Classification of Covid-19 Dataset using Machine Learning

Krishnaveni[1] and Privietha P[2]

[1]Student, Hindusthan College of Engineering and Technology, Coimbatore, Tamil Nadu, India
[2]Assistant Professor, Department of Computer Applications,
Hindusthan College of Engineering and Technology, Coimbatore, Tamil Nadu, India
[1]krishnavenigopala18@gmail.com, [2]priviethaprabhakar@gmail.com

Abstract. The project entitled as "Classification of covid-19 analysis using Machine language" is a dataset analysis. The term data analytics refers to the process of examining datasets to draw conclusions. Data analytic techniques enables raw data and uncover patterns to extract valuable insights. Readily available datasets from Kaggle website is used for classification process. Classification is a supervised learning task and pandas is a python library. In order to construct a classification model machine algorithm is predefined. The training model is used or predict a class for new coming document. Seaborn is a library that users Matplot underneath to plot graphs. It will be used to visualize random distribution Dataset. Training is the process that makes the system 'learn' the pattern typical classification. We use the default scikit-learn implementation of logistic regression and linear support vector machine for multi-label classification, which trains one classifier per class using a one-vs-rest scheme.

Keywords: classification, COVID, Logistic Regression, XGBoost, Random Forest.

1. Introduction

The COVID-19 pandemic has made it a global priority for research on the subject to be developed at unprecedented rates all over the world. Researchers in a wide variety of fields, from clinicians to epidemiologists to policy makers, must all have effective access to the most up to date publications in their respective areas. Automated document classification can play an important role in organizing the stream of articles by fields and topics to facilitate the search process and speed up research efforts[1]. We explore the data efficiency and generalizability of these models as crucial aspects to address for document classification to become a useful tool against outbreaks like this one. During a sudden healthcare crisis like this pandemic, it is essential for models to obtain useful results as soon as possible. Since labelling biomedical articles is a very time-consuming process, achieving peak performance using less data becomes highly desirable. The data efficiency of these models by training each of the ones is evaluated[2].

2. Existing System

The symptoms of COVID-19 fever, cough, difficulty breathing and muscle pain can resemble those of many other diseases, such as influenza, making diagnostic tests therefore essential for identifying people who actually have COVID-19. In addition to this, these tests can also help determine who has recovered from COVID-19, as well as improve our understanding of how the virus spreads and help monitor the effectiveness of control measures.

Testing for the virus itself vs antibodies:

Some test for the virus itself, by looking for the RNA (the genetic blueprint) of the SARS-CoV-2 virus that causes COVID-19. When carried out properly, a result that the virus has been detected is extremely reliable[3]. However, these tests are not very helpful for determining whether someone has recovered from the virus, and can potentially miss the virus if it is present in extremely low levels in a patient's body. Other tests look for antibodies to the virus – evidence that the body has produced an immune response to it. It takes time for such antibodies to be created, so antibody tests are not much use in confirming if someone has COVID-19 in the first few days of infection. However, in contrast to the RNA tests, they can be extremely useful in determining whether someone has previously been infected with the new coronavirus, but no longer has the virus present. A complicating factor, however, is that different people can have different antibody responses to COVID-19. For example, individuals with severe disease seem to develop higher antibody levels than individuals with mild or asymptomatic disease. As a result, a test for antibodies developed using blood samples from individuals with severe COVID-19 may not work as well in detecting antibodies in people with a mild or asymptomatic version of the disease, where there are far fewer antibodies to detect[4].

3. Proposed System

Machine-learning on the data allowed us to construct pre-test models predicting whether a patient would test positive for a particular virus. Text mining improved the predictions for one viral test[5]. Cost-sensitive models optimized for test sensitivity showed reasonable test specificities and an ability to reduce test volume by up to 46% for single viral tests. We conclude that diverse forms of data in the electronic medical record can be used productively to build models that help physicians reduce testing volumes[6].

4. System Specification

Hardware Specification

- Processor : Intel Core i5
- RAM: 8GB
- Hard Disk Drive : 1TB

Software Specification

- Operating System : Windows 7or above
- Application : Jupyter Notebook

5. Software Description

Python

Python is a computer programming language often used to build websites and software, automate tasks, and conduct data analysis. Python is a general-purpose language, meaning it can be used to create a variety of different programs and isn't specialized for any specific problems[7]. Python is used for server-side web development, software development, mathematics, and system scripting, and is popular for Rapid Application Development and as a scripting or glue language to tie existing components because of its high-level, built-in data structures, dynamic typing, and dynamic binding.

Jupyter Notebook:

Jupyter Notebook is the latest web based interactive development environment for notebooks, code, and data. Its flexible interface allows users to configure and arrange workflows in data science, scientific computing, computational journalism, and machine learning. A modular design invites extension to expand and enrich functionality.

Data Set

- **Data Description:** Data collected is from March 2020 - November 2021
- **Symptoms:** Cough, Fever, Sore Throat, Shortness of Breath & Headache.
- **Other Features:** Gender, Age 60 and above, Test indication & Test date.
- **Target** Feature: Corona Result

6. Materials And Methods

Logistic Regression

In the Equation(1), If Y takes on more than two values like in our case, say k of them, we can still use logistic regression. Instead of having one set of parameters β0, β, each class c in 0: (k −1) will have its own offset β(c) 0 and vector β(c), and the predicted conditional probabilities will be

$$P_r\,(Y = c|X = x) = \frac{e^{\beta_0{}^{(c)} + x^T\beta^{(c)}}}{\sum_k e^{\beta_0^{(k)} + x^T\beta^{(k)}}} \qquad (1)$$

Decision Tree

A decision tree is a tree where each node represents a feature (attribute), each link (branch) represents a decision (rule) and each leaf represents an outcome (categorical or continuous value).

Decision trees learn how to best split the dataset into smaller and smaller subsets to predict the target value[8]. The splitting process continues until no further gain can be made or a preset rule is met, e.g. the maximum depth of the tree is reached.

Random Forest

The Random Forest Algorithm is composed of different decision trees, each with the same nodes, but using different data that leads to different leaves. It merges the decisions of multiple decision trees in order to find an answer, which represents the average of all these decision trees. Gini-index is often used to how branching is done by nodes in a decision tree.

This formula uses the class and probability to determine the Gini of each branch on a node, determining which of the branches is more likely to occur[9]. Here, *pi* represents the relative frequency of the class we are observing in the dataset and *c* represents the number of classes.

Support Vector Machine

The main objective of SVM is to find the optimal hyperplane which linearly separates the data points in two components by maximizing the margin. The point above or on the hyperplane will be classified as class +1, and the point below the hyperplane will be classified as class -1. Computing the (soft-margin) SVM classifier amounts to minimizing an expression of the for

We focus on the soft-margin classifier since choosing a sufficiently small value for lambda yields the hard-margin classifier for linearly-classifiable input data.

$$k(x, y) = \tanh(\alpha x^T y + c) \qquad (2)$$

The Equation (2) represents the kernel methods map the data into higher dimensional spaces in the hope that in this higher-dimensional space the data could become more easily separated or better structured. There are also no constraints on the form of this mapping, which could even lead to infinite-dimensional spaces. The Sigmoid Kernel (Hyperbolic Tangent) comes from the Neural Networks field, where the bipolar sigmoid function is often used asan activation function for artificial neurons.

Methodology

Model is proposed based on classification algorithm “Logistic Regression”. Based on variables like cough, fever, sour, Throat pain that, covid 29 viruses infected persons. Comparison with algorithm like KNN, is done to choose the best method. Figure1 represents the flowchart of the methodology to process the model[10].

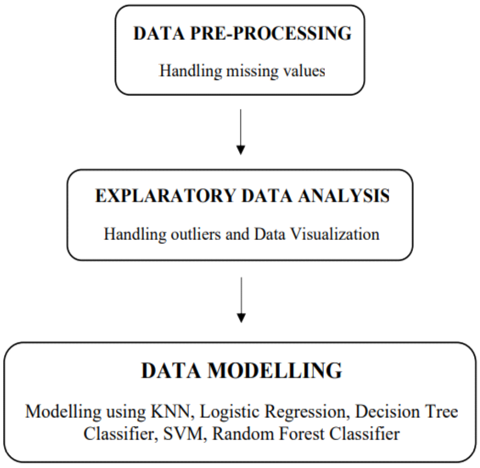

DATA PRE-PROCESSING
Handling missing values
EXPLARATORY DATA ANALYSIS
Handling outliers and Data Visualization
DATA MODELLING
Modelling using KNN, Logistic Regression, Decision Tree Classifier, SVM, Random Forest Classifier

Figure 1: Data modeling.

7. Implementation & Result Analysis

Encoding The Features

Dropped test_indication = other because it is not specified what other.

Dropped test_date because our objective is to detect if a patient is Covid Positive or Negative based on Symptoms, Gender, Age & Test Indications. {\display style (x_{i},\;y_{i})}. Figure 2 shows the pictorial representation of data analysis of Covid dataset.

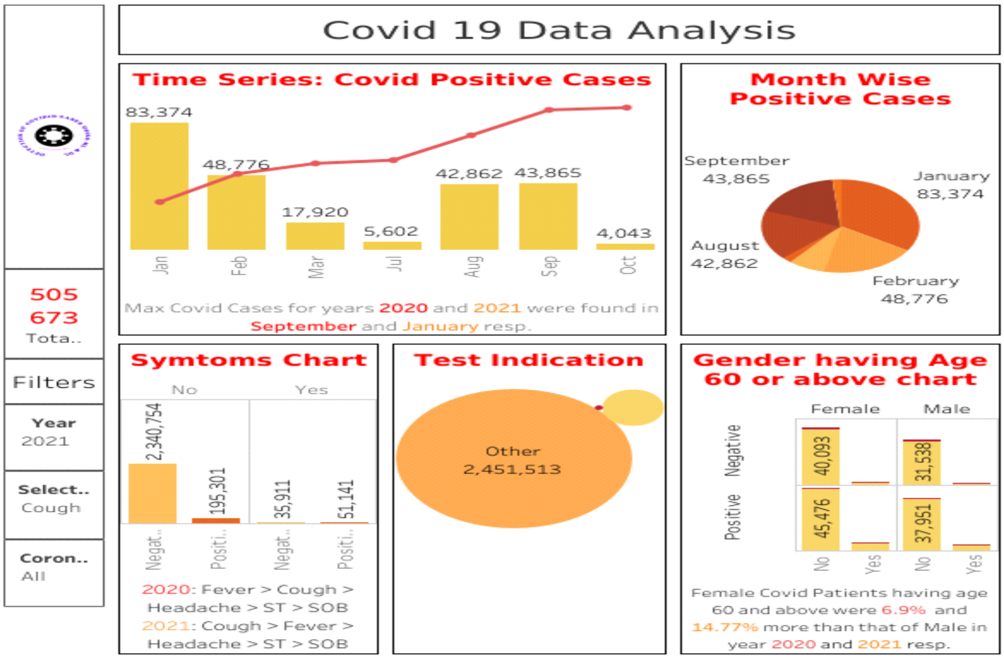


Figure 2: Correlation map.

Extracting Risk Coefficient from the data:

We observed that fever was the most common symptom found among covid positive patients (Links + EDA Dashboard) so we gave 0.2 weight for it.

Contact with covid patient would directly lead to home quarantine so we gave 0.2 weight for it.

Remaining symptoms & Age 60+ features got 0.1 weight. Removed Contradictory records using risk coefficient. And the use of risk coefficient is over Datatype for all features is converted to integer.

- Covid Positive Cases - 162021,
- Covid Negative Cases - 5337010

To Under sampling instead of Oversampling because

- Data is abundant for Negative Cases.
- Increasing Positive Cases by oversampling would be an issue according to real world scenario.

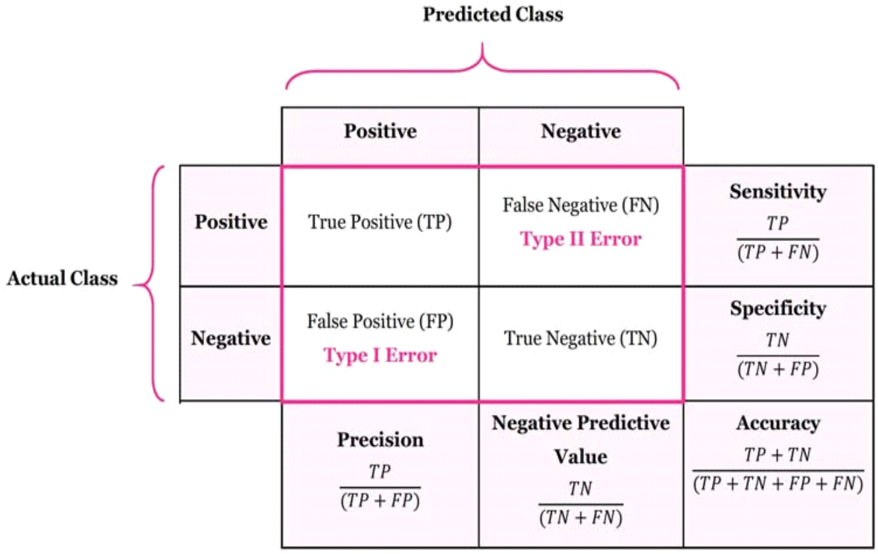
We observe that the data is now balanced.

- Covid Positive Cases - 162021, Covid Negative Cases – 270035

Feature Selection

- Anova Test
- Chi Square Test
- No feature seems unimportant.
- All the features are contributing towards the detection of covid cases.
- Data Modeling
- Train Test Split

Evaluation Metrics



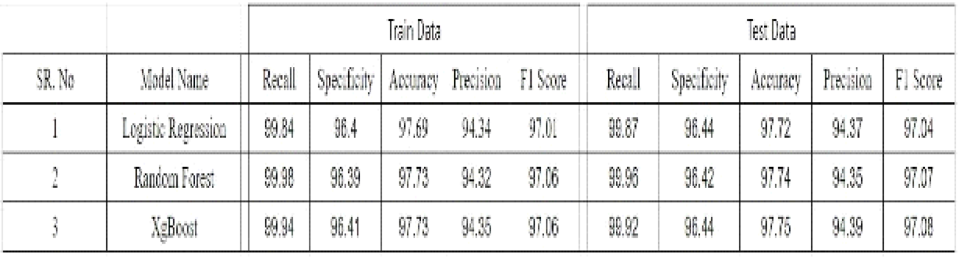
		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

Figure 3: Feature Selection.

Figure 3 represents the cross validation for feature selection using 5 metrics to evaluate our Models. Metrics - Recall, Specificity, Accuracy, Precision & F1 Score.

Model Evaluation - Calculated all the metrics using Confusion Matrix.

Figure 4 represents the accuracy table for the training and testing data of logistic regression, Random Forest and XG Boost.



		Train Data					Test Data				
SR. No	Model Name	Recall	Specificity	Accuracy	Precision	F1 Score	Recall	Specificity	Accuracy	Precision	F1 Score
1	Logistic Regression	99.84	96.4	97.69	94.34	97.01	99.87	96.44	97.72	94.37	97.04
2	Random Forest	99.98	96.39	97.73	94.32	97.06	99.96	96.42	97.74	94.35	97.07
3	XgBoost	99.94	96.41	97.73	94.35	97.06	99.92	96.44	97.75	94.39	97.08

Figure 4: Model Evaluation.

8. Future Enhancements

In future, the reassures has planned to us the model to predict unsupervised data. The dataset can be revised and updates with latest trend of covid 19 in India. The main focus in future work is to refine the parameters to increase the accuracy rate of the model. So that the study helps the authorities to take timely actions and make decisions accordingly.

9. Conclusion

This analysis is provided for document classification models on the LitCovid dataset for the COVID19 .Fine-tuning pretrained language models yields the best performance on this task. We study the generalizability and data efficiency of these models, evaluate the effect of article titles on performance through a data and discuss some important issues to address in future work.

References

1. Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019a. Docbert: Bert for document classification. ArXiv, abs/1904.08398.
2. Simon Baker, Ilona Silins, Yufan Guo, Imran Ali, Johan Hogberg, Ulla Stenius, and Anna Korhonen. ¨ 2016. Automatic semantic classification of scientific literature according to the hallmarks of cancer. Bioinformatics, 32 3:432–40.
3. Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Long former: The long-document transformer. arXiv:2004.0515.
4. Ali S, Patterson M. Spike2vec: An efficient and scalable embedding approach for covid-19 spike sequences. In 2021 IEEE Conference on Big Data (Big Data) 2021 Dec 15 (pp. 1533-1540).
5. Ali S, Bello B, Patterson M (2021a) Classifying covid-19 spike sequences from geographic location using deep learning. arXiv preprint arXiv:211000809 GISAID Website (Accessed: 10-12-2021) . https://www.gisai dorg/
6. Leung CK, Chen Y, Hoi CS, Shang S, Cuzzocrea A (2020a) Machine learning and olap on big covid-19 data. In: 2020 IEEE International Conference on Big Data (Big Data), pp 5118–5127
7. Leung CK, Chen Y, Shang S, Deng D (2020b) Big data science on covid-19 data. In: 2020 IEEE 14th International Conference on Data Science and Engineering (BigDataSE), pp 14–21
8. Ali S, Mansoor H, Arshad N, Khan I (2019a) Short term load forecasting using smart meter data. In: International Conference on Future Energy Systems, pp 419–421
9. Lundberg SM, Erion G, Chen H, Himmelfarb J, Bansal N, Lee SI (2020) From local explanations to global understanding with explainable ai for trees. Nat Mach Intell 2(1):2522–5839
10. Ali S, Shakeel MH, Khan I, Faizullah S, Khan MA (2021) Predicting attributes of nodes using network structure.