International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT21711

A Machine Learning Approach to Analyse the Symptoms of Covid-19 for the Initial Diagnosis of a Patient

K. Anurag Reddy[1], Susant Kumar Rath[1], Advin Manhar[2]
[1]Student, Amity University Chhattisgarh, Raipur, Chhattisgarh, India
[2]Assistant Professor, Amity University Chhattisgarh, Raipur, Chhattisgarh, India

Article Info
Volume 7, Issue 1
Page Number: 264-270
Publication Issue :
January-February-2021

ABSTRACT

The recent outbreak of the respiratory ailment COVID-19 caused by novel corona virus SARS- Cov2 is a severe and urgent global concern. In the absence of vaccine, and also treatment of COVID- 19 WHO (World Health Organization) had informed that Social distancing is the only way to avoid this pandemic and also made clear that Prevention is better than Cure. The main containment strategy is to reduce the contagion by the isolation of affected individuals. Earlier stage this pandemic was declared as a sort of Pneumonia where an individual gets affected by cold, fever and headache. Later, some new symptoms are seen in affected people like sore throat, breathing problems, and sometimes constipation. To make rapid decisions on treatment, and isolation needs, it would be useful to determine which symptoms presented by suspected infection cases are the best predictors of a positive diagnosis. This can be done by analyzing patient's symptoms and its outcome. Here, we developed a model that employed supervised machine learning algorithms to identify the certain features predicting COVID-19 disease diagnosis with high accuracy. Features examined includes details of the concerned individual, e.g., age, gender, observation of fever, breathing difficulty, and clinical details such as the severity of cough and incidence of lung infection and congestion. We had implemented some Machine Learning techniques with algorithms and found out the highest accuracy more than (50 %) of individual patient for all age groups. The following data is collected from COVID-19 positive patients, online survey and social survey done at testing centres. After that we had applied various methods as Data Preprocessing, Model Validation and Statistical analysis, etc. The probability and accuracy of a patient is shown in using various methods of Machine learning algorithm for a better understanding.

Keywords : COVID-19, World Health Organization, Machine Learning, SARS-Cov-2, Coronavirus, Machine Learning, Early Stage Symptom

Article History
Accepted : 01 Jan 2021
Published : 04 Jan 2021

I. INTRODUCTION

Recently there has been a rapid spread of the novel SARS-CoV2 coronavirus all over world. World Health Organization (WHO) has declared the pandemic as COVID-19 virus. It had caused more than three million cases and 1764794 deaths across the world as per WHO statistics of 20 December 2020.The first human coronaviruses, 229E and OC43, were identified during the 1960s from human nasal secretion. In the beginning stages coronavirus infections were viewed as giving rise to innocuous respiratory human conditions which were not fatal. But later on, development of serious and deadly respiratory disorders attributed to beta form of coronavirus with the severe acute respiratory syndrome (SARS) and the middle east respiratory syndrome (MERS). Interesting thing is the SARS-CoV infections was first arose in Foshan, China way back in 2002 and MERS-CoV in 2012 in Saudi Arabia (Zhavoronkov), both causing alarm and containment efforts due to their rapid spread and high mortality rates and also emergency all over nation and also neighboring countries. Both SARS and MERS had caused almost 9.7% and 35.8% mortality rates respectively among the diagnosed patients.

These identified coronaviruses causing a significant threat to human health and has potential to cause extreme and lethal respiratory tract infections in human, animals particularly if infection occurs and can be transferred easily. The development and spread of coronavirus had spread vastly and outpaced the rate of treatment through any type of cure. However, after affecting any individual priority is given to find the infected and isolate to rest of the people to treat them. From the earliest published information had analyzed that affected 262 individuals confirmed COVID-19 by infection of respiratory and respiratory tract routes in Beijing, China.But in February 2020 the virus outbreaked and caused almost 1.5% of Wuhan, China and caused severe problem. However, accurate global estimates are far more challenging due to the different response of spread from country to country. Like, in Italy during March 2020, it showed a case where fatality rate of 8.3%. This may reflect the demographic differences between nations, with 25% of the Italian population being over 64. However, even when stratified by age, infection rates remain higher in Italians over 70 years of age when compared to China.

In this study, we developed a machine learning technique to identify the most important and significant clinical symptoms that will predict true COVID-19 positive cases.

II. MATERIALS & METHODS

2.1) DATA COLLECTION

We had collected raw data from hospital through GitHub repository, this record had helped us about an individual patient's when admitted or getting diagnose in hospitals or in clinics for treatment. In our data we got information from certain parts of China like Anhui, Henan, Jiangsu, Shanxi and also from Zhejiang. The raw data which was available for us through Github repository was in the form of Mandarin Chinese, Which was later translated by Google Translator and by other means to get its accuracy.

As we didn't get enough data from the above-mentioned repository, so we had consulted local testing centres for data though as everyone was able to give some data but it was not sufficient then we

went to nearby hospital and requested them for available data, which made our raw data complete.

2.2 DATA PREPROCESSING

In Machine Learning process, Data Preprocessing is the step in which the data gets transformed, or Encoded, to bring it to a particular form that the machine can easily understand. Simply, the features of the data can now be easily decoded by our algorithm.
The original Chinese datasets which was available for us from above mentioned repository did not include much information about which patients were suspected positive and which were confirmed patients. The definition of a suspected case are the persons who develop symptoms and had communication with confirmed COVID-19 but didn't confirm as COVID-19 after diagnosis.

Moreover, confirmed cases defined as, the patients who are confirmed as positive for COVID-19 in the CDC test report or the doctors mentioned cases after diagnosis.The data contained patient symptoms in different(text) format.So, we applied various process to arrange and decode to make raw data into matching features individuals with respect to the columns as age, congestion, lotas(Loss of taste and smell)etc., The dataset consists of 10, 000 data where each and every data is divided according to the feature and sorted mannerly.Out of that data almost 4908 are suspicious to be positive for Covid- 19 although all 4908 are not having same or common symptoms each one are having some common symptoms but not all.

Even though we collected data from hospitals, centres and also repository we too faced missing value issues almost 4.3% of data was missing regarding their symptoms or by temperature. So, we managed to fill those data by taking the data's average value into consideration.

2.3 METHODS

Now comes the most difficult task, that to identify that which method will be suitable for finding best accuracy or showing suitable result for our data.

2.3.1 DECISION TREE

Decision Tree algorithms can be utilized to optimize both classification regression and data regression. It utilizes tree representation in which each leaf node attributes to a group of data and a branch corresponds to a value.
The main idea of using Decision tree algorithm is to build a tree for entire data and process a unique output for each and every leaf. An attribute at a node with more information gain can split the trained data to improve classification accuracy.

By using Decision tree of the given data, we can see that accuracy is almost more than 50%

2.3.2) SUPPORT VECTOR MACHINE

Support vector machine is another simple Machine learning algorithm that every ML expert should have in his arsenal. SVM is highly preferred by many Machine Learning experts as it produces significant accuracy with less computation power. Support Vector Machine, known as SVM can be used for both regression and classification tasks.

But, SVM is mainly used in classification techniques for better results.

The main objective of the SVM is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points.

By using Support Vector Machine (SVM) of the given data we can see that accuracy is more than (50%).

2.3.3 LOGISTIC REGRESSION

Logistic Regression is an alogorithm which is used for solving classification problems, it is a predictive analysis algorithm and completely based on the concept of probability. The only difference between Linear regression and Logistic regression is that Logistic Regression uses more complex cost function, this cost function also be known as the 'Sigmoid function' or as the 'logistic function' instead of a linear function. The hypothesis of LogReg tends to limit the cost function between 0 and 1.

By using Logistic Regression of the given data we can see that accuracy is almost more than (50%).

2.4. EVALUATION CRITERIA

There are different types of assessment parameters in our approach, Like precision, recall, F1-score, Log loss, and area under the ROC curve (AUC).
These methods are used to estimate our prediction accuracy.

Precision: Precision is the ratio of truely predicted positive observations to the total predicted positive observations. So it depends on True Positive (TP) and False Positive (FP) values.

Precision = TP/(TP+FP)

Recall: Recall is the ratio of truely predicted positive observations to the all observations in actual class. True Positive (TP) and False negative (FN) values are used to measure recall.

Recall = TP/(TP+FN)

F1 Score: F1 Score is known as weighted average of Precision and Recall. Therefore, this F1 score will take both false positives and false negatives into account. It is slightly complex compared to accuracy but more useful than it, if we have any unusual class distribution.

F1 Score = 2 x [(Precision x Recall) / (Precision + Recall)]

3) STATISTICAL ANALYSIS
Importing Dataset:
To use any type of algorithm we need a dataset, for that we used Google Colab as our platform. Codes are shown below:

```
from google.colab import files
uploaded = files.upload()
```

```
import pandas as pd
df=pd.read_csv('dataset.csv')
```

After importing Dataset, the data is cleaned and processed which is called as Data preprocessing.
After successfully importing data we need to split our data into training and testing datasaet.

We have used two methods for splitting the dataset

1. We had created our own method where we had split our data. Codes are shown below:
2. By using predefined method present in SKLEARN library. Codes are shown below

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test=train_test_split(X,y,test_size=0.3,random_state=143)
```

```
X_train
X_test
y_train
y_test
```

```
array([0, 0, 0, ..., 0, 0, 0])
```

```
X_train.shape
y_train.shape
X_test.shape
y_test.shape
```

```
(3000,)
```

After splitting our data, now we are using Machine learning algorithms i.e., Logistic Regression, Decision trees and Support Vector Machine.

III. LOGISTIC REGRESSION

As we discussed above about LogReg, here are some codes, After splitting the data into 70% for training and 30% for testing, we used X_train for training the features like bodypain, Lotas, fever etc. and y_train for target values such as Infection Probability. In 30% testing data we will be testing the above trained model where X_test will be testing features and y_test will target value as Infection Probability. Training data is always fit into created model and the testing data is tested on it.

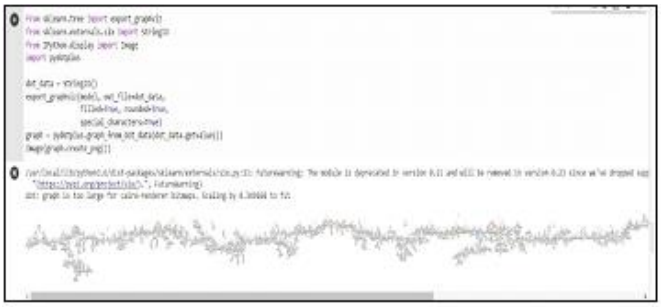
```
from sklearn.linear_model import LogisticRegression
logreg= LogisticRegression(random_state=1)

logreg.fit(X_train,y_train)

LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='auto', n_jobs=None, penalty='l2',
                   random_state=1, solver='lbfgs', tol=0.0001, verbose=0,
                   warm_start=False)

y_pred=logreg.predict(X_test)

y_pred

array([1, 0, 0, ..., 0, 0, 1])

y_pred=logreg.predict_proba(X_test)

y_pred
```

By using this method, we got accuracy more than 50%.

IV. DECISION TREES

```
from sklearn.tree import DecisionTreeClassifier

model = DecisionTreeClassifier(random_state = 1)

model.fit(X_train, y_train)

DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='gini',
                       max_depth=None, max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort='deprecated',
                       random_state=1, splitter='best')

y_pred = model.predict(X_test)
```

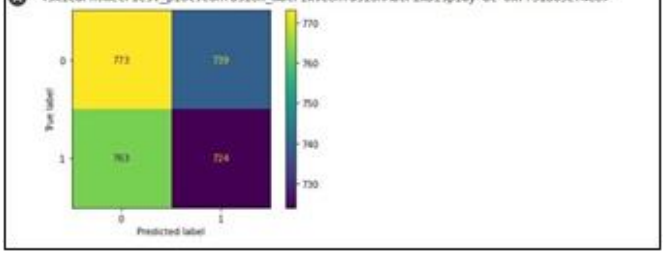
By using Decision trees, we had created a tree model where each leaf node represents a class label and branches represent conjunctions of features that lead to those class labels. The path from root to leaf node represent classification rules.



The accuracy is nothing but ratio of number of correct predictions to that of total number of input samples. Where we had achieved almost same accuracy as Logistic Regression i.e., 50%.

The matrix compares the target values with those predicted by the machine learning model. Here is the graphical representation of confusion matrix given below:

```
import matplotlib.pyplot as plt
metrics.plot_confusion_matrix(model, X_test, y_test, values_format= '.3g')

<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x7f51d63ef400>
```



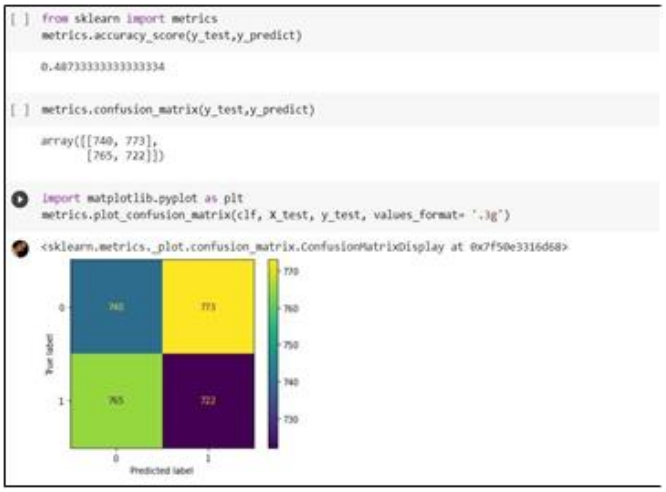
V. SUPPORT VECTOR MACHINE

Support Vector Machine also makes use of test train split as Logistic Regression and helps to predict the train and test values as shown below.

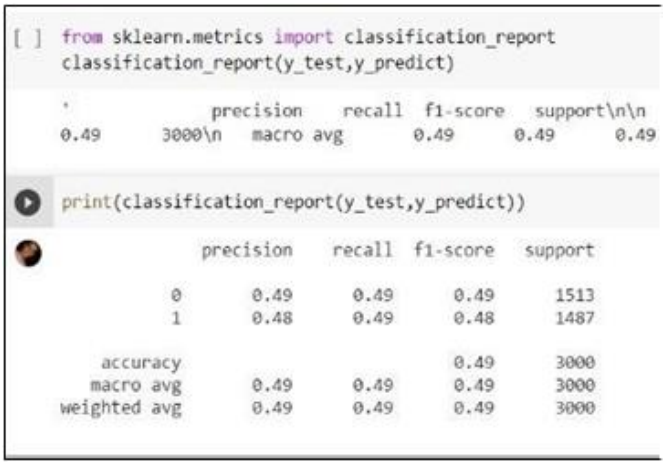
```
CREATING A MODEL

from sklearn import svm
clf=svm.SVC(kernel='linear', C=2.0, gamma='auto', probability=True)
clf.fit(X_train,y_train)

SVC(C=2.0, break_ties=False, cache_size=200, class_weight=None, coef0=0.0,
    decision_function_shape='ovr', degree=3, gamma='auto', kernel='linear',
    max_iter=-1, probability=True, random_state=None, shrinking=True, tol=0.001,
    verbose=False)

y_predict=clf.predict(X_test)
y_predict

array([1, 0, 1, ..., 0, 1, 1])

y_predictprob=clf.predict_proba(X_test)
y_predictprob

array([[0.50935026, 0.49064974],
       [0.51207009, 0.48792991],
       [0.50940267, 0.49059733],
       ...,
       [0.51205863, 0.48794137],
       [0.50937063, 0.49062939],
       [0.50940576, 0.49059424]])
```

Same as Decision trees in SVM confusion matrix is developed and it helps in predicting values for

trained data and it also shows accuracy same as above two methods i.e., 50%.



Here, we can see a clear classification report as mentioned above about precision, F1 score, Recall and support.



VI. CONCLUSION

In our report, we developed and tested a range of machine learning model approaches and found the most significant COVID-19 features were (in descending order): fever, temperature, Bodypain, Lotas (Loss of Taste and smell) etc., Our models were able to predict the stage of COVID-19 based on available patient's information travel and clinical symptoms.

We implemented ML algorithms on different clinical features of patients with COVID-19 infections in a new dataset from mainland China and used different classifiers to examine information and assess performance. Our ability to predict the probability and course of COVID-19 spread will improve the maximum doctors to identify infected patients at an early stage by utilizing clinical features. Some of the classifiers did not show reliable outcomes, presumably because while they demonstrated exactitude, they had created one-sided results for these datasets. However, the size of the COVID19 dataset was probably not extensive enough to give enough statistical power to resolve these issues. In future studies, using much larger datasets, we will improv our capacity to circumvent these limitations and also further improve our predictive accuracy.

VII. REFERENCES

[1]. WHO. Pneumonia of unknown cause – China 2020. Available from : https://www.who.int/csr/don/05-january- 2020-pneumonia-of-unkown-cause- china/en/. [Cited 2020 28 February].

[2]. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus- infected pneumonia. N Engl J Med. 2020;382(13):1199–207.

[3]. Meng Z, Wang M, Song H, Guo S, Zhou Y, Li W, et al. Development and utilization of an intelligent application for aiding COVID-19 diagnosis. medRxiv. 2020 ; https://doi.org/10.1101/2020.03.18.20035 816.

[4]. GitHub Repository on China Covid data of 2019-2020.

[5]. https://youtu.be/7sz4WpkUIIs Support Vector Machine

[6]. https://youtu.be/02eZFXALcl4 Python Extraction of data

[7]. Catriona Manville, Gavin Cochrane, Jonathan Cave, Jeremy Millard, Jimmy Kevin Pederson, Rasmus Kåre Thaarup, . . . Bas Kotterink. (2014). Mapping smart cities in the EU. Retrieved 4/11/2016, 2016, from http://www.smartcities.at/assets/Publikatio

nen/Weitere-Publikationenzum-Thema/mappingsmartcities.pdf.

[8]. Assessment of Linguistics Web Technologies Based on Cyclic Sequential Access Structure, CIKITUSI JOURNAL FOR MULTIDISCIPLINARY RESEARCH, ISSN NO : 0975-6876, Advin Manhar, Mohammed Bakhtawar

[9]. Naming the coronavirus disease (COVID-19) and the virus that causes it. World Health Organization. (2020). http://www.who.int/emergencies/diseases/novel-coronavirus2019/technical-guidance/naming-the-coronavirus-disease-(covid2019)-and-the-virus-that-causes-it Accessed 18 April 2020.

[10]. Gupta, M.. ML: Feature Scaling – Part 2. (2019). GeeksforGeeks. URL https://www.geeksforgeeks.org/ml-feature-scaling-part-2/ Accessed 18 April 2020.

[11]. BDBC-KG-NLP/COVID-19-tracker. GitHub. (2020). https://github.com/BDBC-KG-NLP/COVID-19-tracker Accessed 20 February 2020

[12]. Agarwal, R. The 5 Classification Evaluation metrics every Data Scientist must know.(2019). https://towardsdatascience.com/the-5-classificationevaluation-metrics-you-must-know-aa97784ff226 Accessed 18 April 2020. https://www.aljazeera.com/news/2020/01/countries-confirmed-casescoronavirus-200125070959786.html Accessed 18 April 2020

[13]. Karim, M., & Rahman, R.M. (2013). Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing. Journal of Software Engineering and Applications 06, 196–206. doi:10.4236/jsea.2013.64025.

[14]. Larose, C.D., & Larose, D.T. (2019). Data science using Python and R. Wiley, Hoboken. (Chapter 6).

[15]. Li, F., Li, Y.-Y., & Wang, C. (2009). Uncertain data decision tree classification algorithm. Journal of Computer Applications 29, 3092– 3095. doi:10.3724/sp.j.1087.2009.03092.

[16]. Peeri, N.C., Shrestha, N., Rahman, M.S., Zaki, R., Tan, Z., Bibi, S., Baghbanzadeh, M., Aghamohammadi, N., Zhang, W., & Haque, U. (2020). The SARS, MERS and novel coronavirus (COVID-19) epidemics, the newest and biggest global health threats: what lessons have we learned? International Journal of Epidemiology. doi:10.1093/ije/dyaa033.

[17]. Chan, J.F.-W., Yuan, S., Kok, K.-H., To, K.K.- W., Chu, H., Yang, J., Xing, F., Liu, J., Yip, C.C.-Y., Poon, R.W.-S., Tsoi, H.-W., Lo, S.K.- F., Chan, K.-H., Poon, V.K.-M., Chan, W.-M., Ip, J.D., Cai, J.-P., Cheng, V.C.- C., Chen, H., Hui, C.K.-M., & Yuen, K.-Y. (2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. The Lancet 395, 514–523. doi:10.1016/s0140-6736(20)30154-9.

[18]. Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., Xiao, Y., Gao, H., Guo, L., Xie, J., Wang, G., Jiang, R., Gao, Z., Jin, Q., Wang, J., & Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet 395, 497–506. doi:10.1016/s0140-6736(20)30183-5.

[19]. Nishiura, H., Jung, S.-M., Linton, N.M., Kinoshita, R., Yang, Y., Hayashi, K., Kobayashi, T., Yuan, B., & Akhmetzhanov, A.R. (2020). The Extent of Transmission of Novel Coronavirus in Wuhan, China, 2020. Journal of Clinical Medicine 9, 330. doi:10.3390/jcm9020330

Cite this article as : K. Anurag Reddy, Susant Kumar Rath, Advin Manhar, "A Machine Learning Approach to Analyse the Symptoms of Covid-19 for the Initial Diagnosis of a Patient", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 1, pp. 34-40, January-February 2021. Available at

doi : https://doi.org/10.32628/CSEIT21711

Journal URL : http://ijsrcseit.com/CSEIT21711