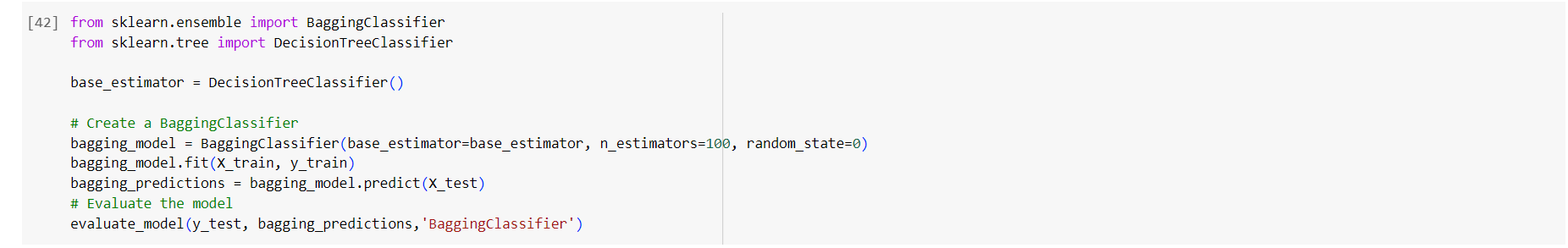
****

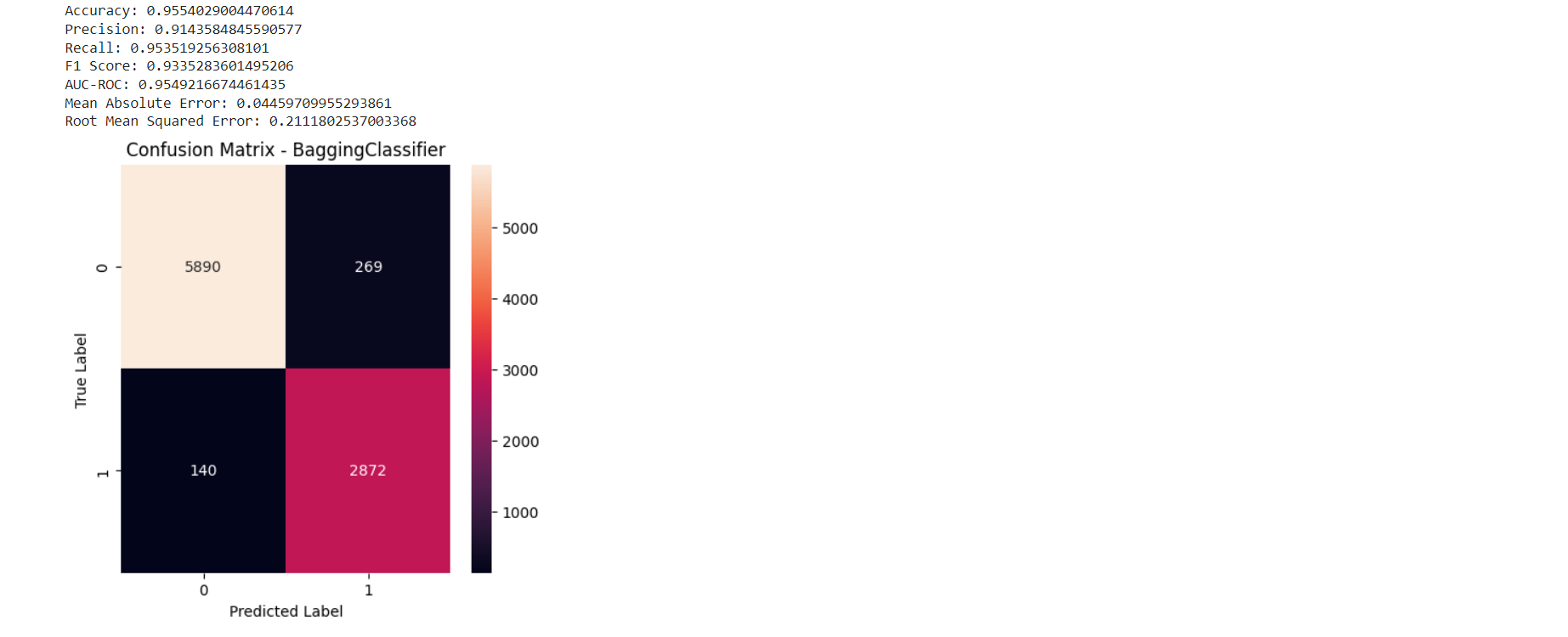
**Random forest**

****

****

*[Fig – 23] Random forest classifier is tested*

****

****

*[Fig – 24] Bagging Classifier*

**7. SOFTWARE TESTING**

Software testing is a testing that is done before actual software is completely executed. The main objective for doing the software testing is the requirements of the expected output is free from errors and defects.

**7.1 Unit Testing:**

Unit testing ensured that each individual function and a part of the process chain was working properly. Each separate feature or transformation function was tested to ensure that its behavior was correct. This testing covered the way these functions work with input at the margins and their handling of strange inputs, such as making sure that the log-transform function does not choke on zeros and ensuring the function get\_data\_ready works in the proper way to transform the data for use by models during input and prediction.

**7.2 Integration Testing:**

This was a test to check whether different parts of the system worked together cohesively. It involved validation of the end-to-end pipeline of data processing from preprocessing to model prediction and checked that data flows were properly engaged and modules like imputation, normalization, and model training had correct interactions. Integration testing confirmed that all steps in the pipeline were correctly applied without data leakage, with consistency across different stages of data preparation and model deployment.

**7.3 Performance Testing:**

Performance testing was conducted with respect to conditions the model is expected to work in, targeting speed and scalability. Load testing estimated how large datasets could be handled, while latency testing measured the time taken to preprocess the data and generate a prediction. This testing ensured that the model was able to process data and deliver predictions promptly; thus, it would be appropriate for real-time applications where timely sepsis detection is paramount.

**7.4 Model Testing:**

All predictive models required testing to understand their performance and reliability. The testing will involve training and validation using cross-validation based on the training set, and models such as Random Forest, Logistic Regression, KNN, and XGBoost will be trained. Computation of key performance metrics such as accuracy, precision, recall, F1 score, and AUC-ROC is to be done to get a better sense of the models' effectiveness. It will also encompass techniques for hyperparameter tuning, like GridSearchCV, in order to optimize the model's parameters for improved performance.

**7.5 Validation Testing:**

Therefore, validation testing would assess how well the model generalizes to other data in hospitals or settings. This will be done through a set used for testing. This validated that the model performs well on another dataset. It confirmed the ability of the model to ensure consistency and accuracy in the predictions it gives in data from different hospitals, hence its robustness and generalizability in real-world scenarios.

**8. RESULTS**

Examining the fundamental parametric requirements, an exploratory data analysis was conducted on the gathered data set.

Plots of probabilities were made between a few characteristics and the sorted values. O2Sat, temperature, MAP, BUN, creatinine, glucose, WBC, and platelets were all plotted. A number of algorithms were used to fit the dataset, and the optimal one was determined using a variety of criteria, including confusion matrix, recall, accuracy, and precision. The log loss and classifier accuracy were plotted.

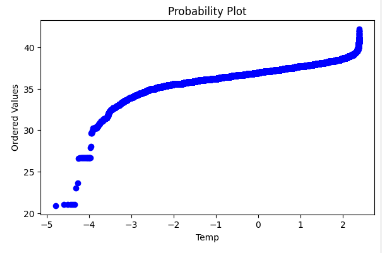
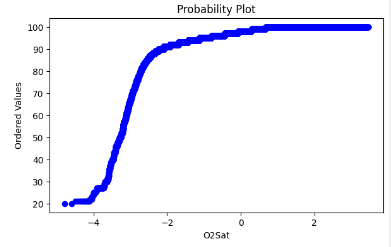


Fig 8(a): 02Sat vs Ordered Values Probability plot Fig 8(b): Temp vs Ordered Values Probability plot

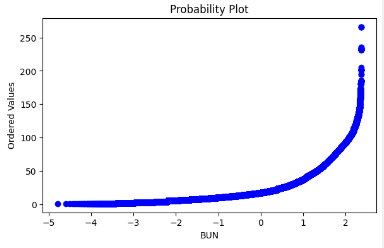
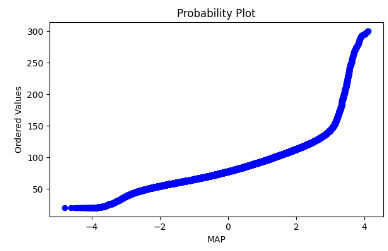


Fig 8(c): MAP vs Ordered Values Probability plot Fig 8(d): BUN vs Ordered Values Probability plot

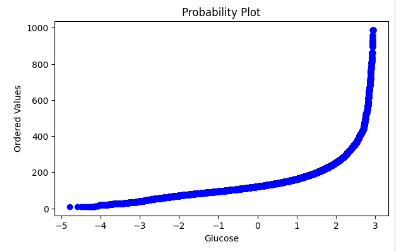
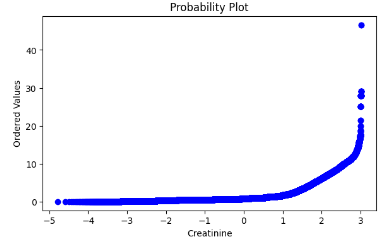


Fig 8(e): Creatinine vs Ordered Values Probability plot Fig 8(f): Glucose vs Ordered Values Probability plot

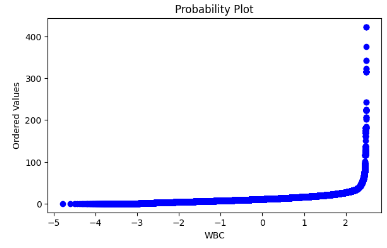
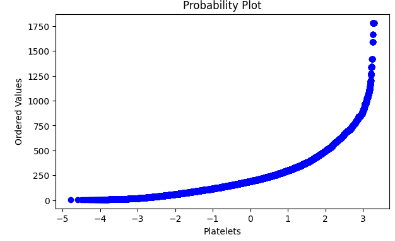
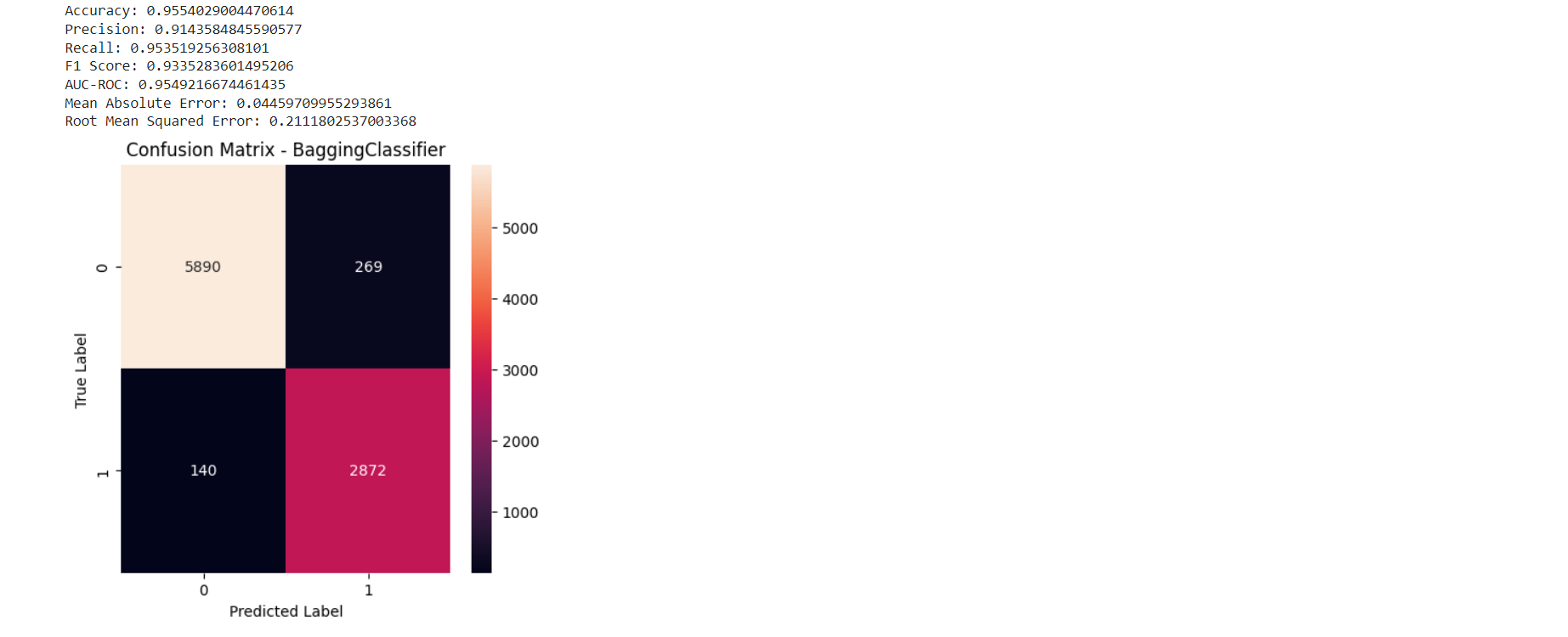
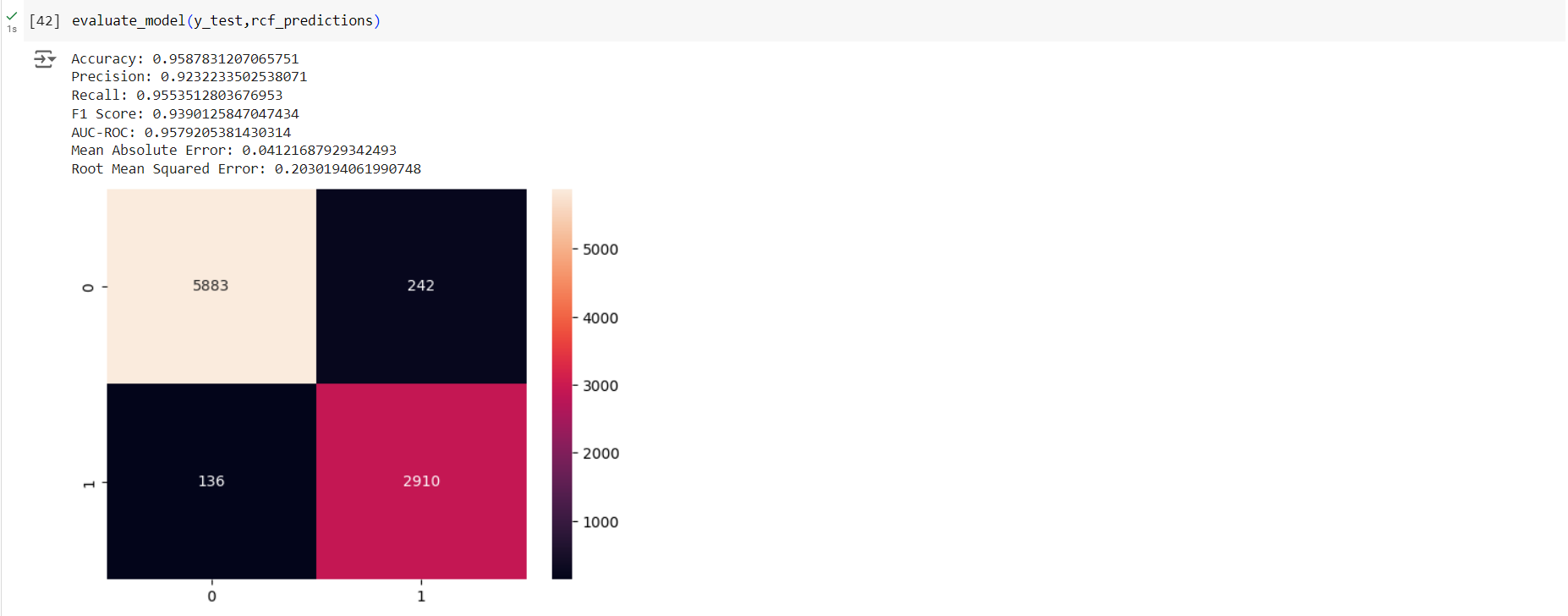
 

Fig 8(g): WBC vs Ordered Values Probability plot Fig 8(h): Platelets vs Ordered Values Probability plot

****

*[Fig-25] Results for Bagging Classifier*

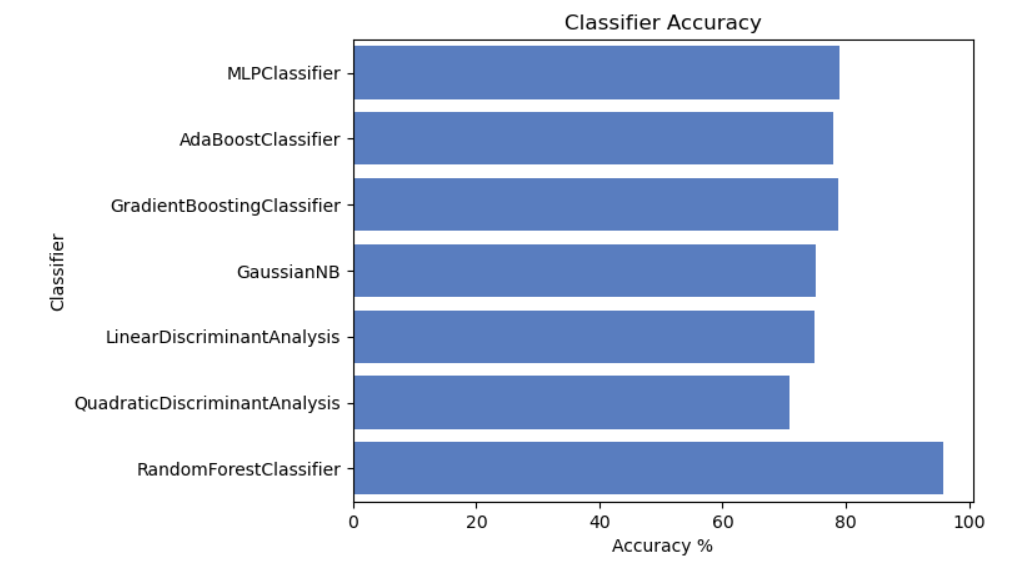
****

*[Fig-26] Results for Random Forest Classifier*

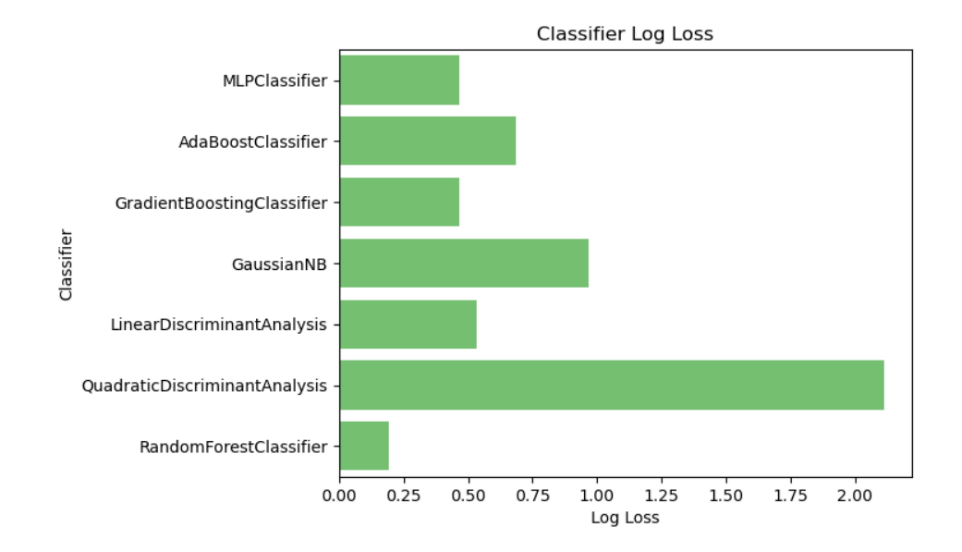
**9. CONCLUSION AND FUTURE ENHANCEMENTS**

Sepsis is a dangerous disease process that may lead to organ failure, tissue damage, or even death. The primary objectives are to diagnose whether a patient has sepsis immediately after entering an emergency department for treatment. In practice, very few methods for sepsis detection exist using LSTM and RNN. However, the major disadvantage of an RNN is that since it's recurrent in nature, the computations are pretty slow with an accuracy of just 82% achieved for an RNN model. For this reason, a solution has been developed where early diagnosis of sepsis can be done using classifiers of Ensemble Techniques like Random Forest Classifier, Bagging Classifier that take into view feature importance, preprocessing, and classification to come over the problems. It returned an accuracy of 96.23%, is fast, and less likely to give false results.

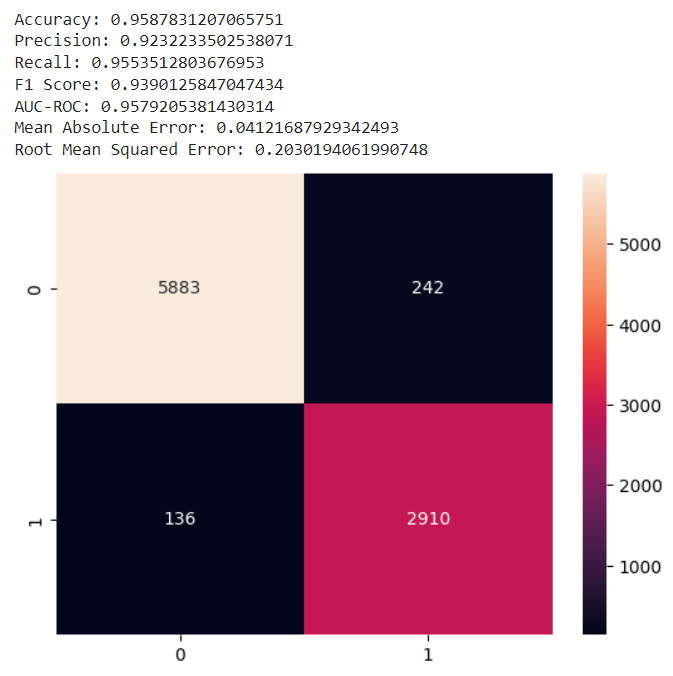
The project also depicts early and more accurate detection of this disease using different classifiers, with primary concern for stacking the classifiers to come up with an accurate prediction model that will not involve much waste of time in laboratory tests. In the near future, we also intend to enhance the working application to become customer-specific and implement this model on any hospital website to help the medical staff to spot any early indication of disease.



*[Fig-26] Classifier Accuracy*



*[Fig-27] Classifier Log Loss*



*[Fig-28] Results for Random Forest Classifier*

**10. BIBLIOGRAPHY**

[1] Effect of a machine learning-based severe sepsis prediction algorithm on patient

survival and hospital length of stay: a randomised clinical trial; David

W Shimabukuro, Christopher W Barton, Mitchell D Feldman

# [2] Practical Machine Learning-Based Sepsis Prediction; Michael J. Pettinati; Gengbo

# Chen; Kuldeep Singh Rajput

[3] A New Effective Machine Learning Framework for Sepsis Diagnosis; Xianchuan

Wang; Zhiyi Wang; Jie Weng

[4] A Machine Learning Algorithm to Predict Severe Sepsis and Septic Shock: Development,

Implementation, and Impact on Clinical Practice; Heather M; Ginestra , Jennifer C MD

# [5] Machine learning for the prediction of sepsis: a systematic review and meta-analysis of

# diagnostic test accuracy; Lucas M.Fleuren, Thomas L.T.Klausch

[6] A Machine Learning Model for Accurate Prediction of Sepsis in ICU Patients; Dong

Wang, Jinbo Li, Yali Sun.

[7] Early Prediction of Sepsis Based on Machine Learning Algorithm; Wenqian Shen, Guanjun

Wang

[8] Early Detection of Sepsis With Machine Learning Techniques: A Brief Clinical

Perspective ; Daniele Roberto, Alessio Singnori

[9] Outcome Prediction of Patients for Different Stages of Sepsis Using Machine Learning

Models;Pankaj Chaudhary,Deepak Kumar

[10] Machine Learning Models for Early Prediction of Sepsis on Large Healthcare

Datasets;Javier Enrique Camacho-Cogollo ,ORCID,Isis Bonet ,ORCID,Bladimir Gil.

# [11] Machine Learning-Based Early Prediction of Sepsis Using Electronic Health Records: A

Systemaric Review: Khandaker Reajul Islam ,Johayra Prithula ,Jaya Kumar.

# [12] Prediction of Sepsis Mortality in ICU Patients Using Machine Learning Methods: Khandaker

# Reajul Islam, Johayra, jAya Kumar