DETECTION OF SEPSIS USING THE ANALYSIS OF MACHINE LEARNING ALGORITHMS

REVATI V
Assistant Professor
Department of Computer Science
GITAM Hyderabad, India

PAVAN KUMAR V
Student

ADARSH C
Student

AKANKSHA B
Student

SAKETH SAI Y
Student

ABSTARCT: Our research uses multiple machine learning algorithms that utilize Electronic Health Records data to predict Sepsis's onset accurately. The ability to precisely classify observations is precious for various medical applications like predicting whether a particular patient needs to hospitalize or forecasting their chances of contracting a severe illness. Sepsis management is highly time-sensitive, and each passing hour of delayed treatment raises the possibility of mortality due to organ damage. The decade's long clinical research never resulted in crucial biomarkers to detect Sepsis on its onset. Therefore, detecting Sepsis early using accurate and reliable EHR data has become a challenge. Recent advancements in Machine learning and data mining have enabled ML scientists to tackle it efficiently.

Key words: Sepsis, Machine Learning, Classifier algorithms, Gradient Boosting classifier, Decision tree, Random Forest

1. INTRODUCTION

Sepsis is an acute immune response to an infection. A person suffering from Sepsis has an immune system that can injure tissues and organs, which can be life-threatening. The signs and symptoms of Sepsis include a high fever, a rapid heart rate, breathing difficulty, and confusion. It is more likely to be observed in older people, younger children, and people with weakened immune systems or specific health issues. Annually, an estimated 48·9 million incident cases of Sepsis were noted globally, and 11·0 million sepsis-related deaths were reported. Around 1 in 3 deaths in hospitals result from Sepsis. It has become imperative to Learn and spot the signs that can help people receive the demanded care quickly. Sepsis is a medical crisis, and quick treatment can save lives.

2. LITERATURE SURVEY

We are highly indebted to the research conducting by scientists in both domains of health and computer science. Our work is a direct extension of several years of study on the disease of Sepsis by the various researcher. There is a significant effort by doctors worldwide to find solutions to diagnose Sepsis before it could take lives. The research article from R. Lakshmi Devi, C. Preetha, C. Kalaiarase, and B. Ponsubashini titled, "Early Prediction of Sepsis using Clinical Data." has been the base for our project. The paper explicitly asserts how we can use clinical data of patients to diagnose Sepsis and how it can help the health infrastructure from collapsing. Another excellent research article that helped us in this project is by Fatima.M and Pasha.M titled "Survey of Machine Learning Algorithms for Disease Diagnostic". The article has explored how we can utilize ML algorithms to identify various diseases. As cited in the bibliography of our paper, the other articles have assisted in carrying out our project in an efficient manner.We have also sourced our knowledge from various articles on Machine Learning from the unrestricted web.

3. RESEARCH METHODOLOGY

3.1. DATASET

The Electronic Health Record (EHR) dataset we have chosen for our research has a patient's medical history that includes various data points like Vital signs, Demographics, Laboratory values, and Sepsis outcomes. The EHR dataset obtained from PhysioNet has a database of 40,000 patients, which facilitates access to information and has the potential to enable smooth research. EHRs contribute to the continued progress of healthcare that can establish the connection between patients and hospitals. The data, timeliness, and availability will enable researchers to make better decisions and provide better care.

3.2. DATA PRE-PROCESSING

3.2.1. CONVERTING .PSV FILES TO .CSV FILES

"Comma-separated value" is a text file format used to describe data in tables where each line in the text file contains data for a single record. A comma or other specific character used to separate field values is called the list delimiter character. Our source's raw dataset was in "Pipe-separated value" format where the values were separated pipes, And it was our task to convert it into a .csv file where Microsoft applications can read it conveniently to process the raw data. Researchers worldwide prefer CSV file format because of its readability, ability to parse, compactness, agility, and size.

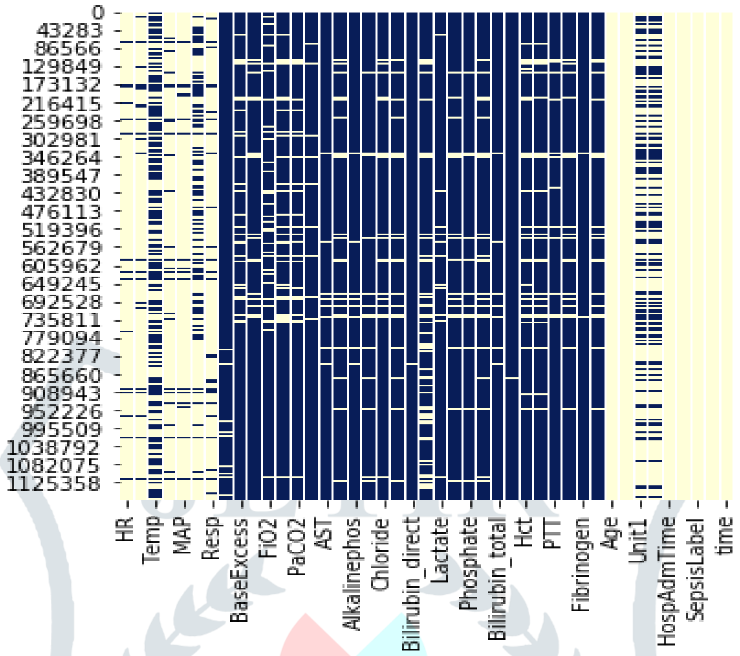


Fig 2.1.1 Dataset before pre-processing

3.2.2. LABEL DISTRIBUTION

Label Distribution Learning is a unique machine learning standard that includes a certain number of labels, representing the extent to which each label describes the occurrence. We have considered the various health markers that are crucial in identifying the symptoms of Sepsis as labels. The creation of labels also enabled us to analyze the data in a much more exciting way with more graphical diagrams. Label distribution helped us to study the spread of each data label. Thus, helping us to remove few health markers from our dataset that are undermining the training efficacy of our machine learning model. The creation of labels also enabled us to analyze the data in a much more exciting way with more graphical diagrams.

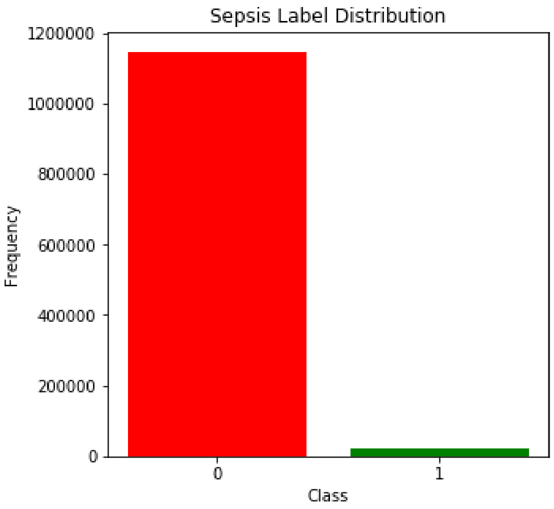


Fig 4.2.2.1 Sepsis Label Distribution

3.2.3. NULL VALUE IDENTIFICATION

The dataset obtained from the real world is expected to be not clean and homogenous. There might be numerous reasons as to why data missed during extraction or collection. Missing values create many difficulties when using the dataset for any machine learning algorithm. They hinder data analysis and data visualization. Missing values must be handled because they diminish the quality, lead to wrong prediction or classification, and cause an exceptional bias for any model. The label distribution helps us to observe the percentage of null values each label has. This process helps us remove health markers that have more than 90% of their values as null. Removing the unnecessary features from the model, in this case, they are the health markers before training the model, will help us avoid the overfitting of the model.

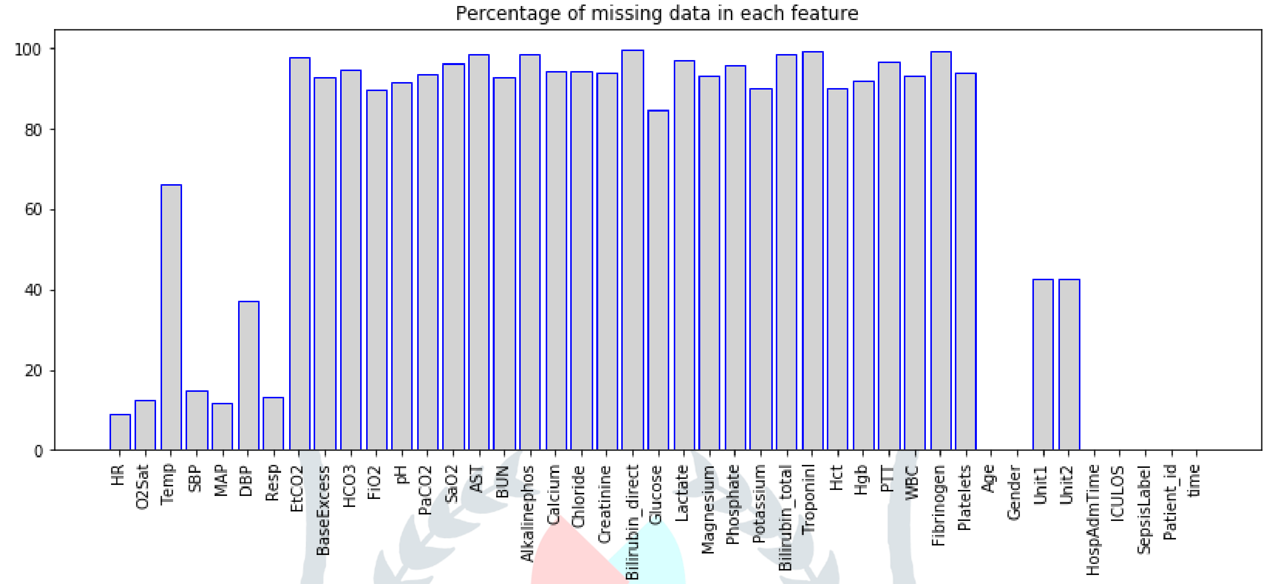


Fig 2.3.1 Null Value Identification

3.2.4. ONE HOTKEY ENCODING

In Machine-learning projects, the data set might contain categorical values, i.e., primarily non-numerical values. Algorithms such as decision trees can manage categorical values great, but most algorithms expect numerical values to produce fruitful outcomes. Most of the algorithms work better with numerical inputs, and hence, the main challenge faced by a researcher is to convert the categorical data into numerical data. One hotkey encoding is one such method used in the conversion of text data into numerical data. Each category value is transformed into a new column in this encoding approach and assigned 1(true) or 0(false) to the column. Since we have used the OneHotEncoder from the SciKit-learn library, we only provided categorical numerical values. Hence, any string type value must be label encoded before one-hot encoding.

3.3. TRAINING

Machine Learning algorithms require a whole collection of data, called a training dataset, to offset further application and utilization. It is the basis for the model's growing archives of knowledge. We require both training and testing data sets to build an ML algorithm. The model trained on a training set is evaluated on a test set obtained from the primary dataset. Commonly, training data is split randomly while capturing known classes upfront. Thus, helping our model prevent overfitting and become more robust. We have used the 80-20 split to divide our original dataset into training and testing datasets, respectively. Since three different algorithms need to be used to train data, we have maintained a similar training dataset to have a level playing field. The remaining 20% data obtained is used to make a testing set, validating older models until the new model provides satisfactory results.

3.3.1. RANDOM FOREST ALGORITHM

Random Forest algorithm is a well-known supervised classification algorithm which, from its name, indicates that the created trees are randomized. The number of trees created in the forests has a direct impact on the results we obtain. This unique algorithm can be used either for classification or regression functions. One essential feature of the Random Forest algorithm is that if there are adequate trees in the forest, we will not obtain an overfitted model. It can manipulate missing values, and it also can be modelled for categorical values. Random Forest algorithm can be used in medicine to identify the correct combination of components in a pill and to recognize diseases by analysing the patient's electronic health records.

3.3.2. GRADIENT BOOSTING ALGORITHM

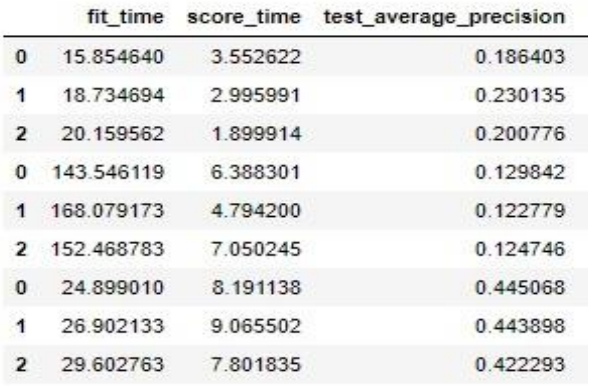
Gradient boosting is a machine learning algorithm adopted for regression and classification tasks, which produces a prediction model as an aggregate of weak decision trees. The concept behind the gradient boosting algorithm is to use the weak learning method several times to get a succession of hypotheses, where each tree is altered upon the patterns that the previous ones found challenging and misclassified. It builds a successive model in a progressive stage-wise fashion; it optimizes arbitrary differentiable loss functions. In every stage, regression trees fit the negative gradient of the binomial or multinomial deviance loss functions. One key feature of the gradient boosting algorithm is that each new tree helps rectify errors caused by the previously trained tree, making the model even more expressive. Each fresh tree helps fix errors made by the earlier trained tree, making the model even more expressive. There are typically three parameters to be weighed while using a gradient boosting algorithm: the number of trees, the depth of trees, and the learning rate. Due to its phenomenal regression abilities, a gradient boosting algorithm fits right for our model to predict Sepsis using health records.

3.3.3. DECISION TREE ALGORITHM

Decision trees learn from previous data to approximate a value of the target variable by learning a set of simple if-then-else decision rules inferred from the data features. The complexity of the decision rules increases with the depth of a tree, making the model more reliable. There are definite actions involved in constructing a decision tree, including splitting, pruning, and tree selection. Decision trees are popular among analysts because of their usability and the need to do more limited data pre-processing. The decision trees can also handle both numerical and text data and can handle multi-output problems. A decision tree functions well even if few assumptions are not per the actual model from which the data was generated. A decision tree classifier can perform multi-class classification on a dataset. Due to its simple nature and the capability to forecast results, the decision tree algorithm fits suitable for our project to predict Sepsis in patients using their health records available at a hospital.

4. RESULTS

The algorithms after training the data three times each, we obtain nine different models, each with unique values of fit time, score time, and the average test precision. In the first column, we can notice that the gradient boosting algorithm takes a significant amount of time to train, followed by Random Forest and Decision tree classifiers. The average test precision in the following column shows that the Random forests have the highest precision among the three, followed by Decision tree and gradient boosting algorithms. The decision tree classifier has the lowest scoring time among the three algorithms, which means it trains and validates the data quickly. The nine models trained using the classes of the respective algorithm in scikit-learn are helpful in various medical applications.



	fit_time	score_time	test_average_precision
0	15.854640	3.552622	0.186403
1	18.734694	2.995991	0.230135
2	20.159562	1.899914	0.200776
0	143.546119	6.388301	0.129842
1	168.079173	4.794200	0.122779
2	152.468783	7.050245	0.124746
0	24.899010	8.191138	0.445068
1	26.902133	9.065502	0.443898
2	29.602763	7.801835	0.422293

Fig 5.1 Results from the three algorithms

5. CONCLUSION & FUTURE WORK

We have achieved our aim to build a machine learning model to predict Sepsis using the patient's electronic health records. The various data pre-processing techniques have helped us shaping efficient models. The large dataset sourced from PhysioNet also played a crucial role in shaping the results obtained by each algorithm. The three algorithms used in our project have performed excellent, but the Random forest classifier has displayed the highest precision and would be the best model to be used in real-time to help save patients' lives from Sepsis. The project is just the beginning of our effort, and we will be appending the machine learning model with a mobile application to provide a successful diagnosis of Sepsis before the onset of symptoms using the health markers of an individual. We want to conclude by stating that the Random forest algorithm has the best precision among the three classifiers and can identify Sepsis more precisely.

6. BIBLIOGRAPHY

1. R. M. Demirer and O. Demirer, "Early Prediction of Sepsis from Clinical Data Using Artificial Intelligence,"2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT),2019, pp. 1-4, doi:10.1109/EBBT.2019.8741834.

2. X. Wang, Z. Wang, J. Weng, C. Wen, H. Chen and X. Wang, "A New Effective Machine Learning Framework for Sepsis Diagnosis," in IEEE Access, vol. 6, pp. 48300-48310, 2018, doi: 10.1109/ACCESS.2018.2867728.

3. R. Lakshmi Devi, C. Preetha, C. Kalaiarase and B. Ponsubashini, "Early Predicton of Sepsis using Clinical Data," 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), 2020, pp. 1-6, doi:10.1109/ICSCAN49426.2020.9262412.

4. Fleuren, L.M., Klausch, T.L.T., Zwager, C.L. et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med 46, 383–400 (2020). https://doi.org/10.1007/s00134-019-05872-y

5. Pepic, I., Feldt, R., Ljungström, L. et al. Early detection of sepsis using artificial intelligence: a scoping review protocol. Syst Rev 10, 28 (2021). https://doi.org/10.1186/s13643-020-01561-w

6. Armando D Bedoya, Joseph Futoma, Meredith E Clement "Machine learning for early detection of sepsis: an internal and temporal validation study" JAMIA Open, Volume 3, Issue 2, July 2020, Pages 252–260, https://doi.org/10.1093/jamiaopen/ooaa006

7. Pierrakos, C., Vincent, JL. Sepsis biomarkers: a review. Crit Care 14, R15 (2010). https://doi.org/10.1186/cc8872

8. Fatima, M. and Pasha, M. (2017) Survey of Machine Learning Algorithms for Disease Diagnostic. Journal of Intelligent Learning Systems and Applications, 9, 1-16. https://doi.org/10.4236/jilsa.2017.91001

9. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Intern Med.* 2018;178(11):1544–1547. doi:10.1001/jamainternmed.2018.3763