Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 7602 - 7607
Received 16 February 2021; Accepted 08 March 2021.

Implementation of Machine Learning Algorithms for Predictive Analysis about Mortality for Covid-19 Data Sets

Shagufta Praveen[1],Dr. Mazhar Afzal[2],Aamir Khan[3]
Department of Computer Science
Saharanpur,India
Glocal University
shaguftasaeed125@gmail.com

Abstract: Covid-19, a viral global pandemic, gave number of question marks for various machine learningbased industries.Collecting data and related solutions for infected crowd worldwide is a biggest challenge today. This article represents contribution of machine learning to fight against COVID-19. Itreflects various factors and other related work that could determine the help of machine learning in this current scenario. In this article decision tree and linear regression, supervised learning algorithms of machine learning are used to calculate various predictions with the help of attributes of collected data. Python 3.8.0 is used for the implementation and model functions are used for data validation, prediction and accuracy**.**

Keywords: decision tree, python , COVID-19,linear regression

INTRODUCTION

Pharmaceuticals and related medicinal departments are unable to provide successful result for this terrible situation today. World Economy dropped, several deaths not only due to disease but also because of hunger and they are relatively increasing every day. Machine Learning[1], the most promising tool from decades giving us some hope for various unseen predictionabout the current scenario of world and different zones. There are thousands of projects and models developing and even developed across the world that mainly based on ML and its algorithm.During drug creation AI again found to be one of the fastest ways to collect right drug. Yes, there are many challenges but there is a big hope that data science and machine learning will contribute something to the world so that current situation could be overcome.

MACHINE LEARNING AND COVD-19

Regarding help, machine learning works on supervised[2] and unsupervised learning methods[3] where it detects through regression[4], classification[5] and clustering[6]. Keeping all this in mind, regression can help us in various predictions like predictionof number of figure of patients, to identify risk and to predict next pandemic. Classification can help us to identify the safe and unsafepatient, symptomatic and asymptomatic patient etc. And Clustering can tell us to know the red zone areas and many morepartitions of group on different basis.The most abrupt part of the scenario is every zone has different people with different physiology where their affects, symptoms and many more related essential things varies. But most of the important factors that are involved is age, hygiene habits, human interaction,location and climate.

Many techniques were proposed by researchers among them one is classification in machine learning. Classification has innumerable application that includes medical diagnosis. In this article, decision tree classifier is being used to discuss the future prediction of COVID -19[7] with help of various attributes(like age and gender).Classification is a dual step process where in first step model is constructed based on previous data and later model's accuracy is determined [8].At times age and location is not enough factors like strength of the immunity, history of that particular patients(diabetes, heart illness) also various factors that can be the cause of the corona and especially to death.The Indian Express news paper recently published about the occurrences of death of age groups 40-64 or 35-64 that are found in more than 50% of patients suffering from covid-19[9]

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 7602 - 7607
Received 16 February 2021; Accepted 08 March 2021.

LITERATURE REVIEW

There are several researchers who included the decision tree as a method in their reseach work for the diagnosis of several diseases and for their related attributes. In 2008 decision tree algorithm was used to predict early phases of illness for dengue disease[10]. In 2015, Decision trees was used to detect the prediction percentage of heart disease for various patients on the basis of some attributes[11]In 2015, some korean researchers research regarding prediction of data-mining based coronary heart disease risk prediction model using fuzzy logic and decision tree[12].In 2016, a clinical decision tree to predict whether a bacteremia Patient Is Infected With an Extended-Spectrum β-Lactamase–Producing Organism[13],In 2018, decision tree was used with neural network classifiers for prediction of heart disease[14].In 2020, prediction over mortality rate has been done withartificial intelligence over covid-19 data set[15].Prediction and data collectionData mining[8] is used for different perception and to summarize useful information for analysis purpose.Prediction in ML is done through various algorithm like regression , decision tree, naïve bayes and many more.Data is gathered, collected,cleaned, processed in models and their output is analyzed so that better prediction can be achieved.

Data for this article is collected from various sources.Data is associated with corona disease where data is collected from various web sources and excels files. These files has all related data regarding death and hospitalized people in respect of days.In regarding consistency of the outcome,missing values are removed and redundant data is also eliminated.

ALGORITHM

Algorithm that is used for prediction of deaths caused by covid 19 can be done through a approach in which decision trees are constructed and follow a top-down approachwhich involves training set of tuples and associated class labels.

Algorithm plays around few parameters like attribute list, attributed selection method that helps in procedure for selecting best attribute and can discriminate wisely with the help of certain criterion.Those criterian are entropy, information gain or gini index etc,
This criterian helps in the best splitting of the tree. Generally tree is split in binary form with the help of raw data.This article follows the criteria: Gini index[8] to evaluate the purity and impurity of the node in a tree.
$Gini(D)=1-\left(\sum_{i=1}^{m} x_i^2\right)$
X_i is the probability that a tuple in D belong to class C_i and is estimated by $C_{i/d}/D$
Here decision tree implemented by python script

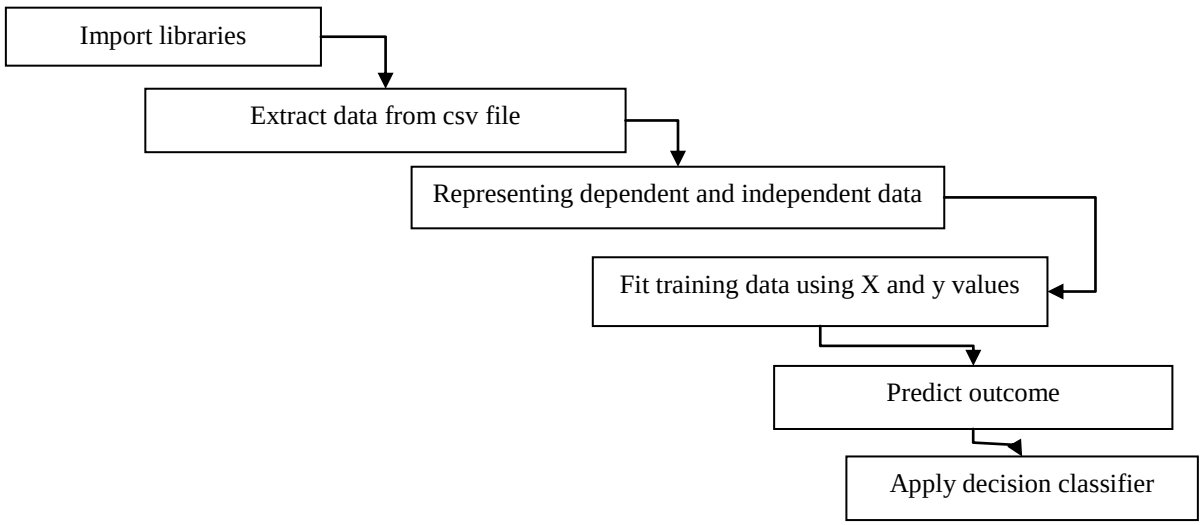

Fig 1. Flow chart for decision tree

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 7602 - 7607
Received 16 February 2021; Accepted 08 March 2021.

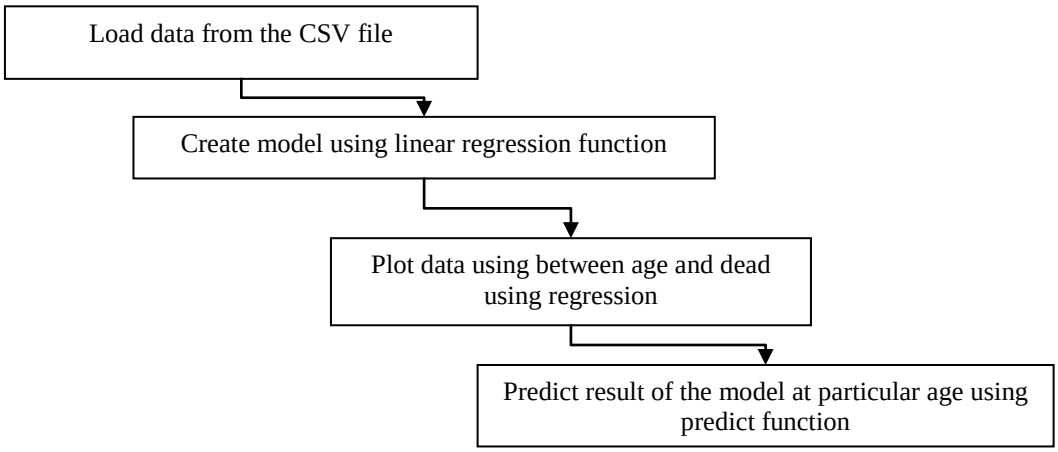

Fig 2.Flow chart for linear regression

Output:digraph Tree {
node [shape=box, style="filled", color="black"] ;
0 [label="age <= 59.5\ngini = 0.123\nsamples = 759\nvalue = [50, 709]", fillcolor="#47a4e7"] ;
1 [label="age <= 43.5\ngini = 0.063\nsamples = 671\nvalue = [22, 649]", fillcolor="#40a0e6"] ;
0 -> 1 [labeldistance=2.5, labelangle=45, headlabel="True"] ;
3 [label="gini = 0.034\nsamples = 523\nvalue = [9, 514]", fillcolor="#3c9fe5"] ;
1 -> 3 ;
4 [label="age <= 52.5\ngini = 0.16\nsamples = 148\nvalue = [13, 135]", fillcolor="#4ca6e8"] ;
1 -> 4 ;
13 [label="age <= 48.5\ngini = 0.12\nsamples = 94\nvalue = [6, 88]", fillcolor="#46a4e7"] ;
4 -> 13 ;
15 [label="gini = 0.198\nsamples = 54\nvalue = [6, 48]", fillcolor="#52a9e8"] ;
13 -> 15 ;
16 [label="gini = 0.0\nsamples = 40\nvalue = [0, 40]", fillcolor="#399de5"] ;
13 -> 16 ;
14 [label="age <= 57.5\ngini = 0.226\nsamples = 54\nvalue = [7, 47]", fillcolor="#56ace9"] ;
4 -> 14 ;
17 [label="Gender <= 0.5\ngini = 0.263\nsamples = 45\nvalue = [7, 38]", fillcolor="#5dafea"] ;
14 -> 17 ;
19 [label="age <= 53.5\ngini = 0.36\nsamples = 17\nvalue = [4, 13]", fillcolor="#76bbed"] ;
17 -> 19 ;
21 [label="(...)", fillcolor="#C0C0C0"] ;
19 -> 21 ;
22 [label="(...)", fillcolor="#C0C0C0"] ;
19 -> 22 ;
20 [label="gini = 0.191\nsamples = 28\nvalue = [3, 25]", fillcolor="#51a9e8"] ;
17 -> 20 ;
18 [label="gini = 0.0\nsamples = 9\nvalue = [0, 9]", fillcolor="#399de5"] ;
14 -> 18 ;
2 [label="age <= 80.5\ngini = 0.434\nsamples = 88\nvalue = [28, 60]", fillcolor="#95cbf1"] ;
0 -> 2 [labeldistance=2.5, labelangle=-45, headlabel="False"] ;
5 [label="age <= 65.5\ngini = 0.442\nsamples = 85\nvalue = [28, 57]", fillcolor="#9acdf2"] ;
2 -> 5 ;
7 [label="age <= 62.5\ngini = 0.408\nsamples = 49\nvalue = [14, 35]", fillcolor="#88c4ef"] ;
5 -> 7 ;
37 [label="Gender <= 0.5\ngini = 0.444\nsamples = 24\nvalue = [8, 16]", fillcolor="#9ccef2"] ;
7 -> 37 ;

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 7602 - 7607
Received 16 February 2021; Accepted 08 March 2021.

```
39 [label="age <= 61.0\ngini = 0.5\nsamples = 10\nvalue = [5, 5]", fillcolor="#ffffff"] ;
37 -> 39 ;
41 [label="(...)", fillcolor="#C0C0C0"] ;
39 -> 41 ;
42 [label="(...)", fillcolor="#C0C0C0"] ;-
39 -> 42 ;
40 [label="gini = 0.337\nsamples = 14\nvalue = [3, 11]", fillcolor="#6fb8ec"] ;
37 -> 40 ;
38 [label="Gender <= 0.5\ngini = 0.365\nsamples = 25\nvalue = [6, 19]", fillcolor="#78bced"] ;
7 -> 38 ;
43 [label="age <= 63.5\ngini = 0.18\nsamples = 10\nvalue = [1, 9]", fillcolor="#4fa8e8"] ;
38 -> 43 ;
47 [label="(...)", fillcolor="#C0C0C0"] ;
43 -> 47 ;
48 [label="(...)", fillcolor="#C0C0C0"] ;
43 -> 48 ;
44 [label="age <= 63.5\ngini = 0.444\nsamples = 15\nvalue = [5, 10]", fillcolor="#9ccef2"] ;
38 -> 44 ;
45 [label="(...)", fillcolor="#C0C0C0"] ;
44 -> 45 ;
46 [label="(...)", fillcolor="#C0C0C0"] ;
44 -> 46 ;
8 [label="age <= 67.5\ngini = 0.475\nsamples = 36\nvalue = [14, 22]", fillcolor="#b7dbf6"] ;
5 -> 8 ;
9 [label="Gender <= 0.5\ngini = 0.444\nsamples = 6\nvalue = [4, 2]", fillcolor="#f2c09c"] ;
8 -> 9 ;
11 [label="gini = 0.32\nsamples = 5\nvalue = [4, 1]", fillcolor="#eca06a"] ;
9 -> 11 ;
12 [label="gini = 0.0\nsamples = 1\nvalue = [0, 1]", fillcolor="#399de5"] ;
9 -> 12 ;
10 [label="age <= 78.0\ngini = 0.444\nsamples = 30\nvalue = [10, 20]", fillcolor="#9ccef2"] ;
8 -> 10 ;
23 [label="age <= 72.5\ngini = 0.426\nsamples = 26\nvalue = [8, 18]", fillcolor="#91c9f1"] ;
10 -> 23 ;
27 [label="(...)", fillcolor="#C0C0C0"] ;
23 -> 27 ;
28 [label="(...)", fillcolor="#C0C0C0"] ;
23 -> 28 ;
24 [label="Gender <= 0.5\ngini = 0.5\nsamples = 4\nvalue = [2, 2]", fillcolor="#ffffff"] ;
10 -> 24 ;
25 [label="(...)", fillcolor="#C0C0C0"] ;
24 -> 25 ;
26 [label="(...)", fillcolor="#C0C0C0"] ;
24 -> 26 ;
6 [label="gini = 0.0\nsamples = 3\nvalue = [0, 3]", fillcolor="#399de5"] ;
2 -> 6 ;
}
```

PATTERN EVALUATION

The output of the decision tree is analyzed with the help of some criterian, these criteria could be entopy, gini or gain information. Gini criteria helps to calculate homogeneous, pure and impure nodes.The less impurity helps in nodes splitting.Value of different criteria reveal about the node splitting with their increasing and decreasing value.Here in the article data is evaluated with decision tree where gini is used as criteria. In this, data collected and showcased with the help of table(collected

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 7602 - 7607
Received 16 February 2021; Accepted 08 March 2021.

data has information about age and gender of infected people).Below decision tree is achieved on the basis of collected data where age is used to represent yes or no and gini critera helps in node splitting.

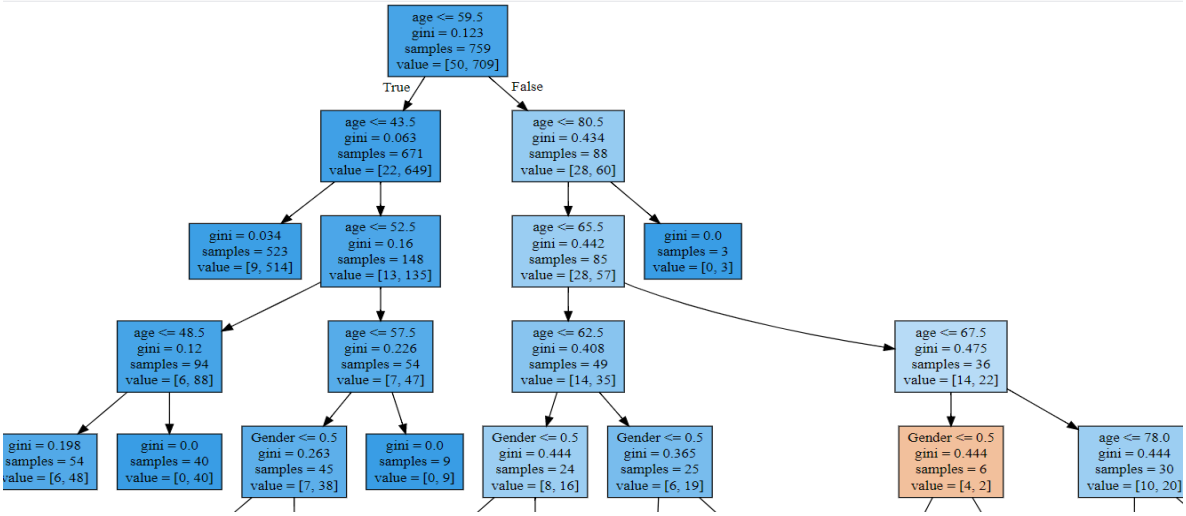

Fig 3. Decision tree

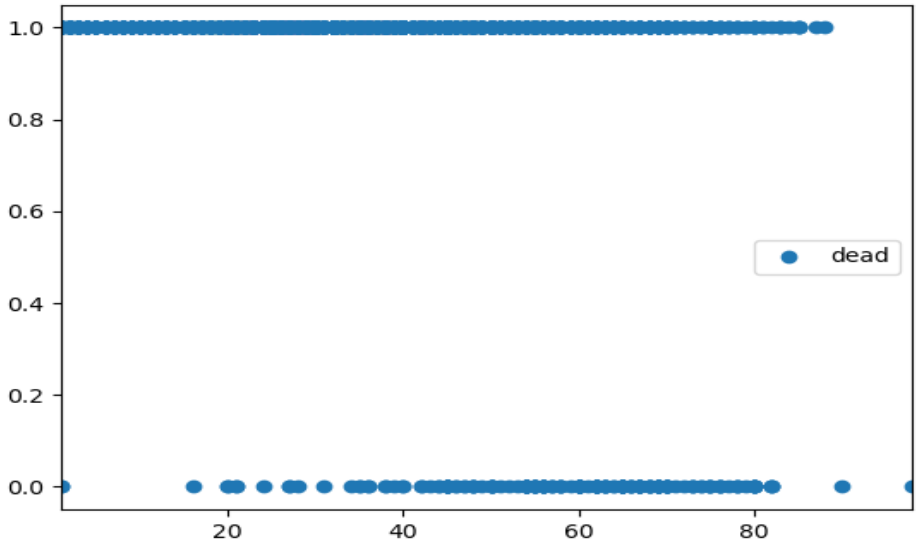

Fig 4. Linear regression graph

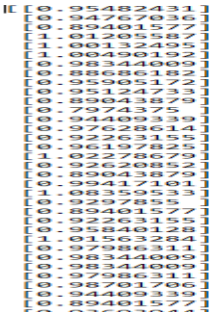

Fig 5.Predict values for every value of the excel

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 7602 - 7607
Received 16 February 2021; Accepted 08 March 2021.

COMPARATIVE STUDY RESULT AND DISCUSSION

In order to calculate prediction of y variable that express about deaths(corona patients) in respect to their age and gender. On the basis of collected data(From WHO,South-Indiahospital Data),decision tree revealed that people of age between 43 and 57 facing more deaths than other aged people(all corona infected and their health history was not revealed in web sources).Regarding gender, females of age less than 53.5 faced more deaths than males whereas females between between 61 and 67 faced less deaths. For accuracy we also tried to do linear regression using same data set where compute a relation between both x and y variable and tried to plot regression values,we achieved class of data: data.frame,dimension:1013,4, coeffecient:[[-0.00357697]],interecept:[1.0871723] and tried to do prediction using prediction function for relation model but result was not found accurate and perfect like decision tree.

REFERENCES

[1] Shalev-shawrtz S., ben-david, S.,"Understanding Machine Learning from theory to AlgorithmsNew york,USA ,Cambridge University Press, "2014.

[2] Online:https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d

[3] **Celebi,** M. Emre, **Aydin**, Kemal (Eds.),"Undupervised learning methods"**,**springer, Computer Science Department, Conway, Arkansas, USA,2016

[4] Rong,S., Bao-wen ,Z.,,"The research of regression model in machine learning field", MATEC web of conference 176,01033,IFID,2018

[5] Online:https://intellipaat.com/blog/tutorial/machine-learning-tutorial/classification-machine-learning/

[6] Jain, A.K., &Dubes, R.C*., "*Algorithm for Clustering Dat"a, Prentice-Hall, 1988

[7] Online:https://www.who.int/publications/i/item/infection-prevention-and-control-during-health-care-when-novel-coronavirus-(ncov)-infection-is-suspected-20200125

[8] Han, J., kamber, M., Pie, J., Kaufmann, M., ,"Data mining and concept techniques" ,Elsevier, Wyman Street, Waltham, MA 02451, USA,2012

[9] Online:https://www.newindianexpress.com/nation/2020/jun/19/most-covid-19-victims-in-india-belong-to-most-productive-age-study-2158680.html

[10] Tanner L, Schreiber M, Low JGH, Ong A, Tolfvenstam T, Lai YL, et al. "Decision Tree Algorithms Predict the Diagnosis and Outcome of Dengue Fever in the Early Phase of Illness",PLoSNegl Trop Dis.;2(3):e196. Published 2008 Mar 12.

[11] Purushottam, K. Saxena and R. Sharma,. "Efficient heart disease prediction system using decision tree," International Conference on Computing, Communication & Automation, Noida, pp. 72-77,2015

[12] Kim MS, J., Lee, J., Lee, Y, et al.(2015),"Data mining based coronary heart disease risk prediction model using fuzzy logic and decision tree", ,Healthc Inform Res. 2015 Jul;21(3):167-174.

[13] Goodman KE, Lessler J, Cosgrove SE, et al." A Clinical Decision Tree to Predict Whether a Bacteremic Patient Is Infected With an Extended-Spectrum β-Lactamase-Producing Organism". Clin Infect Dis. 2016;63(7):896-903.

[14] Mathan, K., Kumar, P.M., Panchatcharam, P. et al. "A novel Gini index decision tree data mining method with neural network classifiers for prediction of heart disease". Des Autom Embed Syst22, 225–242 (2018).

[15] Pourhomayoun M, ShakibiM. "Predicting Mortality Risk in Patients with COVID-19 Using Artificial Intelligence to Help Medical Decision-Making",medRxiv,2020