Death and ICU Requirement prediction of covid patient using Decision Tree and Random forest

Nazia Afrin
2018-2-60-023@std.ewubd.edu
Department of Computer
science and Engineering
East West University

Sumona Yeasmin
2018-2-60-062@std.ewubd.edu
Department of Computer
science and Engineering
East West University

Moumita Das
2018-2-60-097@std.ewubd.edu
Department of Computer
science and Engineering
East West University

Abstract: Corona virus is a respiratory disease that can attack both human and animals. In humans. Corona virus can start causing simple cold like flu. COVID-19 is the contagious disease caused by the most recently discovered coronavirus.

But the Medical symptoms of a patient Can identify whether someone has Covid or not. In this research, a clinical symptom dataset of covid patient who are admitted to hospitals will be used to classify the symptoms using a Decision Tree algorithm and random forest algorithm. Our concept is to cover a dataset into a model such that the covid patient's death can be predicted or be decided via algorithm. We have implemented both of the algorithm but noticed that the Decision Algorithm gives us the best result.

Keywords—covid, corona virus, random forest, decision tree, classification.

I. INTRODUCTION

Covid-19 is a virus that attacks the human respiratory system. This virus was first discovered in the city of Wuhan, China in December last year. This virus is capable of causing death. However, the number of people infected by the corona virus is still increasing every day. This can be concerning because the virus itself can be contagious in a significant level and also it can cause death. The symptoms of Covid-19 are similar to the symptoms of the common cold or flu which can rapidly escalate to severe breathing problem. There can be very little symptoms among the patient but the condition of a patient could be worsening withing days. To control from being affected by this disease firstly we have to detect the already affected people .We can easily do this by using decision tree and random forest . The decision tree can decide whether a covid-19 patient is dead or not based on the symptoms as well as it can also decide if a patient requires ICU.

A. Decision Tree

A decision tree is a classification method that uses a tree structure, where each node represents an attribute and the branch represents the value of the attribute, while the leaves are used to represent the class. Decision Trees are usually being implemented as like human thinking while making a decision. The logic behind the decision tree can be easily understood because it is a tree structure.

The decision tree starts at root which turns into possible outcomes by spreading branches.

The decision tree is used in many real-life sectors, such as engineering, civil planning, law, and business. This can evaluate prospective growth for businesses based on previous historical data. Decision tree helps finding prospective clients in business purposes. Decision trees are easy to read and interpret, easy to prepare and less data is required when a variable is created once.

B. Random Forest Algorithm

Random forest is a supervised learning algorithm. Each "forest" it builds, is an visualization of multiple decision trees. Random forest is easy to use machine learning algorithm which produces, a great result most of the time. It is also one of the most popular algorithms, because of it is simple and easy to interpret.

Overfitting problems can be solved by using random forest algorithm if there is an enough number of trees. Hyperparameters, it uses often produce a good and accurate prediction result. But more accurate prediction requires more trees, which results the algorithm quite ineffective and slower to develop for real time situations, though they can handle a lot of different types feature. In spite of this, the random forest algorithm is widely used in a lot of different fields, like banking, the stock market, medicine and e-commerce.

II. RESEARCH METHOD

Our Work has a sequential flow that was followed to achieve the result. At first when we found the dataset it was needed to perform some preprocessing. In order to implement the algorithm and output the best result this data has been preprocessed by some criteria. After that the dataset has been split into test dataset and trained dataset. This is also an important step for the algorithm to work. And finally, we make the last step by implementing the two algorithms Decision tree and random forest and compare the result to measure the accuracy.

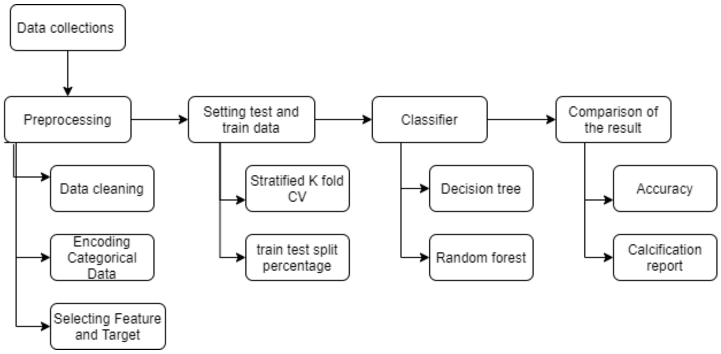


Fig : Process flow

A. Dataset:

The dataset we have used in this analysis is the covid dataset. This dataset was taken from Kaggle website. The dataset has 23 columns with one output. The covid dataset has the health condition of patient who were admitted to hospitals. Here is the column description for the dataset.

1. id: The identification number of the patient
2. sex: Identify gender of the patient, 1 as female and 2 as male.
3. Patient type: Type of patient, 1 for not hospitalized and 2 for hospitalized.
4. Entry date: The date that the patient went to the hospital.
5. Date symptoms: The date that the patient started to show symptoms.

6. Date died: The date that the patient died, 0 means not died and 1 means not dead
7. In tubed: Intubation is a procedure that’s used when you can’t breathe on your own. “1” for yes, “2” for no, “97” for not applicable and “99” for not mentioned.
8. pneumonia: Indicates whether the patient already have air sacs inflammation or not “1” for yes, “2” for no and “99” for not mentioned.
9. age: Specifies the age of the patient.
10. pregnancy: Indicates whether the patient is pregnant or not, “1” for yes, “2” for no, “97” for not applicable and “98” for ignored.
11. diabetes: Indicates whether the patient has diabetes or not, “1” for yes, “2” for no and “98” for ignored.
12. copd: Indicates whether the patient has Chronic obstructive pulmonary disease (COPD) or not, “1” for yes, “2” for no and “98” for ignored.
13. asthma: Indicates whether the patient has asthma or not, “1” for yes, “2” for no and “98” for ignored.
14. inmsupr: Indicates whether the patient is immunosuppressed or not, “1” for yes, “2” and “98” for ignored.
15. hypertension: Indicates whether the patient has hypertension or not, “1” for yes, “2” for no and “98” for ignored.
16. other disease: Indicates whether the patient has other disease or not, “1” for yes, “2” for no and “98” for ignored.
17. cardiovascular: Indicates whether if the patient has heart or blood vessels related disease, “1” for yes, “2” for no and “98” for ignored.
18. obesity: Indicates whether the patient is obese or not, “1” for yes, “2” for no and “98” for ignored.
19. renal chronic: Indicates whether the patient has chronic renal disease or not, “1” for yes, “2” for no and “98” for ignored.
20. tobacco: Indicates whether if the patient is a tobacco user, “1” for yes, “2” for no and “98” for ignored.
21. contact other covid: Indicates whether if the patient has contacted another covid19 patient. “1” for yes, “2” for no and “99” for not mentioned.
22. icu: Indicates whether the if the patient had been admitted to an Intensive Care Unit (ICU “1” for yes, “2” for no, “97” for not applicable and “99” for not mentioned.
23. covid res: 1 indicates person is covid +ve,2 indicates person is covid -ve,3 indicates result is in awaiting process.

B. Preprocessing:

Prepossessing is an important part of the data analysis and researching as it is the key to have the most correct accuracy. At first attributes that are unnecessary for out output was removed (ie. Id, date symptoms etc.) these attributes have no significance on out model output. After the elimination of these attributes, we have worked with 17 columns as input and the target attribute is ‘date died’ and ‘ICU.

For the first Algorithm that predicts the death of a covid patient, we separated the patients who have covid, after that we have set boundaries in the ‘age’ column finally we split the dataset in to feature and target column the feature attributes are 'sex', 'patient_type', 'intubed', 'pneumonia', 'age', 'pregnancy','diabetes', 'copd', 'asthma', 'inmsupr', 'hypertension', 'other_disease', 'cardiovascular', 'obesity', 'renal_chronic', 'tobacco', 'contact_other_covid', 'covid_res', and the target column is ‘date died’. In order to get the prediction weather a covid-19 patient has died or not we also did some pre-processing to the output column making sure the output is 0 for not dead and 1 for dead.

Similarly for the second Algorithm that predicts if a covid-19 patient needs an ICU or not, we also did similar pre-processing method as like as the previous one to predict the target variable.

C. Setting Test Dataset:

This section contains two part: The Cross validation, The splitting of test and train data

i)Applying the Cross validation:

Cross validation is a method that is widely use to find the best optimal result for the dataset. Cross validation is used here so the dataset is more generalized. The purpose of cross-validation is to define a dataset to "test" a set of data in the training phase. It tests the data and outputs the best accuracy. This is useful because as we have a predicted problem and in a predictive problem there are test datapoint and train datapoint are present. As there could be an issue called “model overfitting’, The cross validation prevents this issue.

In our research, we have used Stratified K fold mechanism to find the best accuracy in decision tree setting the n split to 10. We did not cross-validate the model for Random Forest as it takes too much time.

And for specifying our own model we set min_samples_split = 3000 for the Decision Tree for the first case and min_samples_split = 5000 for the second case. We changed this value manually for the better accuracy.

And for the random forest we set the n estimator to 400 for the covid dead prediction and 300 for the ICU Requirement prediction.

ii) Split data into train and test:

The covid dataset will be used as both test and trained dataset. For both of the algorithm we will split this dataset into 20% which implied that 20% data will be for testing purpose and the rest 80% will be for training purpose.

D. Classifier Algorithm:

For this dataset as there are many possible outputs analysis can be done using Decision tree and Random Forest, we will implement this algorithm for a two type of targets.

E. Comparison of the result:

After finishing the classifier implementation for both of the target variable (ICU and Covid Dead) the result found will be compared with each other, For example when the Decision tree and random forest are used for target variable prediction (ICU) the result of these two classifiers will be compared to decide which one is better for this dataset. And for the second output column (Died) the comparison will be done again.

III. Result Analysis

In this section we will discuss about the accuracy of decision tree and random forest algorithm

Classification report: (Predicting ICU requirement in Random forest and Decision tree

Accuracy using random forest: 91.56 %

output	precision	recall	f1-score	support
ICU needed	0.48	0.30	0.37	1130
ICU not needed	0.94	0.97	0.95	12512
Accuracy			**0.92**	**13642**
macro avg	0.71	0.64	0.66	13642
weighted avg	0.90	0.92	0.91	13642

Table: 1.1 classification report for Random forest

Accuracy using Decision Tree: 91.82 %
Maximum Accuracy: 92.0833 %
Minimum Accuracy: 91.0717 %
Overall Accuracy: 91.598 %
Standard Deviation is: 0.002836447539867664

output	precision	recall	f1-score	support
ICU needed	0.51	0.44	0.47	1130
ICU not needed	0.95	0.96	0.96	12512
Accuracy			0.92	13642
macro avg	0.73	0.70	0.71	13642
weighted avg	0.91	0.92	0.91	13642

Table: 1.2 classification report for Decision tree

Algorithm implementation Results

Decision tree: (ICU requirement prediction of covid patient)

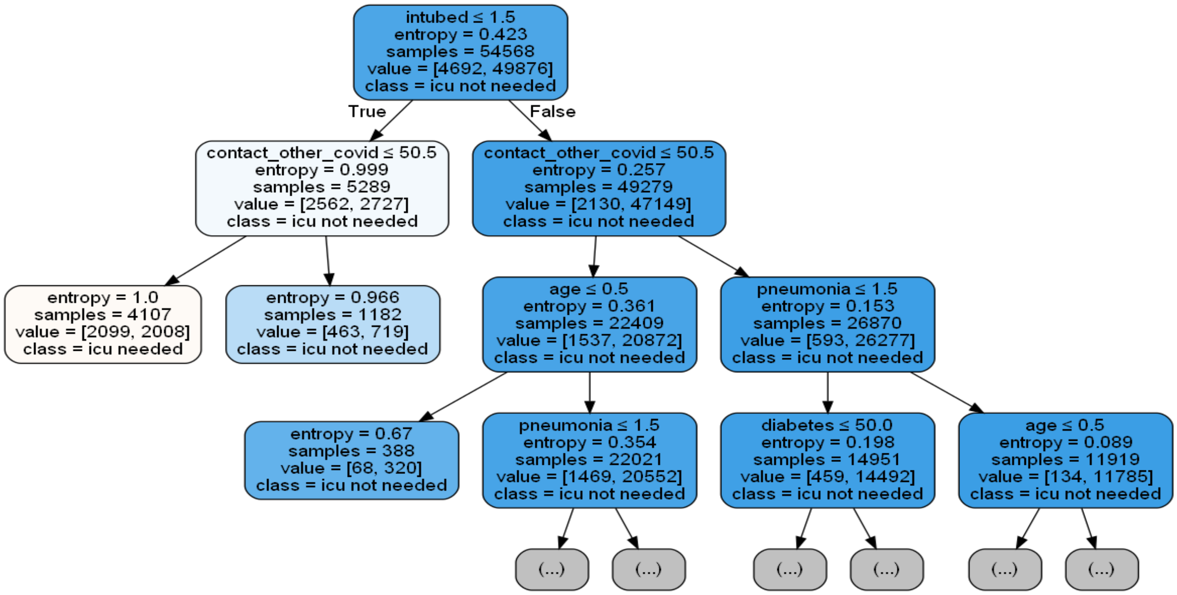
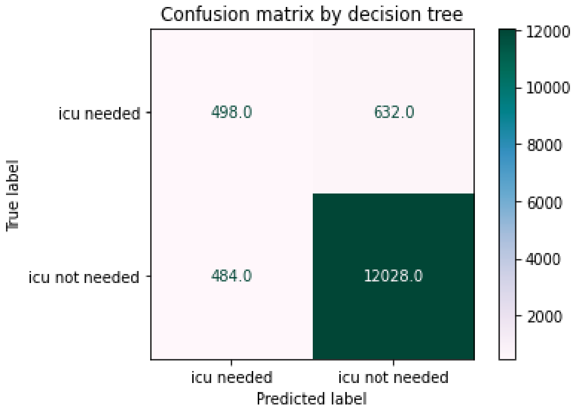


Fig 1: Decision tree for predicting requirement of ICU for covid-19 patient

Confusion Matrix



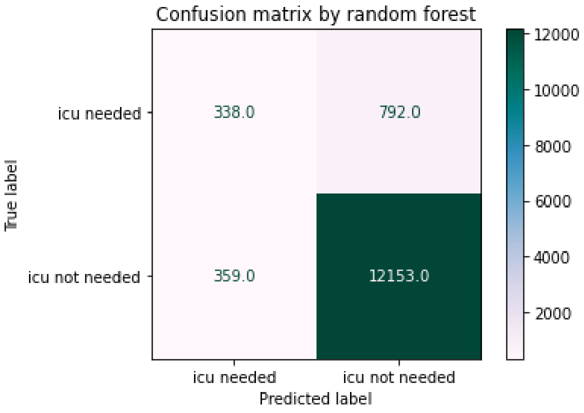


Fig2: Confusion matrix for decision tree

Fig 3: Confusion matrix for Random Forest

Classification report: (Predicting Death in Random forest and decision tree):

Accuracy using random forest: 89.49 %

output	precision	recall	f1-score	support
Not dead	0.93	0.96	0.94	38718
Dead	0.59	0.46	0.52	5414
Accuracy			0.89	44132
macro avg	0.76	0.71	0.73	44132
Weighted avg	0.89	0.89	0.89	44132

Table: 2.1 classification report using random forest

Accuracy using Decision Tree: 89.89 %

Maximum Accuracy: 90.1341 %

Minimum Accuracy: 89.5989 %

Overall Accuracy: 89.8512 %

Standard Deviation is: 0.0019196721035533372

output	precision	recall	f1-score	support
Not dead	0.92	0.97	0.94	38718
Dead	0.64	0.41	0.50	5414
Accuracy			0.90	44132
macro avg	0.78	0.69	0.72	44132
Weighted avg	0.89	0.90	0.89	44132

Table: 2.2 classification report using Decision Tree

Decision tree: (predicting the death of covid patient)

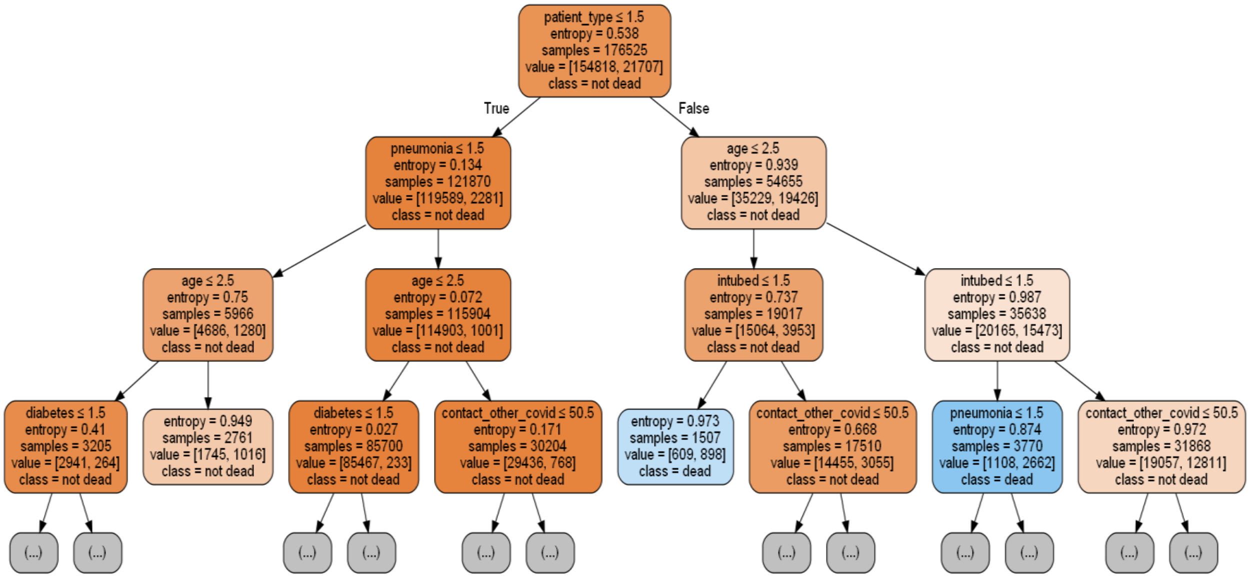
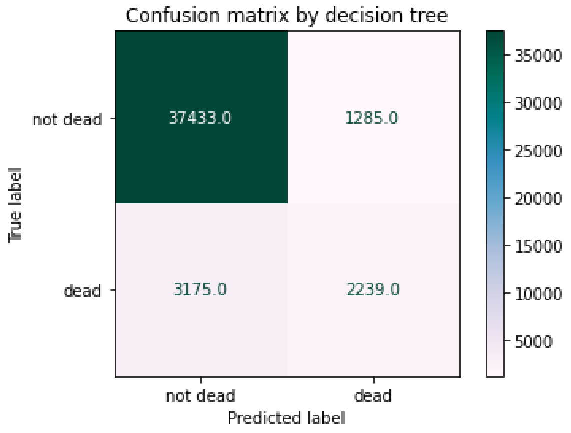


Fig 4: Decision tree for predicting the death of covid-19 patient

Confusion Matrix



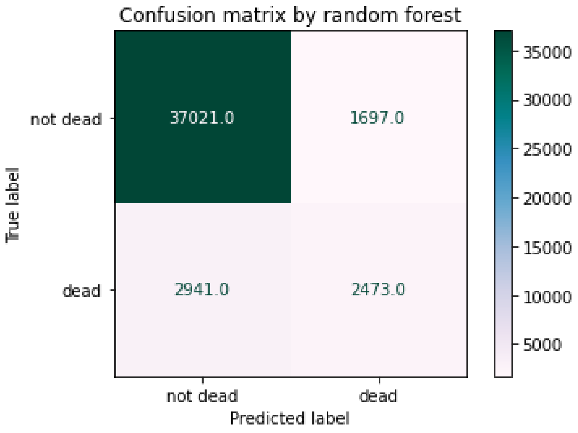


Fig5: Confusion matrix for decision tree *Fig 6: Confusion matrix for Random Forest*

From table 1.1 and table 1.2: The observation can be seen of this classification report for random forest for both ICU prediction and death prediction of covid patient. This report has been generated for random forest and Decision Tree algorithm. It shows the main classification metrics precision, recall and f1-score. Precision is the ability of a classifier not to determine a result positive that is actually negative. The recall value signifies what percentage of positive cases we have managed to catch. The f1 score means the percentage of positive predictions that were correct. The range for F1 score is 1 as in the best and 0 as in the worst.

From these clarification reports we conclude that the overall accuracy of random forest for predicting the ICU requirements of covid- 19

Is 91.56 % and for decision tree it's 91.82% which signifies that while predicting the ICU requirement decision tree does a slightly better job.

From table 2.1 and table 2.2 : The observation shows another set of classification report based on the decision tree and random forest while predicting the Death of a covid-19 patient

The overall accuracy using Decision tree is 89.89 % and Random forest 89.49% and again wee can see the decision tree gives in the more accurate value.

From figure 1 and 4: It shows the visualization of Decision Tree.

From figure 2 and 3: In figure 2, there are 12028 True negative patients who don't need ICU and 498 True positives who do need ICU that is measured correctly by decision tree. Here, 484 False positive patients don't need ICU actually but they need ICU as per prediction and 632 False negatives who do not need ICU according to prediction but do need actually.

But in figure 3, there are 12153 True negative patients who don't need ICU and 338 True positives who do need ICU that is measured correctly by random forest algorithm. Here, 359 False positive patients need ICU as per prediction but they don't need ICU actually and 729 False negatives who need ICU actually but according to prediction they do not need.

From figure 5 and 6: The confusion matrix gives an idea for evaluating the predicted values by a classifier. The confusion matrix of both decision tree and random forest showing the True positives and False negative values.

In figure 5, classified by decision tree correctly, there are 37433 True negative patients who are not dead and 2239 True positives who are dead; again, there are 1285 False positives who are not dead actually but dead as per prediction and 3175 False negatives who are dead actually but not dead as per prediction which are misclassified by decision tree.

But In figure 6, there are 37021 True negative patients who are not dead and 2473 True positives who are dead correctly classified by random forest algorithm. Besides, being misclassified there are 1697 False positives who are not dead actually but dead as pr prediction and 2941 False negative who are dead actually but not dead as per prediction.

IV CONCLUSION AND FUTURE ANALYSIS

Decision tree and random forest are both efficient classifiers, both of them are simply , easy to understand and easy to implement , The comparison of these two classifier in terms of predicting both of the target attribute (ICU requirement and Death prediction) are not very different . This research shows that while dealing with a larger number of datapoints and columns Decision tree can output slightly better accuracy. It is also noted that decision tree is easier to interpret and understand and it has a smaller number of nodes than Random forest.

In future, the research work can also be carried out using the Different dataset with different preprocessing. Research can be conducted using different variation in cross validation and changing the split percentage also.

REFERENCES

[1]https://muthu.co/understanding-the-classification-report-in-sklearn/?fbclid=IwAR2pOVfByMzNFFf_8VnlJ_rCjfMfmLd71YXGavk7gthCY-sBFbCIsLrBJaw#:~:text=A%20Classification%20report%20is%20used,predictions%20from%20a%20classification%20algorithm.&text=The%20report%20shows%20the%20main,positives%2C%20true%20and%20false%20negatives

[2] https://en.wikipedia.org/wiki/COVID-19

[3]https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/