International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 06 | June 2021 www.irjet.net p-ISSN: 2395-0072

Framework for Epidemic Disease Prediction

Jeevan J[1], Rajesh B[2], Smitha G R[3]

[1]*Student, Dept. Information Science and Engineering, RV College of Engineering, Bengaluru, Karnataka, India*
[2]*Student, Dept. Information Science and Engineering, RV College of Engineering, Bengaluru, Karnataka, India*
[3]*Assistant Professor, Dept. Information Science and Engineering, RV College of Engineering, Bengaluru, Karnataka, India*

Abstract - *Epidemic diseases are infectious diseases that can spread to large amount of people in no time. Outbreaks of disease can affect the different aspects such as are human health, economy, lack of medication resources. An infected disease will have a set of feature and factors based on which a machine learning model can be designed to predict the outcome of such diseases. In this paper we look into implementing the model that predicts whether an individual is infected or not based on the set of features and factors of a disease. Decision Tree and Random Forest are two machine learning models which are used and results are observed. Hyper-parameters are provided to the models for better results. The Observations are made on the accuracy scores with different hyper-parameters and its values.*
The Resulting model predicts whether an individual is infected or not in terms of 0 or 1.

Key Words: **Machine Learning, Decision Tree, Random Forest, Hyper-parameters, GridSearchCV.**

1.INTRODUCTION

Epidemic disease is infectious which spreads from person to person and get infected to huge amount of population. Once the disease outbreaks it spreads rapidly in the society and numerous people face the problem of health issues. These infectious diseases come with symptoms that are found in all infected individuals. The infectious disease not only harm the lives of people but it also affects different aspects like economy etc. To predict such infectious diseases a machine learning model is implemented. In this paper we discussion on the infectious disease that has occurred in recent times i,e. COVID-19.

COVID-19 emerged from Wuhan of China in December 2019. An individual infected with COVID-19 has a set of symptoms, based on such symptoms a machine learning models (Decision Tree & Random Forest [1]) are built. The model takes the symptoms as feature set and based on the feature set the model predicts whether an individual is infected or uninfected. The model is also implemented with different hyper-parameters and the results are observed.

The Paper focuses on predicting the COVID-19 for individual whether he/she is infected or uninfected and the observations are made on the model with different hyper-parameter.

2. PROPOSED SOLUTION

In this section we discussion proposed solution for Epidemic Disease Prediction and different steps to follow to do so.

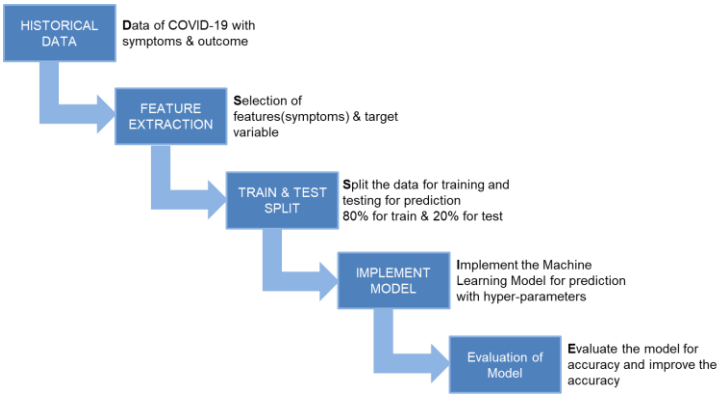

Fig -1: Proposed Methodology

Step 1: Data collection of COVID-19 with symptoms and other details of individual infected patient. The data has information like symptoms, age and status of patient (infected or uninfected).

Step 2: In this step that given data is pre-processed to eliminate the null values and clean the data. Once the data is pre-processed the required features are selected such as symptoms that a COVID-19 patient has, age etc. Along with the feature set the target variable is also identified. In this dataset the status of patient (infected or uninfected) is target variable. These feature set and target variable is visualized separately.

Step 3: The splitting of the data is done for training the model and testing the model. Since training is the important part of machine building 80 percent of data is split for training and only 20 percent of data is utilized for testing the model.

Step 4: Once the data is split into the training and testing data. The model is implemented with training dataset. Decision Tree and Random Forest are two machine learning model implemented with different hyper-parameters such as criterion, max_depth, max_leaf_nodes, min_samples_split, splitter etc.

Step 5: Once the model is implemented and built successfully, feed the testing data to the model and obtain prediction values. Evaluate the model using the metrics provided in the scikit learn library and get the accuracy of the model.

3. MACHINE LEARNING ALGORITHMS

Machine learning is a field where the computers are not programmed to do things but are enabled to learn from the given dataset, analyze the given data and predict the outcome based on the input dataset.
In this paper, the two supervised machine learning algorithms Decision Tree and Random Forest Classifiers are used to classify the given COVID-19 dataset and predict the outcome.
Decision Tree Classifier: Decision Tree Classifier is a supervised machine learning algorithm. It is uses tree like structure to predict the outcome as shown in the below **Fig - 2**. The internal nodes are tested on feature, and the leaf nodes are outcome of class label. The dataset is divided into two parts X and Y, where X is n number of features and Y is the target variable (class label).

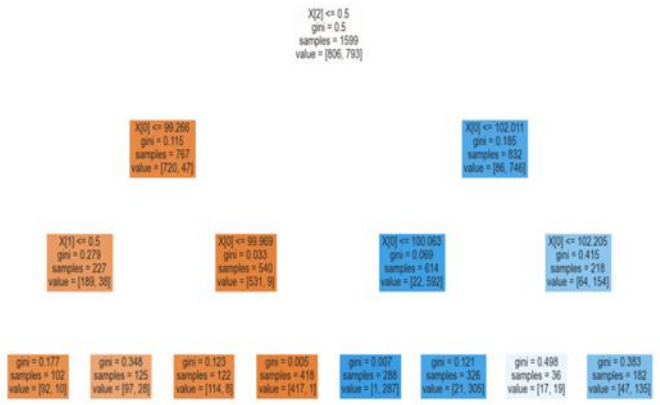

Fig -2: Decision Tree on given feature set

Random Forest Classifier: Random Forest Classifier is a supervised machine learning algorithm. Random Forest uses the ensemble learning technique, where it make use of multiple decision trees and the outcome of each decision tree is considered. The final result is processed by obtaining the outcome of each tree, based on majority of the voting of outcomes.

GridSearchCV: GridSearchCV is used to estimate all the model with different combinations of parameters one by one and select the best one out all models. An array or list of parameters is passed to GridSearchCV where all the list of parameters is evaluated by cross validation and the best model is given as output.

Hyper-parameters: Hyper-parameters are the parameter values feed to the machine learning algorithms, these parameter values control the learning process. Lets look into some of the hyper-parameters.
criterion: This parameter function measures the quality of split. It has two values gini for the gini impurity and entropy for the information gain.

max_depth: It specifies the maximum depth of the tree. If none specified then it expands until all leaves are pure or until all leaves contain less than min_samples_split samples. It takes a number as parameter

min_samples_split: It specifies the minimum samples required to split the internal node. It takes a number as a parameter and must be greater than 1.

Splitter: Strategy used to split at each node, it has two values best for best split and random for random split.

4. RESULTS AND DISCUSSION

The Epidemic Disease Prediction will give the results based on the input feature set. The outcome of the model is infected or uninfected in terms of 0 or 1. The given input to the model is symptoms and age of the patient and based on the input the outcome i,e. 0 or 1 is obtained. Before looking into the predicted values, the accuracy rates of both the Decision Tree and Random Forest without hyper-parameters are shown in the below **Table-1**.

Table -1: Accuracy rates of Decision Tree and Random Forest without hyper-parameters

Model	Accuracy
Decision Tree	85.25
Random Forest	88.75

The data is split into training data and testing data, 80 percent of data is used for training the machine learning model and 20 percent of data is used for testing machine learning model. The machine learning model is trained and built using the training data.

Once the model is built it is feed the with the test data for obtaining the predicted values. The predicted values are used to evaluate the model for the accuracy scores.

The actual and predicted values of Decision Tree and Random Forest are shown in the below **Fig -3**.

	Actual	Random Forest	Decision Tree		Actual	Random Forest	Decision Tree
0	0	0	0	385	1	1	1
1	0	0	0	386	1	1	1
2	0	0	0	387	1	1	1
3	0	0	0	388	1	1	1
4	0	0	0	389	1	1	1
5	1	1	0	390	0	0	0
6	1	1	1	391	0	0	0
7	0	0	0	392	1	1	1
8	1	1	1	393	0	0	0
9	1	1	1	394	0	0	0
10	0	0	1	395	1	1	1
11	1	0	0	396	0	0	0
12	1	1	1	397	0	0	0
13	0	0	1	398	1	1	1
14	0	0	0	399	0	0	1

Fig -3: Actual v Random Forest Predict v Decision Tree Predict

Let's look into hyper-parameters of Decision Tree and its accuracy tuning with changing of parameters. The Decision Tree Classifier is implemented with different hyper-parameters such as criterion, max_depth, max_leaf_nodes, min_samples_split and splitter and the accuracy rates are observed on changing of parameters and its values. The actual and predicted values of Decision Tree with hyper-parameters are shown in the below **Fig -4**.

	criterion	max_depth	splitter	min_samples_leaf	max_leaf_nodes	accuracy
0	gini	none	none	none	none	83.75
1	entropy	none	none	none	none	84.25
2	gini	10	random	2	30	89.25
3	entropy	10	random	2	30	90.75
4	gini	10	best	2	30	89.5
5	entropy	10	best	2	30	89.5
6	gini	10	best	2	none	86.5
7	entropy	10	best	2	none	87.25
8	gini	10	random	2	none	86.25
9	entropy	10	random	2	none	85.25
10	gini	20	random	4	50	90.5
11	entropy	20	random	4	50	89.5

Fig -4: Accuracy rates of Decision Tree Classifier with Different Hyper-parameters and its values

As shown in the above **Fig-4** the accuracy of the model varies when the hyper-parameters and its value gets changed. By this we can observe the varying accuracy rates by tuning hyper-parameters and select the best one with maximum accuracy rate and use that model for prediction of the outcome.

Random Forest is implement using the GridSearchCV and the model with best hyper-parameters is obtained. The results of best hyper-parameters for random forest is shown in the below **Fig-5**.

```
CV_rf.best_params_
```

```
{'criterion': 'gini',
 'max_depth': 3,
 'min_samples_split': 2,
 'n_estimators': 200}
```

Fig -5: GridSearchCV Results for Best Hyper-parameters to Build Random Forest Classifier

5. CONCLUSION

The spread of infectious disease gets bigger and bigger, to determine an individual infected with the infectious disease a machine learning model is trained and implemented. The symptoms data is collected cleaned and is feed to a machine learning model. This model has features set and a target variable. The feature set include different symptoms of a infectious disease, and the target is the status of an individual(infected or uninfected) in terms of 0 or 1. This model is built using two classifier algorithms one is Decision Tree and another is Random Forest. Not only the model is implemented for Epidemic Disease Prediction, but also the model is evaluated with various hyper-parameters and the accuracy rates of model is observed with different hyper-parameters and its values.

Hence the resulting model outcome will be the status of the patient whether an individual is infected or uninfected in terms of 0 or 1.

REFERENCES

[1] A. Ajith, K. Manoj, H. Kiran, P. J. Pillai and J. J. Nair, "A Study on Prediction and Spreading of Epidemic Diseases," 2020 International Conference on Communication and Signal Processing (ICCSP), 2020, pp. 1265-1268, doi: 10.1109/ICCSP48568.2020.9182147.

[2] Y. Zhang, W. K. Cheung and J. Liu, "A Unified Framework for Epidemic Prediction based on Poisson Regression," in IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 11, pp. 2878-2892, 1 Nov. 2015, doi: 10.1109/TKDE.2015.2436918.

[3] A. Zamiri, H. S. Yazdi and S. A. Goli, "Temporal and Spatial Monitoring and Prediction of Epidemic Outbreaks," in IEEE Journal of Biomedical and Health Informatics, vol. 19, no. 2, pp. 735-744, March 2015, doi: 10.1109/JBHI.2014.2338213.

[4] M. Kumar, J. Bareja, M. Singh and R. Sharma, "An Intelligent Prediction Model of COVID-19 in India using Hybrid Epidemic Model," 2020 International Conference on Smart Electronics and Communication (ICOSEC), 2020, pp. 389-396, doi: 10.1109/ICOSEC49089.2020.9215426.