DETECTION OF COVID-19 USING MACHINE LEARNING ALGORITHMS

Main Project work submitted to Avinashilingam Institute for Home Science and Higher Education for Women

MASTER OF SCIENCE IN INFORMATION TECHNOLOGY

Submitted By

A. Kavya (19PIT003)

Under the guidance of

Dr. F. Paulin M.C.A., M.Phil., Ph.D.,

Assistant Professor, Department of Information Technology

AVINASHILINGAM INSTITUTE FOR HOME SCIENCE AND HIGHER EDUCATION FOR WOMEN

SCHOOL OF PHYSICAL SCIENCES AND COMPUTATIONAL SCIENCES

DEPARTMENT OF INFORMATION TECHNOLOGY

Coimbatore-641043

May 2021

DECLARATION

DECLARATION

I hereby declare that the project entitled **"DETECTION OF COVID-19 USING MACHINE LEARNING ALGORITHMS"** is a record of the original work done by A. Kavya (19PIT003) under the guidance of Dr. F. Paulin M.C.A., M.Phil., Ph.D., Assistant Professor, Department of Information Technology, School of Physical Sciences and Computational Sciences, Avinashilingam Institute for Home Science and Higher Education for Women, in the partial fulfilment for the degree of Master of Science in Information Technology and this project has not formed the basis for any Degree/Diploma/Associates.

Place:

Date:

Signature of the Candidate

_____

Countersigned by

Dr. F. Paulin M.C.A, M.Phil., Ph.D.,

Assistant Professor, Department of Information Technology,

School of Physical Sciences and Computational Sciences

CERTIFICATE

CERTIFICATE

This is to certify that this project work entitled "**DETECTION OF COVID-19 USING MACHINE LEARNING ALGORITHMS"** done by A. Kavya(19PIT003) has been submitted to Avinashilingam Institute for Home science and Higher education for women, Coimbatore-43 in partial fulfilment of the requirement for the award of the degree of **MASTER OF SCIENCE IN INFORMATION TECHNOLOGY.** This Project has not found the basis for the award of any Degree/Associate/fellowship or similar title to any Candidate of any University. Certified as a bonafied record of the work submitted for the Viva voce held on _____.

Signature of the HOD

Signature of the Guide

Signature of External Examiner

GATEWAY
SOFTWARE SOLUTIONS

ISO 9001:2015

Date: 30/04/2021

TO WHOMSOEVER IT MAY CONCERN

This is to certify the student **Ms. KAVYA A(19PIT003)** pursuing her final year in **MSC INFORMATION TECHNOLOGY** in **AVINASHILINGAM INSTITUTE FOR HOME SCIENCE & HIGHER EDUCATION FOR WOMEN, COIMBATORE** has completed her project entitled "**DETECTION OF COVID-19 USING MACHINE LEARNING ALGORITHMS**" in our concern starts from February 2021 to April 2021.

Wish her the best

GATEWAY SOFTWARE SOLUTIONS

GATEWAY SOFTWARE SOLUTIONS CBE-12

Manager

Mobile: 7397078885
E-mail : info@gatewaysoftwaresolutions.com / Webiste : gatewaysoftwaresolultions.com

ACKNOWLEDGEMENT

ACKNOWLEDGEMENT

I would like to express my sincere thanks to God Almighty, for his constant love and grace that he has showered upon me, which kept me in good health, and sound mind without which my project would not have reached a successful end.

I would like to express my deep sense of reverential gratitude and sincere thanks to **Dr. S. P. Thyagarajan, Chancellor,** Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for providing all facilities during my course of study.

I owe my great deal of gratitude to **Dr. PremavathyVijayan M.Sc., M.Ed., Dip. Spl. Edn., M.Phil., Ph.D., Vice Chancellor,** Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for extending all resources that facilitated the smooth conduct of the project study.

I express my gratitude to **Dr. S. Kowsalya, Registrar,** Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for providing all facilities and support necessary for the study.

I wish to extend my sincere thanks **Dr. K. Udaya Chandrika M.Sc., M.Phil., Ph.D., Dean School of Physical Sciences and Computational Sciences**, for her support and valuable guidance.

I take this opportunity to express my profound gratitude to **Dr. D. ShanmugaPriya, M.Sc., M.Phil., Ph.D., Head, Department of Information Technology,** School of Physical Sciences and Computational Sciences, for her valuable guidance and encouragement.

I heartily thank my esteemed project guide **Dr. F. Paulin, M.C.A, M.Phil., Ph.D., Assistant Professor, Department of Information Technology,** for imparting tremendous assistance and well-timed support for triumph of our project.

I express my honourable thanks to our project coordinator **Dr. T. Jayamalar M.C.A, M.Phil., Ph.D, Assistant Professor, Department of Information Technology,** for her kind advice and knowledgeable suggestions which helped us to complete our project successfully.

I would like to express my sincere gratitude to all the staff members of the Department of Information Technology, for their constant encouragement and for the opportunity to do our project in this esteemed university. Last yet importantly, I would like tothank my parents, family members, friends and all well-wishers for their kind inspiration, blessings and encouragement during the course of project.

ABSTRACT

ABSTRACT

Technology advancements have a rapid effect on every field of life, it would be in medical field or any other field. Artificial intelligence has shown the promising results in health care through its decision making by analysing the data. COVID-19 attacked more than 100 countries in a matter of no time. People all over the world are vulnerable to its consequences in future. It is imperative to develop a control system that will detect the coronavirus. One of the solutions to control the current havoc can be the diagnosis of disease with the help of various AI tools.

This project classified the textual clinical reports by using classical and ensemble machine learning algorithms. In Data preprocessing, cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency. Feature engineering was performed using some techniques and it is useful to improve the performance of machine learning algorithms and is often considered as applied machine learning. Feature extraction involves choosing a set of features from a large collection. These features were supplied to traditional and ensemble machine learning classifiers. In this work, three classifiers such as Gaussian Naïve Bayes, Random Forest Classifier and SGD Classifier were used and compared to check the accuracy. The best classifier would help to detect the covid-19 from the dataset using the machine learning algorithm. Random Forest showed better results than other Machine Learning algorithms by having 95% testing accuracy.

CONTENT

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 AN INTRODUCTION TO COVID 19

The World Health Organisation (WHO) has declared the coronavirus disease 2019 (COVID-19) a pandemic. A global coordinated effort is needed to stop the further spread of the virus. A pandemic is defined as "occurring over a wide geographic area and affecting an exceptionally high proportion of the population."

On 31 December 2019, a cluster of cases of pneumonia of unknown cause, in the city of Wuhan, Hubei province in China, was reported to the World Health Organisation. In January 2020, a previously unknown new virus was identified, subsequently named the 2019 novel coronavirus, and samples obtained from cases and analysis of the virus genetics indicated that this was the cause of the outbreak. This novel coronavirus was named Coronavirus Disease 2019 (COVID-19) by WHO in February 2020. The virus is referred to as SARS-CoV-2 and the associated disease is COVID-19.

According to W.H.O the signs and symptoms of mild to moderate cases are dry cough, fatigue and fever while as in severe cases dyspnea (shortness of breath), Fever and tiredness may occur. The persons having other diseases like asthma, diabetes, and heart disease are more vulnerable to the virus and may become severely ill. The person is diagnoses based on symptoms and his travel history. Vital signs are being observed keenly of the client having symptoms. No specific treatment has been discovered as on 10th April 2020, and patients are being treated symptomatically. The drugs like hydroxychloriquine, antipyretic, anti-virals are used for the symptomatic treatment. Currently, no such vaccine is developed for preventing this deadly disease, and we may take some precautions to prevent this disease. By washing hands regularly with soap for 20 s and avoiding close contact with others by keeping the distance of about 1 m may reduce the chances of getting affected by this virus. While sneezing, Covering the mouth and nose with the help of disposable tissue and avoiding the contact with the nose, ear and mouth can help in its prevention.

Since no drug is made for curing the COVID-19. Various paramedical companies have claimed of developing a vaccine for this virus. Less testing has also given rise to this disease

as lack the medical resources due to pandemic. Since thousands and thousands are being tested positive day by day around the globe, it is not possible to test all the persons who show symptoms. Apart from clinical procedures, machine learning provides a lot of support in identifying the disease with the help of image and textual data. Machine learning can be used for the identification of novel coronavirus. It can also forecast the nature of the virus across the globe. However, machine learning requires a huge amount of data for classifying or predicting diseases. Supervised machine learning algorithms need annotated data for classifying the text or image into different categories. From the past decade, a huge amount of progress is being made in this area for resolving some critical projects. Recent pandemic has attracted many researchers around the globe to solve this problem. The data consists of clinical reports in the form of text in this project, that are classified and it can help in detecting coronavirus from earlier clinical symptoms.

1.2 MACHINE LEARNING

Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people.

Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve. Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs.

Any technology user today has benefitted from machine learning. Facial recognition technology allows social media platforms to help users tag and share photos of friends. Optical character recognition (OCR) technology converts images of text into movable type. Recommendation engines, powered by machine learning, suggest what movies or television shows to watch next based on user preferences. Self-driving cars that rely on machine learning to navigate may soon be available to consumers.

Machine learning is a continuously developing field. Because of this, there are some considerations to keep in mind as you work with machine learning methodologies, or analyze the impact of machine learning processes.

1.2.1 MACHINE LEARNING METHODS

In machine learning, tasks are generally classified into broad categories. These categories are based on how learning is received or how feedback on the learning is given to the system developed.

Two of the most widely adopted machine learning methods are **supervised learning** which trains algorithms based on example input and output data that is labeled by humans, and **unsupervised learning** which provides the algorithm with no labeled data in order to allow it to find structure within its input data.

1.2.1.1 SUPERVISED LEARNING

In supervised learning, the computer is provided with example inputs that are labeled with their desired outputs. The purpose of this method is for the algorithm to be able to "learn" by comparing its actual output with the "taught" outputs to find errors, and modify the model accordingly. Supervised learning therefore uses patterns to predict label values on additional unlabeled data.

A common use case of supervised learning is to use historical data to predict statistically likely future events. It may use historical stock market information to anticipate upcoming fluctuations, or be employed to filter out spam emails. In supervised learning, tagged photos of dogs can be used as input data to classify untagged photos of dogs.

There are two main types of supervised learning problems: they are classification that involves predicting a class label and regression that involves predicting a numerical value.

- **Classification**: Supervised learning problem that involves predicting a class label.
- **Regression**: Supervised learning problem that involves predicting a numerical label.

Both classification and regression problems may have one or more input variables and input variables may be any data type, such as numerical or categorical.

1.2.1.2 UNSUPERVISED LEARNING

In unsupervised learning, data is unlabeled, so the learning algorithm is left to find commonalities among its input data. As unlabeled data are more abundant than labeled data, machine learning methods that facilitate unsupervised learning are particularly valuable.

The goal of unsupervised learning may be as straightforward as discovering hidden patterns within a dataset, but it may also have a goal of feature learning, which allows the computational machine to automatically discover the representations that are needed to classify raw data.

Unsupervised learning is commonly used for transactional data. You may have a large dataset of customers and their purchases, but as a human you will likely not be able to make sense of what similar attributes can be drawn from customer profiles and their types of purchases. With this data fed into an unsupervised learning algorithm, it may be determined that women of a certain age range who buy unscented soaps are likely to be pregnant, and therefore a marketing campaign related to pregnancy and baby products can be targeted to this audience in order to increase their number of purchases.

Without being told a "correct" answer, unsupervised learning methods can look at complex data that is more expansive and seemingly unrelated in order to organize it in potentially meaningful ways. Unsupervised learning is often used for anomaly detection including for fraudulent credit card purchases, and recommender systems that recommend what products to buy next. In unsupervised learning, untagged photos of dogs can be used as input data for the algorithm to find likenesses and classify dog photos together.

There are many types of unsupervised learning, although there are two main problems that are often encountered by a practitioner: they are clustering that involves finding groups in the data and density estimation that involves summarizing the distribution of data.

- **Clustering: Unsupervised** learning problem that involves finding groups in data.
- **Density Estimation**: Unsupervised learning problem that involves summarizing the distribution of data.

1.2.2 MACHINE LEARNING APPROACHES

As a field, machine learning is closely related to computational statistics, so having a background knowledge in statistics is useful for understanding and leveraging machine learning algorithms.

For those who may not have studied statistics, it can be helpful to first define correlation and regression, as they are commonly used techniques for investigating the relationship among quantitative variables. **Correlation** is a measure of association between two variables that are not designated as either dependent or independent. **Regression** at a basic level is used to examine the relationship between one dependent and one independent variable. Because regression statistics can be used to anticipate the dependent variable when the independent variable is known, regression enables prediction capabilities. The following approaches are used in this project.

1.2.2.1 NAÏVE BAYES ALGORITHM

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

Bayes' theorem is also known as **Bayes' Rule** or **Bayes' law**, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.

The formula for Bayes' theorem is given as:

$$P(A|B)= \frac{P(B|A)P(A)}{P(B)}$$

Where, P(A|B) is Posterior probability, Probability of hypothesis A on the observed event B.

GAUSSIAN NAÏVE BAYES: The Gaussian model assumes that features follow a normal distribution. This means if predictors take continuous values instead of discrete, then the model assumes that these values are sampled from the Gaussian distribution.

1.2.2.2 THE RANDOM FOREST CLASSIFIER

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction The fundamental concept behind random forest is a simple but powerful one — the wisdom of crowds. In data science speak, the reason that the random forest model works so well is:

- **A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.**

The low correlation between models is the key. Just like how investments with low correlations (like stocks and bonds) come together to form a portfolio that is greater than the sum of its parts, uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions. **The reason for this wonderful effect is that the trees protect each other from their individual errors** (as long as they don't constantly all err in the same direction). While some trees may be wrong, many other trees will be right, so as a group the trees are able to move in the correct direction. So the prerequisites for random forest to perform well are:

- There needs to be some actual signal in our features so that models built using those features do better than random guessing.
- The predictions (and therefore the errors) made by the individual trees need to have low correlations with each other.

1.2.2.3 STOCHASTIC GRADIENT DESCENT CLASSIFIER – LINEAR MODEL

Stochastic Gradient Descent (SGD) to the regularized linear methods can help building an estimator for classification and regression problems. Scikit-learn API provides the SGD Classifier

class to implement SGD method for classification problems. The SGD Classifier applies regularized linear model with SGD learning to build an estimator. The SGD classifier works well with large-scale datasets and it is an efficient and easy to implement method.

The word ‘stochastic’ means a system or a process that is linked with a random probability. Hence, in Stochastic Gradient Descent, a few samples are selected randomly instead of the whole data set for each iteration. In Gradient Descent, there is a term called “batch” which denotes the total number of samples from a dataset that is used for calculating the gradient for each iteration. In typical Gradient Descent optimization, like Batch Gradient Descent, the batch is taken to be the whole dataset. Although, using the whole dataset is really useful for getting to the minima in a less noisy and less random manner, but the problem arises when our datasets gets big. Suppose, you have a million samples in your dataset, so if you use a typical Gradient Descent optimization technique, you will have to use all of the one million samples for completing one iteration while performing the Gradient Descent, and it has to be done for every iteration until the minima is reached. Hence, it becomes computationally very expensive to perform. SGD is preferred over Batch Gradient Descent for optimizing a learning algorithm.

1.2.3 PROGRAMMING LANGUAGES

When choosing a language to specialize in with machine learning, the skills listed on current job advertisements as well as libraries available in various languages that can be used for machine learning processes.

From data taken from job ads on indeed.com in December 2016, it can be inferred that Python is the most sought-for programming language in the machine learning professional field. Python is followed by Java, then R, then C++.

1.2.3.1 PYTHON

Python’s popularity may be due to the increased development of deep learning frameworks available for this language recently, including TensorFlow, PyTorch, and Keras. As a language that has readable syntax and the ability to be used as a scripting language, Python proves to be powerful and straightforward both for preprocessing data and working with data directly.

The scikit-learn machine learning library is built on top of several existing Python packages that Python developers may already be familiar with, namely NumPy, SciPy, and Matplotlib.

1.3 ABOUT THE PLATFORM

1.3.1 ANACONDA NAVIGATOR

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda distribution that allows you to launch applications and easily manage conda packages, environments, and channels without using command-line commands. Navigator can search for packages on Anaconda.org or in a local Anaconda Repository. It is available for Windows, macOS, and Linux.

In order to run, many scientific packages depend on specific versions of other packages. Data scientists often use multiple versions of many packages and use multiple environments to separate these different versions.

The command-line program conda is both a package manager and an environment manager. This helps data scientists ensure that each version of each package has all the dependencies it requires and works correctly.

Navigator is an easy, point-and-click way to work with packages and environments without needing to type conda commands in a terminal window. You can use it to find the packages you want, install them in an environment, run the packages, and update them – all inside Navigator.
The following applications are available by default in Navigator:

- JupyterLab
- Jupyter Notebook
- Spyder
- PyCharm
- VSCode
- Glueviz
- Orange 3 App
- RStudio

- Anaconda Prompt (Windows only)
- Anaconda PowerShell (Windows only)

1.3.1 JUPYTER NOTEBOOK

Jupyter Notebook, an open-source, web-based IDE with deep cross-language integration that allows you to create and share documents containing live code, equations, visualizations, and narrative text. Data scientists and engineers love using Jupyter for data cleaning and transformation, statistical modeling, visualization, machine learning, deep learning, and much more. Jupyter Notebook's format (ipynb) has become an industry standard and can be rendered in multiple IDEs, GitHub, and other places.

Jupyter has support for over 40 programming languages, including Python, R, Julia, and Scala. Notebooks can be shared easily with others, and code can produce rich, interactive output, including HTML, images, videos, and custom MIME types. It allows you to leverage big data tools such as Spark and explore that same data with pandas, scikit-learn, TensorFlow, and ggplot2.

Jupyter has become an important part of the workflow for data scientists to process, analyze, and manipulate their data and draw insights from it in a pleasant and effective way. The open and standardized Jupyter notebook file format is designed to capture, display, and share natural language, code, and results in a self-contained and powerful computational narrative.

In 2014, Fernando Pérez announced a spin-off project from IPython called Project Jupyter. IPython continues to exist as a Python shell and a kernel for Jupyter, while the notebook and other language-agnostic parts of IPython moved under the Jupyter name. In 2015, GitHub and the Jupyter Project announced native rendering of Jupyter notebooks file format (.ipynb files) on the GitHub platform.

CHAPTER 2

LITERATURE REVIEW

Akib Mohi Ud Din Khanday et. al. [2020], COVID-19 has shocked the world due to its non-availability of vaccine or drug. Various researchers are working for conquering this deadly virus. We used 212 clinical reports which are labelled in four classes namely COVID, SARS, ARDS and both (COVID, ARDS). Various features like TF/IDF, bag of words are being extracted from these clinical reports. The machine learning algorithms are used for classifying clinical reports into four different classes. After performing classification, it was revealed that logistic regression and multinomial Naı¨ve Bayesian classifier gives excellent results by having 94% precision, 96% recall, 95% f1 score and accuracy 96.2%.

L.J. Muhammad et. al. [2020], Supervised machine learning models for COVID-19 infection were developed in this work with learning algorithms which include logistic regression, decision tree, support vector machine, naive Bayes, and artificial neutral network using epidemiology labeled dataset for positive and negative COVID-19 cases of Mexico. The correlation coefficient analysis between various dependent and independent features was carried out to determine a strength relationship between each dependent feature and independent feature of the dataset prior to developing the models. The 80% of the training dataset were used for training the models while the remaining 20% were used for testing the models. The result of the performance evaluation of the models showed that decision tree model has the highest accuracy of 94.99% while the Support Vector Machine Model has the highest sensitivity of 93.34% and Naïve Bayes Model has the highest specificity of 94.30%.

Kolla Bhanu Prakash et. al. [2020], An analysis on COVID-19 datasets to understand which age group is mostly effected due to COVID-19. Different prediction models are built using machine learning algorithms and their performances are computed and evaluated. Random Forest Regressor and Random Forest Classifier outperformed the other machine learning models like SVM, KNN+NCA, Decision Tree Classifier, Gaussian Naïve Bayesian Classifier, Multilinear Regression, Logistic Regression and XGBoost Classifier. The experiments reveal the persons of age groups 20-30, 30- 40 and 40-50 are suffered with COVID-19. The correlation matrices are built to understand the relationship between the features of the datasets. The feature importance is

computed for the classifiers built. Along with the classifiers and repressors are also built for prediction. The results show that the Random Forest Regressor and Random Forest Classifier has outperformed other models in terms of CoD and Accuracy.

Palash Ghosh et. al. [2020], In this paper, the aim is to analyze data on the number of infected people in each Indian state (restricted to only those states with enough data for prediction) and predict the number of infections for that state in the next 30 days. We hope that such statewise predictions would help the state governments better channelize their limited health care resources. Found that 7 states, namely, Maharashtra, Delhi, Gujarat, Madhya Pradesh, Andhra Pradesh, Uttar Pradesh, and West Bengal are in the severe category. Among the remaining states, Tamil Nadu, Rajasthan, Punjab, and Bihar are in the moderate category, whereas Kerala, Haryana, Jammu and Kashmir, Karnataka, and Telangana are in the controlled category. Also tabulated actual predicted numbers from various models for each state. States with nondecreasing DIR values need to immediately ramp up the preventive measures to combat the COVID-19 pandemic. On the other hand, the states with decreasing DIR can maintain the same status to see the DIR slowly become zero or negative for a consecutive 14 days to be able to declare the end of the pandemic.

Yazeed Zoabi et. al. [2021], In this paper, proposed a machine-learning model that predicts a positive SARS-CoV-2 infection in a RT-PCR test by asking eight basic questions. The model was trained on data of all individuals in Israel tested for SARS-CoV-2 during the first months of the COVID-19 pandemic. Thus, our model can be implemented globally for effective screening and prioritization of testing for the virus in the general population. Using predictions from the test set, the possible working points are: 87.30% sensitivity and 71.98% specificity, or 85.76% sensitivity and 79.18% specificity.

PeipeiWang et. al. [2020], In this article, a forecasting method with Logistic and Prophet model is proposed to analysis the COVID-19. The cap value is fitted by Logistic model through determining the fastest growing point, which is then feed into Prophet model for forecasting. We conducted experiments for global pandemic and also in some particular countries to forecast the epidemic peak, growing fasted point and the turning point of recovered. The Results are plotted and the predictive trend of global and five particular countries, and demonstrate the effectiveness of our model to predicting the turning point and epidemic size of COVID-19.

Hoyt Burdick et. al. [2020], In this project, Currently, physicians are limited in their ability to provide an accurate prognosis for COVID-19 positive patients. Existing scoring systems have been ineffective for identifying patient decompensation. Machine learning (ML) may offer an alternative strategy. A prospectively validated method to predict the need for ventilation in COVID-19 patients is essential to help triage patients, allocate resources, and prevent emergency intubations and their associated risks. 197 patients were enrolled in the REspirAtory Decompensation and model for the triage of covid-19 patients: a prospective studY (READY) clinical trial. The algorithm had a higher diagnostic odds ratio (DOR, 12.58) for predicting ventilation than a comparator early warning system, the Modified Early Warning Score (MEWS). The algorithm also achieved significantly higher sensitivity (0.90) than MEWS, which achieved a sensitivity of 0.78, while maintaining a higher specificity ($p < 0.05$).

Ricardo C. Cury et. al. [2021], Coronavirus disease 2019 (COVID-19) has spread quickly throughout the United States (US) causing significant disruption in healthcare and society. Tools to identify hot spots are important for public health planning. The goal of our study was to determine if natural language processing (NLP) algorithm assessment of thoracic computed tomography (CT) imaging reports correlated with the incidence of official COVID-19 cases in the US. The NLP algorithms were applied to 450,114 patient chest CT comprehensive reports gathered from January 1st to October 3rd, 2020. The best performing NLP model exhibited strong correlation with daily official COVID-19 cases (r^2=0.82, $p<0.005$). The NLP models demonstrated an early rise in cases followed by the increase of official cases, suggesting the possibility of an early predictive marker, with strong correlation to official cases on a weekly basis (r^2=0.91, $p<0.005$). There was also substantial correlation between the NLP and official COVID-19 incidence by state (r^2=0.92, $p<0.005$).

Jiangpeng Wu et. al. [2020], In this study, 11 key blood indices were extracted through random forest algorithm to build the final assistant discrimination tool from 49 clinical available blood test data which were derived by commercial blood test equipments. The method presented robust outcome to accurately identify COVID-19 from a variety of suspected patients with similar CT information or similar symptoms, with accuracy of 0.9795 and 0.9697 for the cross-validation set and test set, respectively. The tool also demonstrated its outstanding performance on an external validation set that was completely independent of the modeling process, with sensitivity,

specificity, and overall accuracy of 0.9512, 0.9697, and 0.9595, respectively. Besides, 24 samples from overseas infected patients with COVID-19 were used to make an in-depth clinical assessment with accuracy of 0.9167. After multiple verification, the reliability and repeatability of the tool has been fully evaluated, and it has the potential to develop into an emerging technology to identify COVID-19 and lower the burden of global public health. The proposed tool is well-suited to carry out preliminary assessment of suspected patients and help them to get timely treatment and quarantine suggestion.

Dac-Nhuong Le et. al. [2020], In this paper presents a novel IoT enabled Depthwise separable convolution neural network (DWS-CNN) with Deep support vector machine (DSVM) for COVID-19 diagnosis and classification. The proposed DWS-CNN model aims to detect both binary and multiple classes of COVID-19 by incorporating a set of processes namely data acquisition, Gaussian filtering (GF) based preprocessing, feature extraction, and classification. Initially, patient data will be collected in the data acquisition stage using IoT devices and sent to the cloud server. Besides, the GF technique is applied to remove the existence of noise that exists in the image. Then, the DWS-CNN model is employed for replacing default convolution for automatic feature extraction. Finally, the DSVM model is applied to determine the binary and multiple class labels of COVID-19. The diagnostic outcome of the DWS-CNN model is tested against Chest X-ray (CXR) image dataset and the results are investigated interms of distinct performance measures. The experimental results ensured the superior results of the DWS-CNN model by attaining maximum classification performance with the accuracy of 98.54% and 99.06% on binary and multiclass respectively.

Jim Samuel et. al. [2020], In this research article, covered four critical issues: (1) public sentiment associated with the progress of Coronavirus and COVID-19, (2) the use of Twitter data, namely Tweets, for sentiment analysis, (3) descriptive textual analytics and textual data visualization, and (4) comparison of textual classification mechanisms used in artificial intelligence (AI). The rapid spread of Coronavirus and COVID-19 infections have created a strong need for discovering efficient analytics methods for understanding the flow of information and the development of mass sentiment in pandemic scenarios. While there are numerous initiatives analyzing healthcare, preventative, care and recovery, economic and network data, there has been relatively little emphasis on the analysis of aggregate personal level and social media

communications. McKinsey ecently identified critical aspects for COVID-19 management and economic recovery scenarios. In their industry-oriented report, they emphasized data management, tracking and informational dashboards as critical components of managing a wide range of COVID-19 scenarios.

Frank S. Heldt et. al. [2021], In this retrospective study, we analysed data of 879 confirmed SARS-CoV-2 positive patients admitted to a two-site NHS Trust hospital in London, England, between January 1st and May 26th, 2020, with a majority of cases occurring in March and April. Extracted anonymised demographic data, physiological clinical variables and laboratory results from electronic healthcare records (EHR) and applied multivariate logistic regression, random forest and extreme gradient boosted trees. To evaluate the potential for early risk assessment, we used data available during patients' initial presentation at the emergency department (ED) to predict deterioration to one of three clinical endpoints in the remainder of the hospital stay: admission to intensive care, need for invasive mechanical ventilation and in-hospital mortality. Based on the trained models, we extracted the most informative clinical features in determining these patient trajectories. Considering our inclusion criteria, have identified 129 of 879 (15%) patients that required intensive care, 62 of 878 (7%) patients needing mechanical ventilation, and 193 of 619 (31%) cases of in-hospital mortality. Our models learned successfully from early clinical data and predicted clinical endpoints with high accuracy, the best model achieving area under the receiver operating characteristic (AUC-ROC) scores of 0.76 to 0.87 (F1 scores of 0.42–0.60). Younger patient age was associated with an increased risk of receiving intensive care and ventilation, but lower risk of mortality. Clinical indicators of a patient's oxygen supply and selected laboratory results, such as blood lactate and creatinine levels, were most predictive of COVID-19 patient trajectories. Among COVID-19 patients machine learning can aid in the early identification of those with a poor prognosis, using EHR data collected during a patient's first presentation at ED. Patient age and measures of oxygenation status during ED stay are primary indicators of poor patient outcomes.

S.NO	TITLE OF THE PAPER	AUTHOR'S NAME & YEAR	TECHNIQUES USED	OBSERVATIONS
1.	Machine learning based approaches for detecting COVID 19 using clinical text dataset	Akib Mohi Ud Din Khanday, Syed Tanzeel Rabani, Qamar Rayees Khan, Nusrat Rouf, Masarat Mohi Ud Din June 2020	Logistic regression and Multinomial Naive Bayes	ML algorithms by having testing accuracy.
2.	Supervised Machine Learning Models for prediction of COVID 19 Infection using Epidemiology datasets	L.J. Muhammad, Ebrahem A. Algehyne, Sani Sharif Usman, Abdulkadir Ahmad, Chinmay Chakraborty November 2020	Logistic regression, Decision tree, SVM, Naive Bayes and ANN	The model developed with decision tree happened to be the best model among all models developed in terms of accuracy with 94.99%.
3.	Analysis, Prediction and Evaluation of COVID-19 Datasets using Machine Learning Algorithms	Kolla Bhanu Prakash S. Sagar Imambi, Mohammed Ismail, T Pavan Kumar, YVR Naga Pawan May 2020	Random Forest Regressor and Random Forest Classifier	The experiments reveal the persons of age groups 20-30, 30- 40 and 40-50 are suffered with COVID-19.

4.	COVID 19 in India – Statewise Analysis and Prediction	Palash Ghosh, Rik Ghosh, Bibhas Chakraborty August 2020	Logistic and Exponential models	Found that 7 states namely, Maharashtra, Delhi, Gujarat, Madhya Pradesh, Andhra Pradesh, Uttar Pradesh, and West Bengal are in the severe category.
5.	Machine learning – based prediction COVID 19 diagnosis based on symptoms	Yazeed Zoabi, Shira Deri- Rozov, Noam Shomron January 2021	Developed a model that detects COVID 19 cases by simple features accessed by asking basic questions	Predictions from the test set, the possible working points are: 87.30% sensitivity & 71.98% specificity or 85.76% sensitivity & 79.18% specificity
6.	Prediction of epidemic trends in COVID 19 with logistic model and machine learning techniques	PeipeiWang, Xingi Zheng, JiayangLi, BangrenZhu October 2020	Logistic model to fit the cap of epidemic trend, and then feed the cap value into FbProphet model, a machine learning based time series prediction model to derive the epidemic curve and predict the trend of the epidemic.	As the research in this paper shows that a hybrid logistic and prophet model has a valuable advantage in terms of forecasting the epidemic trend

7.	Prediction of respiratory decompensation in Covid-19 patients using machine learning	Hoyt Burdick, Carson Lam, ,Anna Siefkas, Gregory Braden, R. Phillip Dellinger, Andrea McCoy, Jean-Louis Vincent, Abigail Green-Saxena, Gina Barnes, Jana Hoffman, Jacob Calvert, Emily Pellegrini, Ritankar Das September 2020	The model was created using the XGBoost Classifier method for fitting "boosted" decision trees in Python. Gradient boosting, which XGBoost implements, is an ensemble learning technique that combines results from multiple decision trees to create prediction scores.	A machine learning algorithm for prediction of mechanical ventilation of COVID-19 patients within 24 h of their initial hospital encounter demonstrated a high sensitivity (0.90) and specificity (0.58) and outperformed a commonly used early warning scoring system; the algorithm is capable of detecting 16% more patients than the Modified Early Warning Score ($p < 0.05$) while simultaneously reducing false positive alerts. This algorithm may therefore help to improve patient care, minimize clinician burden, and minimize morbidity

				and mortality during the COVID-19 pandemic.
8.	Natural Language Processing and Machine Learning for Detection of Respiratory Illness by Chest CT Imaging and Tracking of COVID-19 Pandemic in the US	Ricardo C. Cury, Istvan Megyeri, Tony Lindsey, Robson Macedo, Juan Batlle, Shwan Kim, Brian Baker, Robert Harris, Reese H. Clark Feb 2021	NLP algorithms	Using big data, we developed a novel machine-learning based NLP algorithm that can track imaging findings of respiratory illness detected on chest CT imaging reports with strong correlation with the progression of the COVID-19 pandemic in the US.
9.	Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results	Jiangpeng Wu, Pengyi Zhang, Liting Zhang, Wenbo Meng, Junfeng Li, Chongxiang Tong, Yonghong Li, JingCai, Zengwei Yang, Jinhong Zhu, Meie Zhao, Huirong Huang, Xiaodong Xie, Shuyan Li April 2020	Random forest algorithm	The method presented robust outcome to accurately identify COVID-19 from a variety of suspected patients with similar CT information or similar symptoms, with accuracy of 0.9795 and 0.9697 for the cross-

				validation set and test set, respectively.
10.	IoT enabled depthwise separable convolution neural network with deep support vector machine for COVID-19 diagnosis and classification	Dac-Nhuong Le, Velmurugan Subbiah Parvathy, Deepak Gupta, Ashish Khanna, Joel J. P. C. Rodrigues & K. Shankar January 2021	Deep support vector machine	The experimental results ensured the superior results of the DWS-CNN model by attaining maximum classification performance with the accuracy of 98.54% and 99.06% on binary and multiclass respectively.
11.	COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification	Jim Samuel, G. G. Md. Nawaz Ali, Md. Mokhlesur Rahman, Ek Esawi, Yana Samuel June 2020	Natural Language Processing (NLP)	Given the easy availability of COVID-19 related big data, an extensive array of analytics and state of the art machine learning driven solutions needs to be developed to address the pandemic's global information complexities. While the current research stream contributes to the strategic process,

				a lot more needs to be done across multiple social media, news and public and personal communication platforms. Such solutions will also be critical in identifying a sustainable pathway to recovery post-COVID-19
12.	Early risk assessment for COVID-19 patients from emergency department data using machine learning	Frank S. Heldt, Marcela P. Vizcaychipi, Sophie Peacock, Mattia Cinelli, Lachlan McLachlan, Fernando Andreotti, Stojan Jovanović, Robert Dürichen, Nadezda Lipunova, Robert A. Fletcher, Anne Hancock, Alex McCarthy, Richard A. Pointon, Alexander Brown, James Eaton, February 2021	Logistic regression, random forest and Extreme Gradient Boosted Trees (XGBoost)	On all three cohorts, our models reach good performance with the best model showing AUC-ROC between 0.76 and 0.87. Overall, machine learning methods can thus reliably predict poor outcomes for COVID-19 patients from early clinical data, available during the ED stay of patients.

Table 2.1 A Survey of datasets and Techniques Used for Analysis

Observation

From the literature survey it is find out there was only less accuracy results using some Machine learning techniques. The object of this work is to solve the above problem and get the best accuracy using machine learning techniques.

CHAPTER 3

METHODOLOGY

The Detection of COVID – 19 data set is taken for the work from the **github data base**. Jupyter Notebook tool is used for Classification of Machine Learning algorithms and detecting covid 19. This dataset is split for training and testing purposes. Test data set is used to check the performance. The performance of different algorithm such as Gaussian Navie bayes, Random Forest Classifier and Linear model – SGD Classifier are compared on the basis of accuracy. **Figure 3.1** shows the frame work of this methodology.

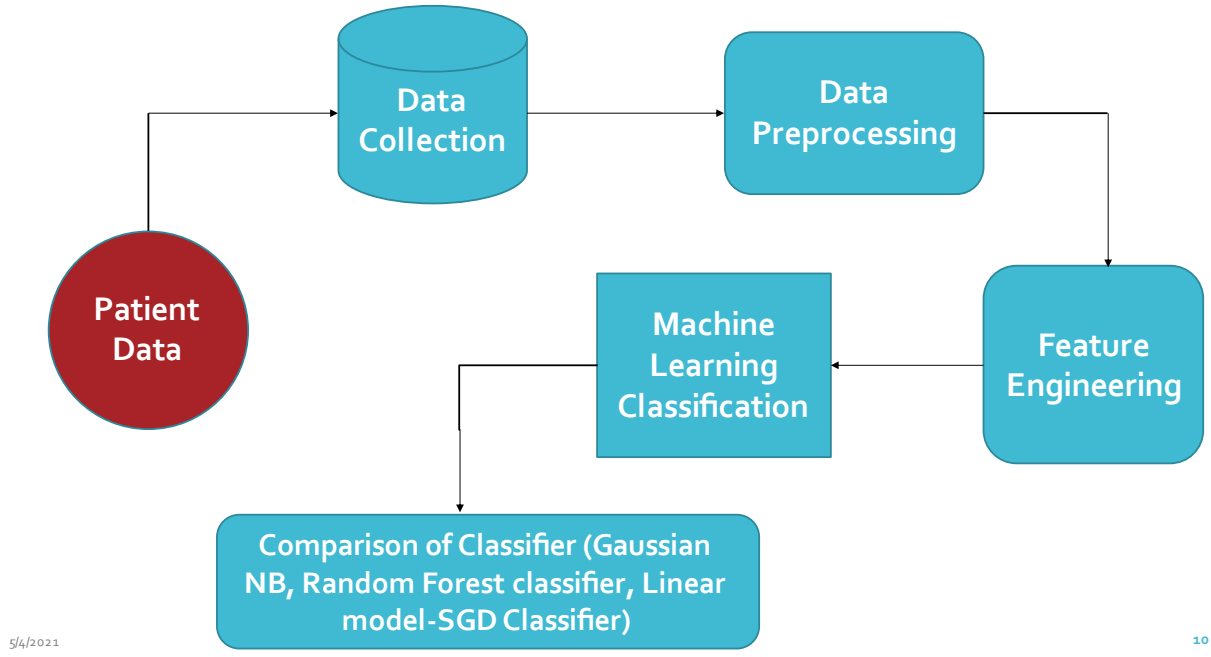

Figure 3.1: Methodology Diagram

The entire methodology is divided into five phases namely, Data collection, Data Preprocessing, Feature Extraction, Machine learning classification using Gaussian NB, Random forest Classifier and Linear model – SGD Classifier. Analysis of data based on clinical reports and Data Visualization done by the Scatter plots. The analysis done using Visualization phases.

3.1 DATA COLLECTION

The Data collection involves two major processes. They are understanding the data set and loading the data set. Understanding the dataset refers to acknowledgement of the information given in the dataset. And loading refers to import the dataset into the Jupyter Notebook. As W.H.O declared Coronavirus pandemic as Health Emergency. The researchers and hospitals give open access to the data regarding this pandemic. The data collected from an open-source data repository GitHub. In which about 316773 patient's data is stored which have shown symptoms of corona virus and other viruses.

Data consists of about 27 attributes namely Temperature, Tiredness, finding, Difficulty-in-Breathing, Sore-Throat, None_Symptom, RT_PCR_positive, Nasal-Congestion, Runny-Nose, Diarrhea, intubated, Age_0-9 Age_10-19, Age_20-24, Age_25-59, Age_60+, Gender, Transgender, Severity_Mild, Severity_Moderate, Severity_None, Severity_Severe, Contact_Dont-Know, Contact_No, Contact_Yes, Country. Data' s in dataset are converted into probabilistic values.

3.2 DATA PREPROCESSING

The dataset is unstructured and generally contains noises, missing values so it needed to be refined such that machine learning can be done. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model. Here removing the unnecessary columns Severity_Mild, Severity_None, Severity_Moderate, Severity_Severe, Contact_No, Contact_Yes, Contact_Dont-Know.

3.3 FEATURE ENGINEERING

Feature engineering is useful to improve the performance of machine learning algorithms and is often considered as applied machine learning. Features are also referred to as 'variables' or

'attributes' as they affect the output of a process. Feature extraction involves choosing a set of features from a large collection. From the preprocessed clinical reports, various features are extracted as per the semantics and are converted into probabilistic values. Here, the extracted features are temperature, Tiredness, finding, Difficulty-in-Breathing, Sore-Throat, RT_PCR_positive, Nasal-Congestion, Runny-Nose, Diarrhea, intubated, Gender.

3.4 MACHINE LEARNING CLASSIFICATION

3.4.1 ALGORITHM FOR GAUSSIAN NAÏVE BAYES

Gaussian Naive Bayes is a variant of Naive Bayes that follows Gaussian normal distribution and supports continuous data. Naive Bayes are a group of supervised machine learning classification algorithms based on the Bayes theorem. It is a simple classification technique, but has high functionality.

Algorithm:

Step 1: Create the training and test data

Step 2: Fit the model on training data and predict dist on test data

Step 3: Review diagnostic measures

Step 4: Calculate prediction accuracy and error rates

3.4.2 ALGORITHM FOR RANDOM FOREST CLASSIFIER

Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. It has nearly the same hyperparameters as a decision tree or a bagging classifier. Random forest adds additional randomness to the model, while growing the trees.

Algorithm:

Step 1: Select random samples from a given dataset.

Step 2: Construct a decision tree for each sample and get a prediction result from each decision tree.

Step 3: Perform a vote for each predicted result.

Step 4: Select the prediction result with the most votes as the final prediction.

3.4.3 ALGORITHM FOR STOCHASTIC GRADIENT DESCENT

Stochastic Gradient Descent Classifier is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x). By calculating accuracy measures (like min_max accuracy) and error rates (MAPE or MSE).

Algorithm:

Step 1: Create the training and test data

Step 2: Fit the model on training data and predict dist on test data

Step 3: Review diagnostic measures

Step 4: Calculate prediction accuracy and error rates

3.5 DATA ANALYSIS

In statistics, data analysis is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily exploratory data analysis is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task. Exploratory data analysis is different from initial data analysis (IDA), which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed.

Accuracy

Accuracy can be estimated using one or more test sets that are independent of the training set. The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. Accuracy refers to the percentage of correct predictions made by the model when compared with the actual classifications in the test data.

3.6 DATA VISUALIZATION

Data visualization is the presentation of data in a pictorial or graphical format.A primary goal of data visualization is to communicate information clearly and efficiently via statistical graphics, plots and information graphics. Numerical data may be encoded using dots, lines, or bars, to visually communicate a quantitative message. Effective visualization helps users analyze and reason about data and evidence. It makes complex data more accessible, understandable and usable. Processing, analyzing and communicating this data present ethical and analytical challenges for data visualization.

3.6.1 THE REQUIRED PACKAGES

matplotlib: This Python package used for data plotting and visualization. It is a useful complement to Pandas, and like it is a very feature-rich library which can produce a large variety of plots, charts, maps, and other visualizations.

sklearn. model_selection: It is a Python library that offers various features for data processing that can be used for classification, clustering, and model selection. Model_selection is a method for setting a blueprint to analyze data and then using it to measure new data.

sklearn. metrics: The module implements several loss, score, and utility functions to measure classification performance. Some metrics might require probability estimates of the positive class, confidence values, or binary decisions values.

CHAPTER 4

EXPERIMENTAL RESULT AND DISCUSSION

4.1 DATA SET

The data collected from an open-source data repository GitHub. In which about 316801 patients data is stored which have shown symptoms of corona virus and other viruses. The Data consists of about 27 attributes.

S. No	Name of the attribute	Type
1.	**Temperature:** If the person temp is high than normal it consider as value 1 otherwise 0	Numeric
2.	**Tiredness:** If the person having tiredness then it consider as value 1 otherwise 0	Numeric
3.	**Findings:** Type of pneumonia, if similar to covid then the value is 1 otherwise 0	Numeric
4.	**Difficulty-in-Breathing:** If yes then the value is 1 otherwise 0	Numeric
5.	**Sore-Throat:** If the person having sore throat then the value is 1 otherwise 0	Numeric
6.	**None_Symptom:** If the person didn't have any symptoms then the value is 1 otherwise 0	Numeric
7.	**RT_PCR_positive:** If yes then the value is 1 otherwise 0	Numeric
8.	**Nasal-Congestion:** If yes then the value is 1 otherwise 0	Numeric
9.	**Runny-Nose:** If yes then the value is 1 otherwise 0	Numeric
10.	**Diarrhea:** If the person having diarrhea then the value is 1 otherwise 0	Numeric
11.	**Intubated:** If yes then the value is 1 otherwise 0	Numeric

12.	**Age_0-9:** If the person is in between the age 0-9 then the value is 1 otherwise 0	Numeric
13.	**Age_10-19:** If the person is in between the age 10-19 then the value is 1 otherwise 0	Numeric
14.	**Age_20-24:** If the person is in between the age 20-24 then the value is 1 otherwise 0	Numeric
15.	**Age_25-59:** If the person is in between the age 25-59 then the value is 1 otherwise 0	Numeric
16.	**Age_60+:** If the person is in between the age 60+ then the value is 1 otherwise 0	Numeric
17.	**Gender_Female:** If Gender is female then the value is 0	Numeric
18.	**Gender_Male:** If Gender is male then the value is 1	Numeric
19.	**Gender_Transgender**	Numeric
20.	**Severity_Mild:** If yes then the value is 1 otherwise 0	Numeric
21.	**Severity_Moderate:** If yes then the value is 1 otherwise 0	Numeric
22.	**Severity_None:** If yes then the value is 1 otherwise 0	Numeric
23.	**Severity_Severe:** If yes then the value is 1 otherwise 0	Numeric
24.	**Contact_Dont-Know:** If don't know then 1	Numeric
25.	**Contact_No:** If contact number not having then the value is 1 otherwise 0	Numeric
26.	**Contact_Yes:** If contact number having then the value is 1 otherwise 0	Numeric
27.	**Country**	String

Table 4.1 Data Description

4.2 TRAINING AND TESTING

Separating data into training and testing sets is an important part of evaluating data mining models. Typically, when separate a dataset into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing. Analysis services randomly samples the data to ensure that the training and testing sets are similar. After a model has been processed by using the training set, test the model by making predictions against the test set. Because the data in the testing set already contains known values for the attribute that want to predict, it is easy to determine whether the model's guesses are correct. In the classification, the dataset is split as 70% of training data and 30% of testing data.

Dataset	Number of Rows
Detection of COVID-19 Dataset	3,16,801
Training set (70%)	2,21,760
Test set (30%)	95,040

Table 4.2 No of records in dataset

The above table describes the number of instances on overall data set, training set and test set. Out of 3,16,801 total records, the training set contains 70% of data that is 2,21,760 instances and 30% of records are taken for test set. First we train the Gaussian Naïve Bayes algorithm for testing the accuracy. Then move on to the next algorithms such as Random Forest classifier and SGD classifier.

4.3 IMPLEMENTING AND INTERPRETING THE RESULTS

Importing required libraries and Data Preprocessing

First, importing libraries which are required for this project and also importing the dataset 'covid.csv'. Then, displaying the head of the covid dataset. And then data preprocessing is done. In data preprocessing, the dataset is cleaned. After data preprocessing, the feature extraction process takes place. Here, the features extracted are temperature, Tiredness, finding, Difficulty-in-Breathing, Sore-Throat, RT_PCR_positive, Nasal-Congestion, Runny-Nose, Diarrhea, intubated, Gender. Then it continued to the classification of machine leaning algorithms.

Figure 4.1 Importing required libraries and 'covid.csv' dataset

In [8]:
```
df.head()
```

Out[8]:

	temperature	Tiredness	finding	Difficulty-in-Breathing	Sore-Throat	None_Sympton	RT_PCR_positive	Nasal-Congestion	Runny-Nose	Diarrhea	...	Age_60+	Gender_Female	Gender_
0	1	1	1	1	1	0	1	1	1	1	...	0	0	
1	1	1	1	1	1	0	1	1	1	1	...	0	0	
2	1	1	1	1	1	0	1	1	1	1	...	0	0	
3	1	1	1	1	1	0	1	1	1	1	...	0	0	
4	1	1	1	1	1	0	1	1	1	1	...	0	0	

5 rows × 25 columns

Figure 4.2 Displaying head of dataset 'df.head()'

jupyter covid_prediction Last Checkpoint: 04/20/2021 (autosaved)

File Edit View Insert Cell Kernel Widgets Help — Not Trusted — Python 3 — Logout

Run — Code

In [16]:
```
df = df.drop(['Severity_Mild'],axis = 1)
df = df.drop(['Severity_None'],axis = 1)
df = df.drop(['Severity_Moderate'],axis = 1)
df = df.drop(['Severity_Severe'],axis = 1)
df = df.drop(['None_Sympton'], axis=1)
#df = df.drop(['None_Experiencing'], axis=1)
df = df.drop(['Contact_No'],axis = 1)
df = df.drop(['Contact_Yes'], axis =1)
df = df.drop(['Contact_Dont-Know'], axis =1)
```

In [17]:
```
df.insert(16, "symptomes", symptomes, True)
```

In [18]:
```
len(df.columns)
def acc_score(y,Y):
    set_value = y
    preset_value = Y
    return accuracy_score(y,Y)*thresh_val
```

In [19]:
```
df.head(100)
```

Out[19]:

- - g	Sore-Throat	RT_PCR_positive	Nasal-Congestion	Runny-Nose	Diarrhea	intubated	Age_0-9	Age_10-19	Age_20-24	Age_25-59	Age_60+	Gender_Female	symptomes	Gender_Male
1	1	1	1	1	1	0	1	0	0	0	0	0	0	1
1	1	1	1	1	1	0	1	0	0	0	0	0	0	1
1	1	1	1	1	1	0	1	0	0	0	0	0	0	1
1	1	1	1	1	1	0	1	0	0	0	0	0	1	1
1	1	1	1	1	1	0	1	0	0	0	0	1	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1	1	0	0	1	1	0	1	0	0	0	0	0	0	1
1	1	0	0	1	0	0	1	0	0	0	0	0	0	1
1	1	0	0	1	0	0	1	0	0	0	0	0	0	1

Figure 4.3 Data Preprocessing

Fig 4.1, 4.2 & 4.3 shows the importing libraries and dataset, displaying head of the dataset and data preprocessing respectively.

Gaussian Naïve Bayes

The Gaussian NB is defined lr() function. Train data first perform the model and followed by the test data perform. And calculate the accuracy.

```
jupyter  covid_prediction Last Checkpoint: Last Tuesday at 12:43 PM  (autosaved)                Logout
File  Edit  View  Insert  Cell  Kernel  Widgets  Help                      Not Trusted  | Python 3
                                                    Run   Code

In [23]: from sklearn.naive_bayes import GaussianNB

In [24]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.1, random_state=1)

In [25]: lr = GaussianNB()

         lr.fit(X_train, y_train)

Out[25]: GaussianNB()

In [26]: y_lr = lr.predict(X_test)

In [27]: # Gaussian Native bayes Result

In [28]: # Importation des méthodes de mesure de performances

         target_names = ['0','1','2']
         classification_report(y_test,y_lr, target_names=target_names)

         g:\movies\archive (1)\covid\lib\site-packages\sklearn\metrics\_classification.py:1245: UndefinedMetricWarning: Precision and F-
         score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this b
         ehavior.
           _warn_prf(average, modifier, msg_start, len(result))
         g:\movies\archive (1)\covid\lib\site-packages\sklearn\metrics\_classification.py:1245: UndefinedMetricWarning: Precision and F-
         score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this b
         ehavior.
           _warn_prf(average, modifier, msg_start, len(result))
         g:\movies\archive (1)\covid\lib\site-packages\sklearn\metrics\_classification.py:1245: UndefinedMetricWarning: Precision and F-
         score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this b
         ehavior.
           _warn_prf(average, modifier, msg_start, len(result))

Out[28]: '              precision    recall  f1-score   support\n\n           0       0.50      1.00      0.67     15791\n           1
         0.00      0.00      0.00      7992\n           2      0.00      0.00      0.00      7897\n\n    accuracy
         0.50     31680\n   macro avg       0.17      0.33      0.22     31680\nweighted avg       0.25      0.50      0.33     31680\n'
```

Figure 4.4 Gaussian Naïve Bayes Classification report

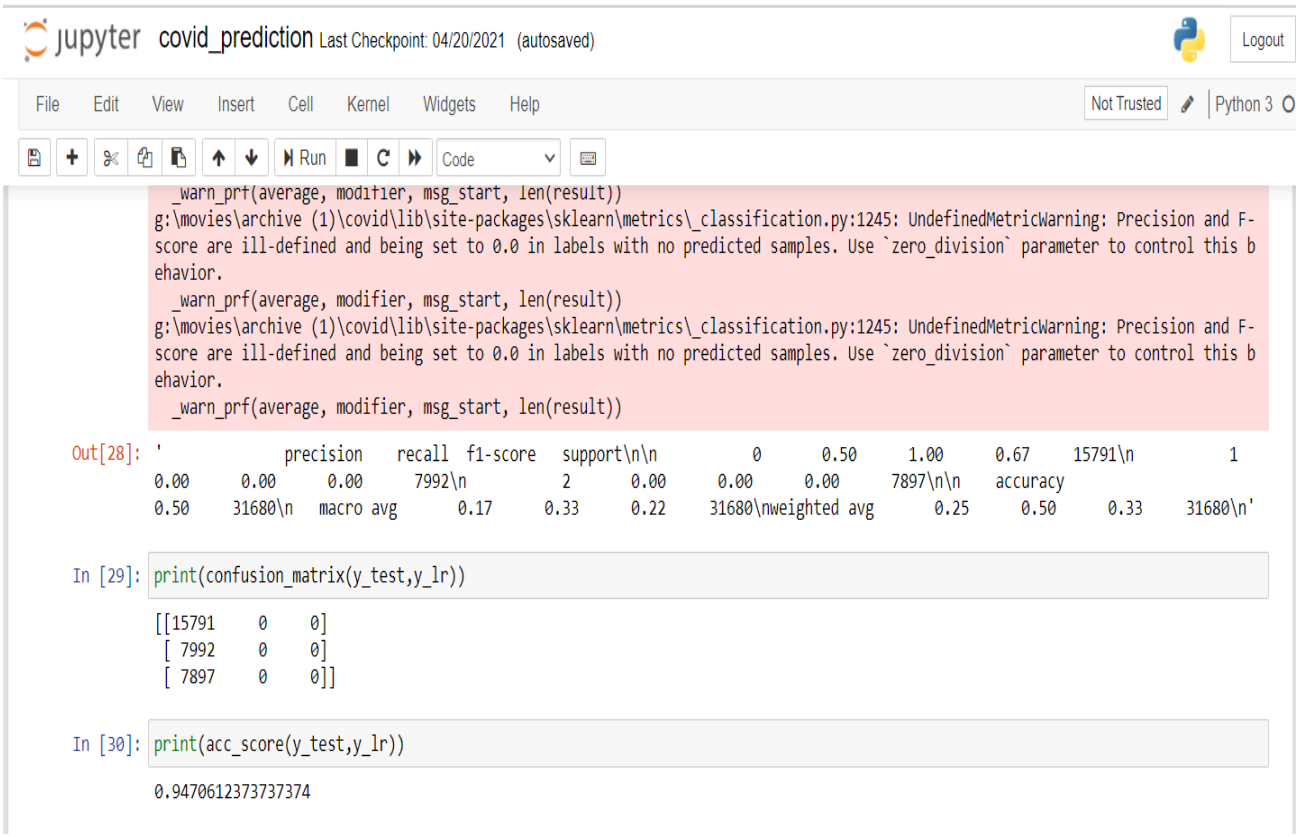

Figure 4.5 Gaussian Naïve Bayes acc_score

Fig 4.4 & 4.5 shows that the Gaussian Naïve Bayes accuracy with 94% comes as output using the code acc_score(y_test, y_lr). This accuracy tells us that 94% is correct result for detection of covid-19 while using this algorithm.

Random Forest Classifier

The Random forest is defined rf() function. Train data first perform the model and followed by the test data perform. And calculate the accuracy. Then the scatterplot shows the covid-19 detection.

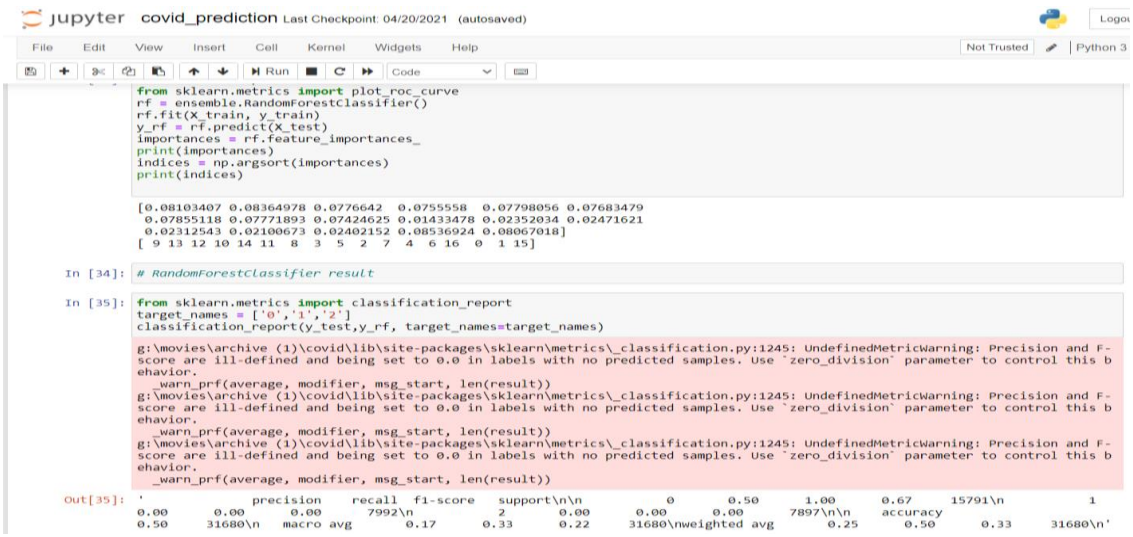

Figure 4.6 Random Forest Classifier classification report

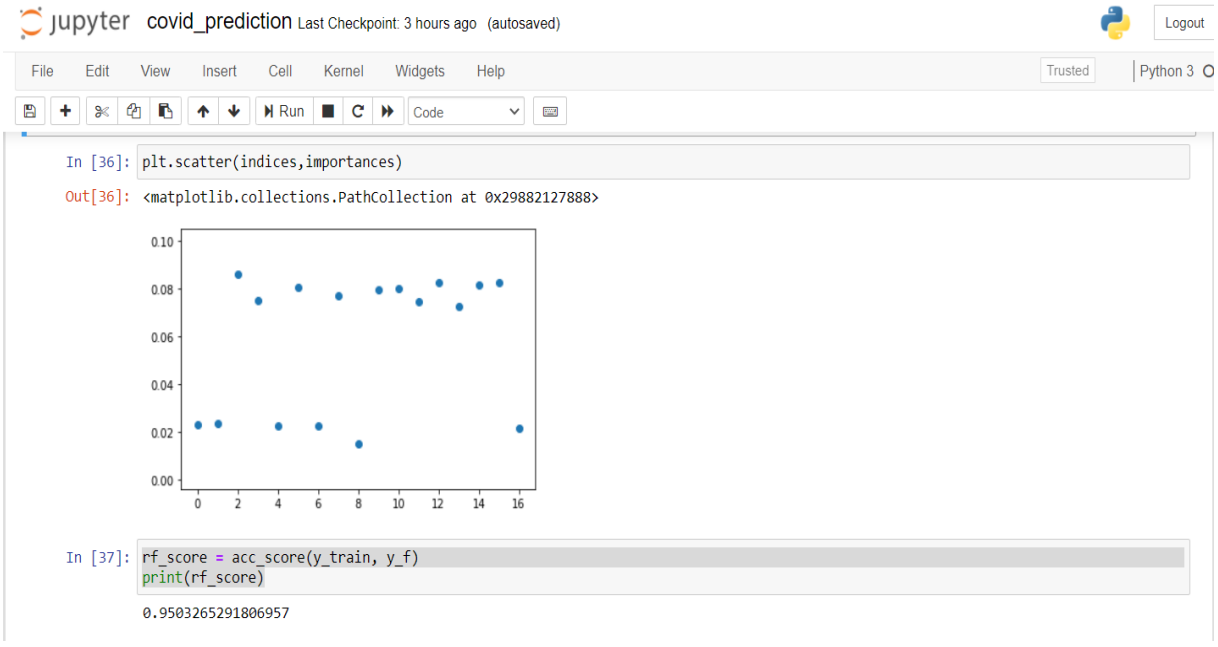

Figure 4.7 Random Forest classifier – Scatterplot & acc_score

Fig 4.6 & 4.7 Shows that the Random Forest classifier accuracy with 95% comes as output using the code acc_score(y_train, y_f). This accuracy tells us that 95% is correct result for detection of covid-19 while using this algorithm. Also the scatterplot tells that scattered area in the plot gives confirm of covid-19.

Stochastic Gradient Descent Classifier

The Stochastic Gradient Descent Classifier is a linear model and is defined as lr() function. Train data first perform the model and followed by the test data perform. And calculate the accuracy. Then the scatterplot shows the covid-19 detection.

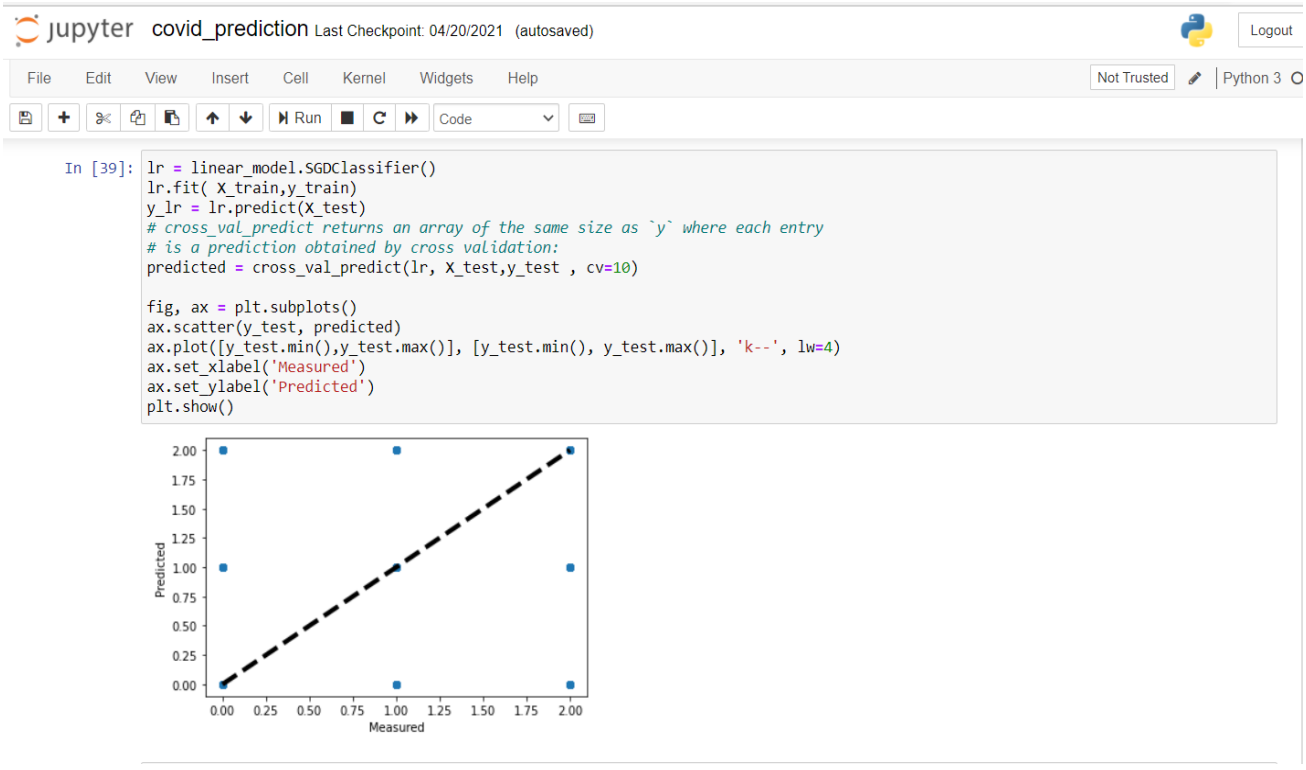

Figure 4.8 Stochastic Gradient Descent Classifier - scatterplot

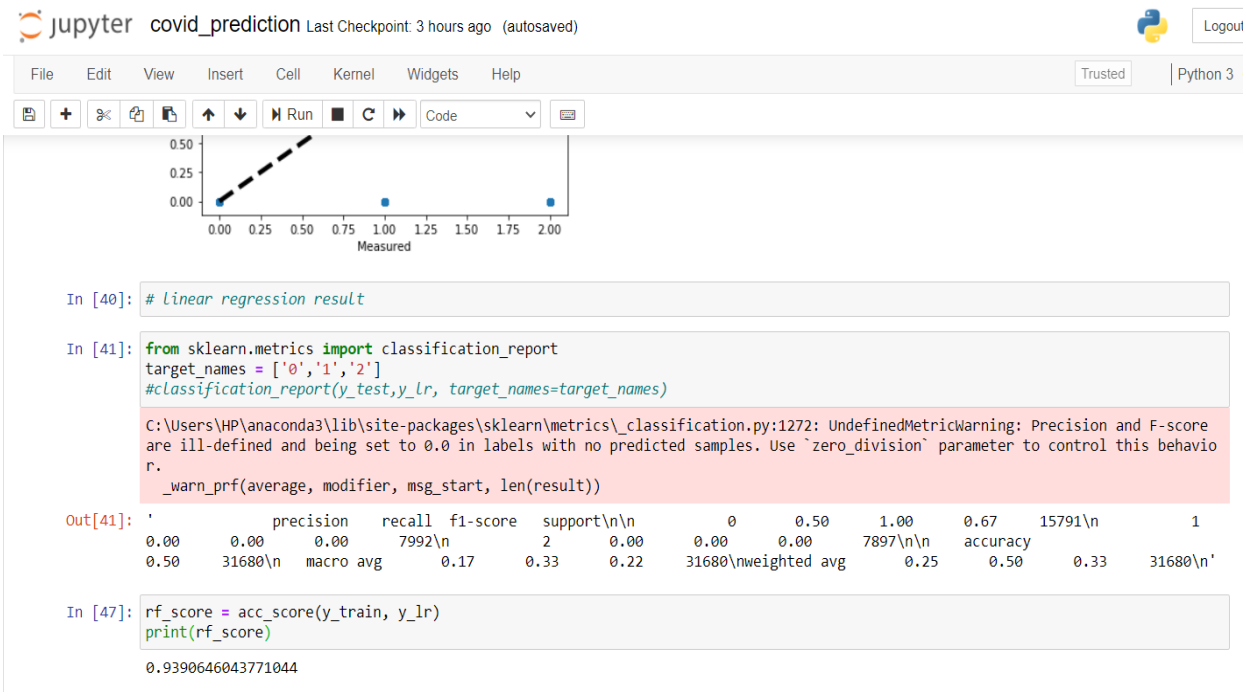

Figure 4.9 Stochastic Gradient Descent classification report & acc_score

Fig 4.8 & 4.9 Shows that the Stochastic Gradient Descent Classifier accuracy with 93% comes as output using the code acc_score(y_train, y_lr). This accuracy tells us that 93% is correct result for detection of covid-19 while using this algorithm. Also in the scatterplot tells that above the line gives us the predicted cases of covid-19.

4.3 PARAMETERS FOR ANALYSIS

The following parameters are used to:

- Predictive accuracy: It is defined as the percentage of correct prediction made by a classification algorithm.

4.3.1 PREDICTIVE ACCURACY

- Predictive accuracy is expressed as the correlation between the prediction and actual score. Accuracy is often the starting point for analyzing the quality of predictive model.

Machine Learning Classification	Accuracy
Gaussian Naïve Bayes	94%
Random Forest Classifier	95%
SGD Classifier	93%

Table 4.3 Table for accuracy of algorithms

The above table describes the accuracy value of three Machine Learning classification algorithms on Detection of COVID-19 data set. The accuracy level is calculated by using the formula:

Accuracy = (TP+TN)/TP *100

TP- True positive

TN- True negative

CHAPTER 5

CONCLUSION

COVID-19 has shocked the world due to its non-availability of vaccine or drug. Various researchers are working for conquering this deadly virus. The machine learning algorithms are used for classifying clinical reports. After performing classification, it was revealed that all three algorithms such as Gaussian NB, Random Forest Classifier and Linear regression SDG classifier gives the output as precision 94%, 95% and 93% respectively. So, here in this project the conclusion is that among the three classifiers, the Random forest classifier gives the best accuracy result.

CHAPTER 6

SCOPE FOR FUTURE ENHANCEMENT

The efficiency of models can be improved by increasing the amount of data. Also, the disease can be classified on the gender-based such that to get information about whether males are affected more or females. Additional features can be added for improving the results and deep learning approaches can be used in future.

CHAPTER 7
REFERENCES

[1] Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML, Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, Zhang YZ (2020) A new coronavirus associated with human respiratory disease in china. Nature 44(59):265–269

[2] World health organization: https://www.who.int/new-room/g-adetail/q-a-corronaviruses#:/text=symptoms. Accessed 10 Apr 2020 5.

[3] Wikipedia coronavirus Pandemic data: https://en.m.wikipedia. org/wiki/Template:2019%E2%80%9320_coronavirus_pan demic_data. Accessed 10 Apr 2020

[4] Medscape Medical News, The WHO declares public health emergency for novel coronavirus (2020) https://www.medscape. com/viewarticle/924596

[5] Description of Boosting Algorithm: https://towardsdatascience. com/boosting. Accessed 10 July 2019

[6] Katuwal R, Suganthan PN (2018) Enhancing Multi-Class Classification of Random Forest using Random Vector Functional Neural Network and Oblique Decision Surfaces, Arxiv:1802.01240v1

[7] Friedman JH (2002) Stochastic gradient boosting. Comput. Stat. Data Anal. 38(4):367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

[8] Accuracy, Recall, Precision, F-Score & Specificity https://towardsdatascience.com/accuracy-recall-precision-f-score-specificity-which-to-optimize-on-867d3f11124

[9] Ayyoubzadeh SM, Ayyoubzadeh SM, Zahedi H, et al. Predicting COVID-19 incidence through analysis of Google trends data in Iran: data mining and deep learning pilot study. JMIR Public Health Surveill. 2020;6(2):e18828.

[10] Daniel R, Schrider A, Kern D. Supervised machine learning for population genetics: a new paradigm. Trend Genet. 2018;34–4:301–12.

[11] Verma P, Khanday AMUD, Rabani ST, Mir MH, Jamwal S (2019) Twitter Sentiment Analysis on Indian Government Project using R. Int J Recent Tech Eng. https://doi.org/10.35940/ijrte.C6612.098319

[12] Chakraborti S, Choudhary A, Singh A et al (2018) A machine learning based method to detect epilepsy. Int J Inf Technol 10:257–263. https://doi.org/10.1007/s41870-018-0088-1

[13] Sarwar A, Ali M, Manhas J et al (2018) Diagnosis of diabetes type-II using hybrid machine learning based ensemble model. Int J Inf Technol. https://doi.org/10.1007/s41870-018-0270-5