A Novel Corona Virus Detection and Validation Measures using Machine Learning Techniques

***G. Dinesh*[1], *Dr Ali Mirza Mahmood*[*2]**
Department of Computer Science & Engineering,
Krishna University, Andhra Pradesh, India.
****Corresponding Author***
E-mail Id:*-alimirza.md@gmail.com*

ABSTRACT

Data mining is a process of extracting unknown or hidden knowledge from the existing data. This is mainly used for predicting the future based on the past data. Classification in data mining is a common technique that classifies data instances into different classes. It allows you to organize data sets of all sorts, including complex and large datasets as well as small and simple one. Decision tree uses the data to generate sequence of if else rules for decision making. In this paper, we are discussing the pandemic covid-19 related dataset of 1,81,884 instances with 9 attributes. The real world covid-19 data is used to build model to extract the important rules about who had a likely chance to get covid-19 positive. This paper includes one of the algorithms of the decision tree known as C4.5.The experimental results provide are good set of rules for corona virus detection.

Keywords:-*Data mining, Classification, decision tree, C4.5, Covid19 dataset.*

INTRODUCTION

A decision tree consists of root, branches and leaves with class labels. The data is sent in a top down approach to classify to its respective label. Decision tree is one of the simplest and most generalized structures for data mining and knowledge discovery [1-4].

Figure 1 represents the decision tree model constructed using the weather dataset on C4.5 algorithm. The label results yes, no as leaf nodes.

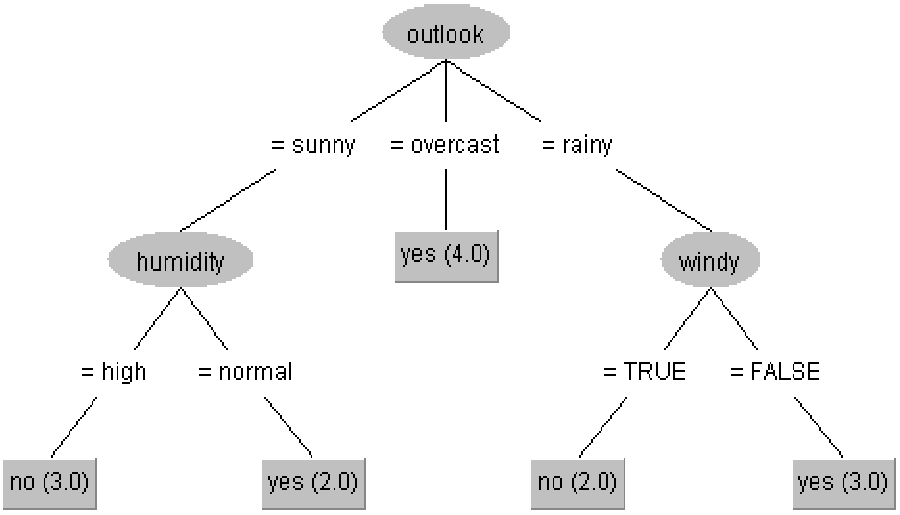

Fig.1:*-Example of Decision Tree on what to do when different situations occur in weather.*

RELATED WORK

Data mining is a research technique to find interesting patterns from hidden information in a database [5-7]. In the health sector, data mining can be used to diagnose a disease from the patient's medical data record. This research used a Chronic Kidney Disease (CKD) dataset obtained from UCI machine learning repository. The major important technique of mining classification algorithm is C4.5 which increases by applying discretion and correlation based Feature section [13].

One of the applicable areas of data mining is to search useful knowledge from the large data sources. There are different techniques used for novel knowledge discovery such as classification, clustering, association and pattern analysis etc. Data mining is playing a critical role in the recent decade for detection of fast spreading diseases. The rate of detection for pandemic diseases should be very fast with huge amount of data sources. COVID-19 is one of pandemic disease which have shaken the existing disease prediction and defending systems.

Analyzing the results obtained from experiments, Random Forest (RF) was identified to perform better compared to other algorithms [11].

Today the health industry holds hidden information essential for decision-making. For predicting heart problems, data extraction algorithms like K-star, J48, SMO, Naïve Bayes, MLP, Random Forest, Bayes Net, and REPTREE are used for this study (Weka) software. The results of the predictive accuracy, the ROC curve, and the AUC value are combined using a standard set of data and a collected dataset.

Data mining functions and techniques are used to identify the level of risk factors to help the patients in taking precautions in advance to save their life [12]. Covid-19 analysis is performed efficiently using the weka tool and different decision tree algorithms such as J4.8, support vector machine etc. and a good accuracy of result is obtained [8].

Machine learning technologies can be integrated with the IoT data for better generation of results. Artificial intelligence, machine learning techniques are also very much useful for intelligent and smart applicability on real world scenarios. In this study, authors key contribution is to study on the pandemic situation of COVID-19 for avoiding and restricting the disease spread.

For decision-makers in various real-world situations and application areas, particularly from the technical point of view, machine learning is very much useful [9]. In the epidemiological COVID-19 research, artificial intelligence is a unique approach to make predictions about disease severity to manage COVID-19 patients.

We investigated the skill of data mining and machine learning, two advanced forms of artificial intelligence, to predict severe COVID-19 pneumonia based on routine laboratory tests. Laboratory datasets analyzed by the R software and WEKA workbench. The C4.5 software, a supervised learning algorithm based on an objective-predefined variable (severity) that generated a decision tree with 89.4% precision [10].

DATASET

In this experiment we have used covid-19 dataset. This data set consists of 1,81,884 instances with 9 attributes. Each instance consists of nine attributes like cough, fever, sore throat, shortness of breath, head ache, age 60 and above, gender, test indication respectively.

Table 1:-*Corona Virus Dataset Properties*

Attributes	**Attribute values**
Cough	Numeric(1,0)
Fever	Numeric(1,0)
Sore throat	Numeric(1,0)
Shortness of breath	Numeric(1,0)
Headache	Numeric(1,0)
Age 60 and above	None, Yes, No
Gender	female, male, None
Test indication	Other, Abroad, Contact with confirmed
Corona	negative, positive

ATTRIBUTES VALUES

In the above table 1 means positive i.e., suffering from those symptoms and 0 means negative i.e., not suffering from those symptoms. In the test indication Abroad means the person traveled abroad from the last 30 days, Contact with confirmed means he is contacted by a corona positive person.3 Negative in Corona result is the person who got a negative result in the test. Positive in Corona result is the person got a positive result in the test.

RESEARCH METHODOLOGY

The Decision Tree is similar to the human decision-making process and so that it is easy to understand. It can solve in both situations whether one has discrete or continuous data as input.

Talking about the characteristics of Decision Tree, the C4.5 algorithm is simulated on the WEKA tool and the data type of the data set is only categorical. C4.5 can take continuous data set as input for simulation purpose. Table 2 presents the simulation details.

The decision tree makes explicit all possible alternatives and traces each alternative to its conclusion in a single view, to make easy comparisons among the various alternatives.

Table 2:-*Decision Tree Simulation*

Decision Tree Algorithm	Data Types	Possible Tool
C4.5	Categorical ,numerical	WEKA

RESULTS

C4.5 Tree

```
test_indication = Other
|  head_ache<= 0
|  |  shortness_of_breath<= 0
|  |  |  sore_throat<= 0: negative (157779.0/2472.0)
|  |  |  sore_throat> 0
|  |  |  |  age_60_and_above = None: negative (6.0)
|  |  |  |  age_60_and_above = Yes: positive (12.0)
|  |  |  |  age_60_and_above = No: positive (70.0/2.0)
|  |  shortness_of_breath> 0
|  |  |  age_60_and_above = None: negative (5.0/1.0)
|  |  |  age_60_and_above = Yes: positive (36.0/2.0)
|  |  |  age_60_and_above = No: positive (56.0/2.0)
|  head_ache> 0
|  |  age_60_and_above = None
|  |  |  sore_throat<= 0
|  |  |  |  fever <= 0: positive (3.0)
|  |  |  |  fever > 0: negative (7.0/2.0)
|  |  |  sore_throat> 0: negative (2.0)
|  |  age_60_and_above = Yes: positive (25.0)
|  |  age_60_and_above = No: positive (158.0/4.0)
test_indication = Abroad
|  head_ache<= 0
|  |  sore_throat<= 0: negative (18113.0/403.0)
|  |  sore_throat> 0
|  |  |  age_60_and_above = None: negative (7.0)
|  |  |  age_60_and_above = Yes: positive (3.0/1.0)
|  |  |  age_60_and_above = No: positive (21.0/8.0)
|  head_ache> 0
|  |  age_60_and_above = None: negative (2.0)
|  |  age_60_and_above = Yes
|  |  |  sore_throat<= 0: positive (6.0/1.0)
|  |  |  sore_throat> 0: negative (2.0)
|  |  age_60_and_above = No: positive (29.0/3.0)
test_indication = Contact_with_confirmed
|  head_ache<= 0
|  |  cough <= 0
|  |  |  gender = female: positive (1325.0/303.0)
|  |  |  gender = male: positive (1274.0/299.0)
|  |  |  gender = None: negative (40.0/10.0)
|  |  cough > 0
|  |  |  age_60_and_above = None
|  |  |  |  gender = female: positive (471.0/119.0)
|  |  |  |  gender = male: positive (483.0/171.0)
|  |  |  |  gender = None: negative (10.0/1.0)
|  |  |  age_60_and_above = Yes: positive (138.0/60.0)
|  |  |  age_60_and_above = No
|  |  |  |  sore_throat<= 0
|  |  |  |  |  shortness_of_breath<= 0
|  |  |  |  |  |  fever <= 0: negative (1023.0/327.0)
|  |  |  |  |  |  fever > 0
|  |  |  |  |  |  |  gender = female: positive (160.0/74.0)
|  |  |  |  |  |  |  gender = male: positive (225.0/105.0)
|  |  |  |  |  |  |  gender = None: negative (14.0)
|  |  |  |  |  shortness_of_breath> 0: positive (31.0/2.0)
|  |  |  |  sore_throat> 0: positive (57.0/3.0)
|  head_ache> 0: positive (291.0/16.0)

Number of Leaves  :   34
Size of the tree:      57
```

***Table 3:**-Overall analysis metrics*

S.no	Validation metrics	value
1.	IR_precision	0.982
2.	IR_recall	0.993
3.	F_measure	0.987
4.	Area_under_ROC	0.795
5.	True_positive_rate	0.993
6.	True_negative_rate	0.530
7.	False_negative_rate	0.007
8.	False_positive_rate	0.470

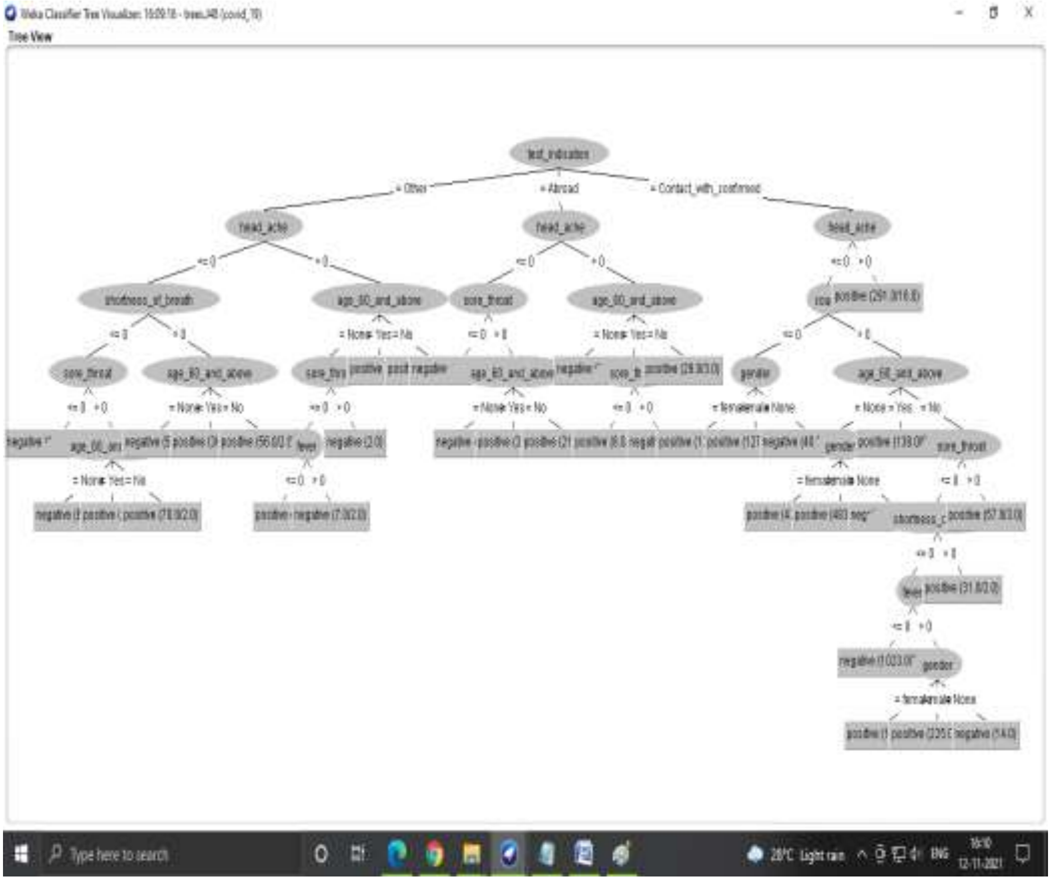

Fig.2:- Tree View

DISCUSSION

In this experiment simulation, we have used 1,81,884 instances having 9 attributes related to covid_19 data. In this experiment, we visualize tree using the C4.5 algorithm with 10 cross-validation folds in the Weka tool.

We got 34 rules to decide if the person has positive or negative and the size of the tree is 57. In this weka tool we go through an experiment and calculate IR precision, IR recall, F measure, Area under ROC, True positive rate, True negative rate, False negative rate, False positive rate.

In a task, the precision for a class is the number of true positives (i.e. the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class).

Precision is defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP}$$

In our experiment we got IR-precision as 0.982 which is pretty good.
Now going to the definition of Recall is the ratio of correctly predicted positive observations to the all observations in actual class.

$$\text{Recall} = \frac{TP}{TP + FN}$$

We have got a recall of 0.993which is good for this model as it’s above 0.5.Now moving our experiment to a single measure which is capable of projecting the metrics of both precision and recall is known as F-measure.

The reason for the inclusion of f-measure evaluation metric is as a single measure of precision or recall is not enough to give a precise picture. In some of the applications, a good value of precision is there with very bad value of recall and vice versa.

Which is a synonym of Harmonic meaning of recall and precision, F-measure provides a way to express both concerns with a single score.
F-Measure = (2 * Precision * Recall) / (Precision + Recall)

The result is a value between 0.0 for the worst F-measure and 1.0 for a perfect F-measure. Here we got 0.987 as F-measure so it is near to perfect F-measure.

An **ROC curve** (**receiver operating characteristic curve**) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

- True Positive Rate
- False Positive Rate

True Positive Rate (**TPR**) is a synonym for recall and is therefore defined as follows:

$$TPR = \frac{TP}{TP + FN}$$

False Positive Rate (**FPR**) is defined as follows:

$$FPR = \frac{FP}{FP + TN}$$

AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0. We got 0.795 Area under ROC which give at least 80% accuracy our prediction.

True Negative: the truth is negative, and the test predicts a negative. The person is not sick, and the test accurately reports this. False Negative: the truth is positive, but the test predicts a negative. The person is sick, but the test inaccurately reports that they are not also called a Type II error in statistics.

CONCLUSION

In this paper, we are taken the real world available data in online. By this data we are extracted 34 conditions to get covid-19 positive or negative from 181884 instances.

We are taken the attributes as corona symptoms like cough, fever, sore-throat, shortness-of-breath, head-ache, age-60 and above, gender, test-indication and corona-result. Here corona-result is taken as class.

Because we have taken real world data the outcome is (i.e..,34 conditions) are very much useful to say that person has some percentage chance to get corona positive or negative. Future work includes more refine set of rules for more accurate detection in presence of noise and uncertainty. Future work includes more refine set of rules for more accurate detection in presence of noise and uncertainty.

REFERENCES

1. Mahmood, A. M., & Kuppa, M. R. (2010, December). Early detection of clinical parameters in heart disease by improved decision tree algorithm. In *2010 Second Vaagdevi International Conference on Information Technology for Real World Problems* (pp. 24-29). IEEE.
2. Mahmood, A. M., & Kuppa, M. R. (2012). A novel pruning approach using expert knowledge for data-specific pruning. *Engineering with Computers*, *28*(1), 21-30.
3. Mahmood, A. M., & Kuppa, M. R. (2012). A novel pruning approach using expert knowledge for data-specific pruning. *Engineering with Computers*, *28*(1), 21-30.
4. Mahmood, A. M., & Kuppa, M. R. (2010, December). Early detection of clinical parameters in heart disease by improved decision tree algorithm. In *2010 Second Vaagdevi International Conference on Information Technology for Real World Problems* (pp. 24-29). IEEE.
5. Mahmood, A. M., Rao, K. M., & Reddi, K. K. (2010). A novel algorithm for scaling up the accuracy of decision trees. *International Journal on Computer Science and Engineering*, *2*(2), 126-131.
6. Mahmood, A. M., Kuppa, M. R., & Reddi, K. K. (2010). A New decision Tree Induction Using Composite Splitting Criterion. *Journal of Applied Computer Science & Mathematics*, (9).
7. Reddi, K. K., Mahmood, A. M., & Rao, K. M. (2010). Generating optimized decision tree based on discrete wavelet transform. *(IJEST) International Journal of Engineering Science and Technology*, *2*(3), 157-164.
8. Villavicencio, C. N., Macrohon, J. J. E., Inbaraj, X. A., Jeng, J. H., & Hsieh, J. G. (2021). COVID-19 Prediction applying supervised machine learning algorithms with comparative analysis using WEKA. *Algorithms*, *14*(7), 201.
9. Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, *2*(3), 1-21.
10. Pulgar-Sánchez, M., Chamorro, K., Fors, M., Mora, F. X., Ramírez, H., Fernandez-Moreira, E., & Ballaz, S. J. (2021). Biomarkers of severe COVID-19 pneumonia on admission using data-mining powered by common laboratory blood tests-datasets. *Computers in biology and medicine*, *136*, 104738.
11. Matta, D. M., & Saraf, M. K. (2020). Prediction of COVID-19 using machine learning techniques.

12. Sultana, M., Haider, A., & Uddin, M. S. (2016, September). Analysis of data mining techniques for heart disease prediction. In *2016 3rd international conference on electrical engineering and information communication technology (ICEEICT)* (pp. 1-5). IEEE.
13. Cahyani, N., & Muslim, M. A. (2020). Increasing Accuracy of C4. 5 Algorithm by applying discretization and correlation-based feature selection for chronic kidney disease diagnosis. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, *12*(1), 25-32.