**An efficient technique for Disease prediction from medical data using data mining algorithm**

**Abstract:**

The purpose of this study was to introduce a technique for disease prediction by using data mining algorithm. Novel Corona virus (COVID-19) pandemic has created chaos on the world. The people of developed countries and developing countries are facing many death tolls due to insufficient way to prevent COVID-19. It is now clear that the environment requires a fast and effective way to control and combat the spread of COVID-19 around the world using non-clinical methods such as data mining approaches. To better reduce the massive pressure on the healthcare system while ensuring the enhanced considerably medium for patients' detection and prognosis of the 2019-nCoV pandemic. In this study, using an epidemiological dataset of COVID-19, data mining models were built in this analysis to forecast the prediction of COVID-19 infected patients. To construct the models, the decision tree, logistic regression and random forest algorithms were implemented directly on the dataset using the Python programming language. The results shows that the model developed with Random Forest data mining algorithm is more efficient to predict the possibility of infected patients from COVID-19 with an accuracy of up to 80%.

|  |
| --- |
| **Chapter 1: Introduction** |

**1.1 Motivation**

The World Health Organization (WHO) has named the COVID-19 virus as a pandemic, with over fourteen million outbreaks and 3116444 deaths worldwide as of 27 April,2021[2]. COVID-19 is triggered by the SARS Coronavirus 2 (SARS-CoV-2) and was deemed a pandemic by WHO on March 11, 2020. The treatment for COVID-19 could be more postponed owing to the virus's potential genetic mutations [3]. The pandemic crisis affects billions of citizens on a societal, economic, and medical level, causing dramatic shifts in social interactions, health systems, commerce, employment and educational environments. The global pandemic is a challenge to human civilization, and urgent intervention is needed. In order to prevent more people from dying, the scientific community has been brainstorming solutions to better limit the epidemic and deter possible pandemics. We have tried to come up with a solution to predict a person being infected with COVID-19 based on the symptoms and ages. So, the goal of our thesis is to help the healthcare system with the non-clinical approach. So that we could prevent more people from dying due to this pandemic.

**1.2 Objectives**

The main goal of this thesis is to find an efficient way to predict if the patient is infected with COVID-19 or not. For this, a model had to be created using various data mining algorithm which is used in machine learning. To fulfill our objective, various steps had to be taken. Those steps are:

1. To understand the problem in the existing models and find out about the issues that caused for it.
2. Stating problems for determining a COVID infected patient.
3. Propose model for the prediction of the disease.
4. Proposing solutions for the problem.
5. Finally, developing an efficient model for the solution.

**1.3 Justifications**

The main purpose of this thesis is to provide an efficient technique to predict a disease (COVID-19). As novel corona virus pandemic is one of the biggest issues in the recent time, it has affected all sectors such as societal, economic, health systems, commerce, employment and educational environments. By conducting our research, we will be able to find out the efficient way to find the solution to predict corona virus infected patients. In the end, we will propose a model which will efficiently predict the corona virus disease. We will also analyze the symptoms of the affected patients to improve our model. Our proposed model will ensure the efficiency and accuracy for detecting the corona virus disease. This will provide a beneficial improvement on the research and also provide a solution for the recent major problem of the world.

**1.4 Background Study**

**1.4.1 Novel Corona Virus**

The world's biggest modern outbreak was uncovered in the mid-twentieth century, and since then, the world's populace has been haunted by the Severe Acute Respiratory Syndrome Coronavirus, or literally, as the citizens affectionately term it, "Covid-19." The negative effects of this disease have harmed the psychological tranquillity of the whole world. People are combining their skills to develop numerous ideas and techniques for constructing a network of machines with sensors and sensors in the cloud to monitor various data streams. Previous studies to see whether an individual had this drug in their blood were costly and time-consuming, but they were not effective. According to this rationale, since it is important to reduce death rates and classify high-risk individuals through blood tests at a quicker pace, experts are discussing novel ways of recognizing high-risk patients.

The viruses in the families Coronaviridae and Betacoronaviridae live as infectious agents in many animal species; when they mutate, they can be viruses causing disease in mammals [4]. These respiratory infection outbreaks have been triggered by SARS in 2002, MERS in 2016, and a new respiratory infection, rhinovirus, in 2018. CoVID-19 is the illness transmitted by the SARS-CoV2 virus. It is a highly contagious virus that has become the world's most deadly. It is also known as SARS-CoV2. It is also affected by SARS-CoV2, a huge satellite CoV that triggers the disease. Unlike the previous events, it induces extreme lower respiratory collapse and can affect the central nervous system in the early stages [5]. This virus, like other members of the Coronavirus family, has a sheath called the shell in its fat structure, which has spiky protrusions of its protein structure on its outer surface. The spiky protrusions on the "corona" are referred to as a cap. The leaf is referred to as a "crown," and may also refer to the fingertips. The spiky proteins of the SARS-CoV2 virus vary from those of the SARS virus by 2% and have even greater human cell adhesion [6]. The virus is a non-living pathogen agent because of its nucleic acid chain (a kind of helical amino acid chain that holds genetic codes) in the envelope. When a virus infects another cell, it can initiate replication, cause harm by reproduction, and breed the virus in the process. The exterior shell must be sturdy in order for the SARS-CoV2 virus to enter cells. Too much soap or detergent, like too much bleach, breaks the envelope and deactivates the infection. If it does not enter the cell's body but does solidify into a bumper in the envelope, it may be infectious for a few days. The virus deforms and loses its ability to infect over time. When a large quantity of SARS-CoV2 reaches the human body, the infection starts when the virus clings to the surface counterpart (a skin-like yet much thinner single layer of cell protecting the inner cavities of our bodies) and then infiltrates the cell [7].

SARS stands for severe acute respiratory syndrome. Coronavirus two (SARS-CoV-2), the causative expert of novel (COVID-19 or 2019-nCoV), emerged in late 2019 and is thought to have originated in Wuhan, Hubei Province, China. [8, 9] Formalized adverbial adverbial SARS-CoV-2 is the name given to the virus that triggers COVID-19 scourge illness [10]. Covids (CoV) are a large community of pathogens that trigger illnesses such as Middle East Respiratory Syndrome (MERS-CoV) and severe acute respiratory syndrome (SARS) (SARS-CoV). Coronavirus is another strain discovered in 2019 that has yet to be identified in humans [11]. "2019 innovative coronavirus," which is actually a modern coronavirus that has not been described before. As time passed, health authorities named it as COVID-19, and it is now linked to SARS-CoV-2. Similarly, to how the general population has been acquainted with this language, authorities have started to use SARS-CoV-2 in conjunction with the latest epidemic. The virus that triggers COVID-19 is known as SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2). This virus is genetically similar to the SARS-associated coronavirus (SARS-CoV), which triggered an epidemic of extreme acute respiratory syndrome (SARS) in 2002-2003, although it is not the same virus. Coronavirus has already reached a tipping point and pandemic potential, killing many people around the planet, and human-to-human transmission of COVID-19 by infected people with mild symptoms has been recorded [12].

More than 42% of adult patients with a final diagnosis are between the ages of 18 and 40. A closer glance at the percentage of people afflicted in a given age demographic reveals that the number of those diagnosed with this condition increases significantly as persons mature. In this study, researchers discovered that the condition is more acute and tends to have more effects in individuals who have more significant medical conditions, such as cardiac and renal diseases, tumors, or where patients are suffering immunosuppression, as in previous reports. Patients of this condition are often under the age of 55 and under the age of 30. 4 percent were between the ages of 45 and 55 [13]. As a result, it would be misleading to believe that it just impacts the elderly. On the other side, it is more common in older persons, particularly if they are already afflicted with a disease.

According to the CDC, 2.4 percent of adolescents living with the condition have it, equivalent to 7.9 percent of adults above the age of 18 [14]. This is because there are more instances that go unnoticed because the condition is much milder in children. That is largely attributed to the reality that infants are often vaccinated as a form of disease prevention. As mentioned in the quotation, pregnant women's lack of direct interaction with their unborn child can play a role in the natural protection they now have. Even though researchers have not decided how this operates, every mother and child who has been exposed to hCMV should be aware of the situation and take precautions to shield themselves from possible injury.

Men are more susceptible to the disease than women. Men have a greater chance of tobacco addiction, which contributes to their higher mortality rate. According to research, this may be attributed to the reality that men have these lifelong diseases and that men are more prone to smoke as a habit. It is known that estrogen, hormones generated by female glands, has an impact on the brain in the same way that estrogen does in males. Furthermore, women have on infections than men and they have more of the immune system-related genes present on the X chromosome, which is the female chromosome. Since the X chromosome has just one copy, there are only two females in the world that have the same number of X chromosomes [15].

According to studies, the virus is transmitted by airborne droplet exposure, but it is more likely spread by contacting polluted materials. Furthermore, although it has been demonstrated that the live virus is excreted from the human body through diarrhoea within 12 days, it has not been shown (or even demonstrated) that it is transmitted in this manner [7]. Another major problem is that the live virus has not been detected in blood or urine. Finally, the live virus may not have any infectious properties. Furthermore, whether the disease is transmitted by blood, it has not yet been shown that it is caused by variations in the environment or temperature. The illness may be passed from person to person, particularly though the recipient has not yet been infected. There is little consensus on the effect of the climate and environmental conditions on the disease's contagiousness. Several experiments have shown that elevated air temperature and humidity reduce the rate of disease transmission. Some experts, however, reject this point, claiming that it would not be successful in treating the disease [16].

SARS, or serious acute respiratory syndrome, is a major condition. SARS started in China in 2003 and spread to two other countries before ceasing in 2004; it was eradicated after a few years [17]. The strain that triggers this epidemic seems to be similar to the SARS epidemic. They are both coronaviruses, and are most likely triggering the foreign body reaction. It is still unclear; however, COVID-19 could spread quicker than SARS in 2003, and it is also likely to trigger less severe illness than other disease epidemics.

**1.4.2 Symptoms:**

The incubation duration is the interval between the entry of the virus into the body and the start of symptoms, which is typically four to seven days in the majority of people but can extend from two to fourteen days if the COVID-19 virus is infected. 10 Symptoms include fever in 80 to 90% of cases, cough in 60% to 70% of cases, gastrointestinal eruptions in 40% to 50% of cases, lack of smell in 30% to 40% of cases, and shortness of breath in 20% of cases [18]. Not all of these symptoms are common in these people. First, it is predicted that an individual will become sick with the disease but will not show symptoms for years. Often after an individual becomes ill does the entire scope of the condition become apparent. Some manifestations of this disease only benefit such that they may not present with fever or respiratory problems, and in these cases, the patient can take longer to recover. The most prominent signs include a burning fever, shortness of breath, and shortness of breath. Beneficiaries of our treatment rebound in around a week whether they are not suffering respiratory problems. Due to the patient's fever, it can take up to two weeks for the patient to heal.

Fever, dry cough, and tiredness are the most frequent signs of corona virus around the world. Aches and pains, sore throat, diarrhoea, conjunctivitis, fever, lack of taste or scent, a rash on the scalp, or discoloration of fingers or toes are less frequent signs of the infection. This infection causes serious symptoms such as trouble breathing or shortness of breath, chest discomfort or strain, and lack of voice or movement.

The manifestations of the disorder can take four distinct types based on the indications [19]. The most difficult ones are those of almost non-existent signs: People are hardly aware of the problems and continue to breathe normally. Moderate cases have patients that have a fever and perhaps have x-ray signs with pneumonia. This disease has affected about 80% of all patients in the mild to severe range.

Patients with a resting respiratory rate of more than 30 breaths per minute and a concentration of less than 93 percent are considered extreme cases. When pneumonia worsens, the radiologic signs increase more than twice in fewer than two days. 14 percent of the patients in the group are deemed malignant.

Ventilators are used in the ICU in cases of high danger. Shock and/or multiple organ dysfunction begin to occur in the patient. Although most instances of this condition are now uncommon, this number was low several years ago. It takes 5–6 days on average after becoming infected with the virus for signs to appear, although it may take up to 14 days.

**1.5 Data mining**

The advancement in information technology has resulted in the development of a massive array of databases, archives, and documents in various fields. Database and computer management research has yielded a tool for preserving and manipulating this important data for future decision-making. Data mining is a technique for collecting useful information and trends from large volumes of data. This approach is also known as information discovery, data processing, intelligence retrieval, and data/pattern study.

Data mining is a computerized technique that can be used to sift through vast amounts of data in order to extract useful and constructive facts. The aim of this technique is to detect previously unknown trends. If these patterns are identified, their success as a metric is used to determine what types of goods the business can pursue in the future. Investigating the issues is done first and then the similarities inside them.

**Exploration:** During the first step of data exploration, data is cleaned and converted into a different format, and essential variables and the quality of data are determined depending on the challenge.

**Pattern Recognition:** Once the raw data has been properly interpreted and refined into a more systematic type, the next move is to form a pattern of relations between the findings. Sort the data sets (or inputs) that can be used for prediction.

**Pattern deployment:** Patterns are applied based on expected results.

Data mining provides impressive results in disease identification where the proper tools and techniques are utilized. As a consequence, data mining is a promising field in healthcare forecasting. Data mining has played an important role in the healthcare industry, especially in disease prediction. Researchers must develop hybrid models to enhance the prediction in order to obtain the highest forecast precision. Growing the size of the data set would therefore improve the reliability of the findings.

**1.5.1 Data mining algorithms**

There are various types of data mining algorithm. A few of the algorithms which are used for prediction are stated below.

**Logistic Regression (LR)**

Logistic Regression is used to evaluate the relationship between the dependent variable of existing and the independent variable of the category or rank [20].

Logistic regression, which is best known in its ANOVA form, is commonly used to study correlations between independent and categorial dependent variable variables. LR is a type of likelihood ratio test in which the dependent variable has two values such as zero and one, yes and no or true and false and hence it is called binary logistic regression [21]. However, when the dependent variable may have two or more values, then multinomial logistic regression is used for statistical analysis. A mathematical model of the set of independent variables for LR is used to produce a desired predicted output by predicting the dependent variable.

**Support Vector Machine (SVM)**

One of the supervised learning algorithms used for classification and regression is the Support Vector Machine (SVM) [22]. For classification tasks in SVM, testing and training data with some instances of the data are required [23]. Because each instance in the training dataset includes one or more target values, the primary objective of SVM is to create a model that predicts the target value or values [22]. SVM is used for regression by using alternate loss functions, which may be linear or nonlinear [23].

**Decision Tree (DT)**

Decision Tree (DT) is a popular data mining strategy because of its capacity to manage both categorical and continuous data, flexibility, and comprehensibility. DT divides the tree into phases that involve growth and pruning phases [24]. A tree is constructed in the first step by partitioning data into smaller sets until each partition is pure; however, the split form of the data is solely based on the data type [25]. The splits for a numerical attribute C result in the value of (C) y, where y is a value in the C domain. Shape the values of (D), BG, where G is a subset of domain (D), to break a categorical D [26]. The pruning technique is used to have the nal tree installed when it is completely developed to eliminate noise in the dataset [27]. However, the development process of the decision tree is computationally more costly than the pruning phase [28].

**Naive Bayes (NB)**

Naive Bayes is a data mining classification algorithm that is used to differentiate dataset instances based on listed features or attributes [29]. NB is a probabilistic classifier that employs the Bayes principle to perform classification tasks [30].

**Random Forest (RF)**

The Random Forest (RF) algorithm is an ensemble learning method for classification and regression tasks in data mining. During preparation, the algorithm generates a large number of decision trees, which it then outputs [31]. The RF data mining algorithm is the strongest for any decision tree that has obvious links to its training dataset [29].

**K-Nearest Neighbour (K-NN)**

K-Nearest Neighbour (K-NN) is a supervised and non-parametric data mining classifier that is used for regression and classification tasks [32]. The input variables in both tasks are the K closes training dataset in the function room. K-NN relies on labelled input data to learn a feature that produces acceptable performance when unlabelled data is inputted [33]. The performance of K-NN classification is a class membership in which data instances are categorized by a majority vote of its neighbours, with the data instance being allocated to the class most popular among its K-nearest neighbours, while the output of K-NN regression is the property value of data instance and this value is the average of the value of K-nearest neighbours [34].

Prediction may be accomplished using the regression technique. To model the relationship between one or more independent variables and one or more dependent variables, regression analysis may be used. Independent variables are proven attributes in data mining, and answer variables are what we want to forecast. Unfortunately, several real-world issues are unpredictable. Sales amounts, market values, and product failure rates, for example, are all impossible to estimate due to dynamic relationships with various predictor variables. As a consequence, more sophisticated forecasting techniques (such as logistic regression, decision trees, or neural nets) might be needed. The same model styles are often used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm may be used to construct both classification trees (to define categorical answer variables) and regression trees (to forecast continuous response variables). Classification and regression models may also be produced by neural networks.

**1.6 Research Methodology**

This paper is based on solution-based approach, focusing on creating a model and an efficient way to predict covid 19 from a patient. The study follows both open-source data collection on the internet. The process was fully participatory ensuring the efficiency in the model. The research considered as systematic review research methodology. The necessary steps to perform the research are presented below.

**1.6.1 Problems of clinical methods**

A blood examination or an antibody test may be used to assess if anyone is afflicted with the Virus. A diagnostic examination will diagnose an infection if anyone is already infected. The FDA has authorized several diagnostic tests for covid-19, including molecular tests [35].

Molecular experiments scan for the virus's genetic material in a human sample. This sample is then usually analysed in a laboratory. On certain samples, the material may be analysed at the point of diagnosis, such as a doctor's clinic. The FDA-approved molecular studies are extremely accurate, and the findings are returned in a matter of minutes to several days, depending on the test.

Another method of diagnostic examination is antigen checking, which uses a swab to see whether viral proteins are found in a sample obtained from within the nose. These experiments are frequently easier and can have findings quicker than certain molecular tests, often within minutes in a doctor's office. While antigen tests are less reliable and can need clarification with another diagnostic test, they can be made more readily accessible due to their ease of use and simplicity.

The other kind of test is an antibody test, also known as a serology test. When a person is infected with a virus, the body produces antibodies that aid the immune system in fighting the infection. An antibody screening identifies antibodies to the virus using a blood sample. If an antibody examination detects antibodies in the blood, it is likely that the person has previously been infected with the virus. Antibody testing cannot tell you whether you have a new infection and cannot be used to detect a Covid-19 infection. The test results for antibodies developed in response to Covid-19 infection will help us further understand questions about virus exposure by identifying people who may be infected with the virus, those who have developed antibodies, whether the antibodies may protect those who are infected from more infection, and whether those who may be infected may be able to donate blood.

**Limitations of Molecular test:**

A molecular diagnosis can be made quickly and provides highly sensitive, accurate, and usually quantitative identification of the SARS-CoV-2 Virus RNA. However, it is difficult, expensive, and time-consuming to implement. A single DNA-based research package cost more than $100. The processing time is 2 to 3 days until the lab is prepared. The test kit study, on the other hand, requires 2 to 3 hours. Furthermore, the much-hyped molecular diagnostics are not available to end-users, but rather to extremely trained clinical laboratory personnel and these medium to high-complexity laboratories. Certain immunoanalytical experiments have a strong false-negative rate by using the RT (primer-PCR) and PCR (polymerase chain reaction) processes. Incorrect RT-PCR findings may be affected by poor compilation, handling, transport, purification, and processing [36]. The existence of the RNA derived from the swabs also influences the results. Other conditions, such as filtered RNA degradation, purification resistance, the involvement of nucleic acid cross-linking reagents, or genomic mutations, may trigger false-negative results. Furthermore, it is critical to note that if blood samples are mishandled during collecting, sorting, and pipetting, false-positive findings will occur. Although the likelihood of these unfavourable outcomes, these diagnostic tests are currently the most accurate, responsive, and easily accessible methods for early and large-scale diagnosis of Serious Acute Respiratory Syndrome-Coronavirus-2 [36].

This can only be calculated using incredibly costly laboratory-based assays that measure a highly unpredictable degree of "dissolvable additives" inside the liquid. These immunoassays can fail to detect whether they have poor sensitivity (and a large LoD), however the number of steps involved in handling the samples and analysing the data increases turnaround time and increases the chance of technical failures and cross-contamination. For further factors are used to measure the total utility of molecular diagnostics as combined with genetics.

**Limitations of Antigen Test:**

This simple antigen tests make obtaining findings simpler, but they are not without limitations. While rapid antigen tests that identify SARS-CoV-2 often focus on nasopharyngeal specimen, the rapid antigen test is more often based on a context of a large number of patients [37]. These specimens are now needed for professional sampling and the use of personal protective equipment during sampling and processing. At this time, the self-sampling method has not been statistically proved or confirmed. Rapid antigen analyses, unlike RT-PCR tests, do not provide controls and therefore are less reliable. Since some of the quick antigen samples are processed separately, analysing vast numbers of specimens at the same time is impossible, and multiplex detection of other respiratory pathogens is currently not possible. Another disadvantage to accelerated antigen tests is that specimens are seldom submitted to public health labs for more characterisation, such as sequencing.

**Limitations of Antibody test**:

A laboratory professional must understand the underlying problem with all serological testing for antibodies, particularly COVID-19 antibodies, which contributes to high analytical error rates. This form of test's reliability is subjective, varied, spontaneous, insidious, and often unreliable. It is caused by different groups and subclasses of distinct antibodies that are produced over the course of an activated immune response (in this case COVID-19 infection). The existence of additional eosinophils as well as other types of normal human antibodies complicates allergic reactions. Any of these passes undetected through the body, and if they do, they can interact with or combine with the tested reagent [38]. This suggests that even with a really strong methodology, the right reagents, and the most stringent internal and external measures in operation, there is always a small margin of unreliability. It is deceptive to claim that such a basic and precise tool is as simple and reliable as a pregnancy test. On these samples, inaccuracy attributable to cross-reactivity is not a possibility.

It is an inherent and variable sensation that false-positive and false-negative outcomes will be obtained in different patient cohorts. When details such as the faulty precision of the procedure, the coexistence of COVID-19 antibodies, and the percentage of individuals with COVID are taken into account, the chances of inaccurate findings being found in different classes of patients may be estimated.

**Limitations of PCR Test:**

The major limitation of RT-PCR research is that it cannot be used to detect previous infection with SARS-CoV-2, which is essential for recognizing the spread of the infection since pathogens only exist in the body for a limited period of time.

Different techniques are required to identify, monitor, and research previous infections, especially those that may have originated and spread without causing symptoms. When an individual recovers, the virus is eradicated, and these checks will no longer determine whether or not the person was infected.

**1.7 Analysing the problem**

There are several limitations on detecting the corona virus inside the humane body. The clinical way is much more hard and painful to obtain. This creates a heavy burden on the people of healthcare system.

**1.8 Developing a model**

Because of these weaknesses we want to introduce the concept of data mining.

And then we will propose a model which will testify the prediction capabilities.

|  |
| --- |
| **Chapter 2: Literature Review** |

**2.1 Chapter Introduction**

In this chapter the summary from the reviewed research papers will be written briefly. From the literature review we understood about the domain of research and got idea about the process of building model and also analyzed the dataset. This helped us to provide a better solution for our own proposed model.

**2.2 Literature review**

By the use of a decision tree, HCV polyprotein cleavage sites have been predicted and these predictions have proved to be satisfactory.

The final results are of the process isn't as good as one would be desired, but the decision tree is what helps produce those results. In the future, they will add more components to the models, including the secondary structure, to study their effects on human body.

The Naive Bayes and decision tree classifier algorithm are the classifiers they used to construct their models. Patients who are health care workers are more likely to survive are first and foremost those who have a good prognosis. In this study, the age of the patient was a strong predictor of how long it would take to find a cure. When patients age between the ages of 66 and 87, they have a higher likelihood of suffering from serious complications. The models were tested and compared, so they were judged to all be equal. The estimated model-accuracy percentage ranges between 53.6% and 71.58% depending on the dataset used [39].

With respect to remove measurements used to evaluate the proximity of neighbours in KNN and its subsidiaries, the normal measurement is Euclidean distance. In any case, Mahalanobis distance turns out to be more reasonable if the information is slanted or the highlights are corresponded, as it mulls over information dissemination. Jaafar et al. [40] report that Euclidean distance crumbles KNN precision if the information is uneven. In this manner, they propose Mahanalobis distance for more precise order. Yi et al. [41] propose a characterization framework dependent on KNN appropriate for automated frameworks. Since robots work in certifiable conditions, where highlights are emphatically associated, they also use Mahalanobis distance. To moderate the computational intricacies associated with Mahalanobis distance, because of figuring the reverse covariance lattice of information, they utilize guideline segment examination (PCA) for information decrease. Fan et al. [42] additionally use Mahalanobis distance with KNN with regards to a structure to improve the security of force frameworks, where highlights are normally profoundly connected.

Information mining has been generally utilized for the forecast and conclusions of numerous illnesses. Ferreira et al. [43] utilized information mining to improve the determination of neonatal jaundice in babies. In their trial, the dataset comprised of 70 variables gathered for 227 sound babies. Numerous information mining strategies were applied, including: J48, Truck, naive Bayes classifier and straightforward calculated. The best prescient models were gotten by utilizing Credulous Bayes, multi-facet perceptron, and basic calculated. For coronary illness analyses, Venkatalakshmi and Shivsankar [44] looked at the presentation of decision tree calculation and Naive Bayes. The trial results utilizing a dataset of 294 records with 13 ascribes showed that the presentation of the two calculations is equivalent. FP-development, Affiliation rule mining, and decision trees were utilized for the determination and visualization of bosom malignant growth [45]. The arrangement models were constructed utilizing a dataset of 699 records and 9 credits and the best exactness was accomplished utilizing decision trees acceptance calculations.

Table 1-A: Existing method with advantages and disadvantages

|  |  |  |
| --- | --- | --- |
| Methods | Advantages | Disadvantages |
| Logistic Regression | LR mainly finds the probability that a new object belongs to a certain class [46] | Linear decision is limited for logistic regression. |
| Random Forest | RF provides noticeable improvements in the classification accuracy of a model through building a  group of trees that generate results individually. [47] | RF is not suitable for regression tasks. |
| Decision Tree | Missing values in the dataset also do not affect the process of building a decision tree to a considerable extent. | DT algorithm is inadequate for applying regression and predicting continuous values. |

The existing methods that are provided with have some advantages and disadvantages. We have to consider those points and provide a suitable model with less disadvantages.

Table 1-B: Comparisons between previous research performance measures

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Methods | Accuracy | Precision Rate | Recall Rate | f-score |
| Proposed Method | 80% | .80 | 1.0 | .89 |
| Villavicencios proposed method [47] | 98.81% | .988 | .988 | .988 |
| Ahamads proposed method [48] | 89% | - | - | - |
| Iwendis proposed method [49] | 94% | 1.0 | .75 | .94 |
| Khandays proposed method [50] | 94% | .93 | .94 | .93 |
| Zhangs proposed method [51] | 75% | - | - | - |

From the existing models that the other researchers have built, majority of them have a better accuracy than our model. Due to not having a proper dataset might have caused our model to have a slight low accuracy than others. But we believe that with the proper dataset we might have better accuracy than this accuracy.

|  |
| --- |
| **Chapter 3: Exploratory Data Analysis** |

* 1. **Chapter introduction**

Open-source secondary data was searched for our research. Since the dataset with symptoms of the patients are quite limited on the internet. We had to carefully gather as much data as possible at any aspect. And we studied the data to identify the issues with the model and proposed a new paradigm to address those issues. In this part, we explored how we collected data, where and who we collected data from, which methods we used for data collection, which methods were used to interpret the collected data, and what the results were after review.

* 1. **Dataset Description**

Data is an integral component of every AI model and, in essence, the prime explanation for the current surge in popularity of machine learning. Because of the availability of data, scalable ML algorithms have become feasible as real goods that can add value to an enterprise, rather than becoming a by-product of the primary processes. A dataset is similar to a database table or a spreadsheet in Microsoft Excel. This is a traditional data framework that is widely used in the area of machine learning. Other types of records, such as photographs, videos, and text, are not considered at this point. Here is a figure showing instance and features of the dataset:

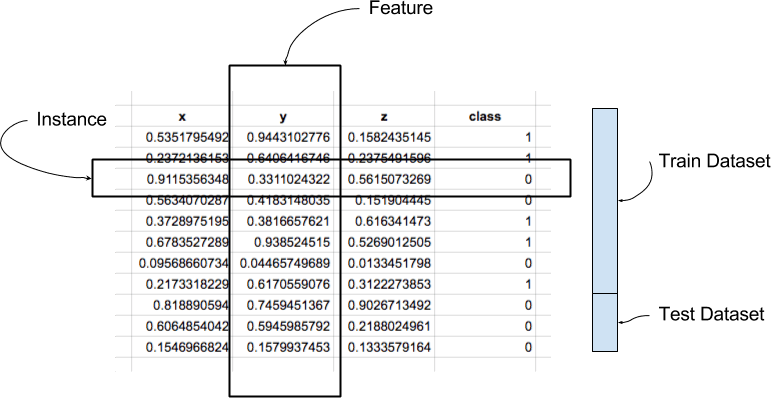


Figure 1-1: Instances and features of dataset [52]

On the figure (0), there are instances, features, train and test dataset are shown that is available on all the datasets.

Instance: An instance is a single row of data. It's a domain-specific phenomenon.

Feature: A feature is a single column of results. It is a part of an observation which is often referred to as a data instance attribute. Some features may be predictors in a model, whereas others may be outcomes or features to be projected.

Data type: A feature has a data type. They may have a true or numerical value, as well as a categorical or ordinal value. Lists, days, periods, and more complicated forms may exist, but when dealing with conventional machine learning approaches, they are usually limited to real or categorical values.

Dataset: A dataset is a list of instances and when dealing with machine learning techniques, it usually requires a few datasets for various purposes.

A secondary data is released on Kaggle [53] is used for creating the model in order to detect and predict virus dissemination throughout the COVID-19 epidemic. The dataset combined with data from various sources to include individual-level data rather than composite data as presented by most data warehouses.

* 1. **Details about dataset**

As we have already used the cleaned covid data with symptoms which was provided in the Kaggle [53]. In the cleaned dataset there are 27 columns and 3,16,801 rows with instances. This cleaned data contains all the possible combinations data from the raw dataset. But this dataset contains dummy variables after combination. The combined data was applied in chatbot, supervised learning and unsupervised learning. As the dataset had nominal values, the train data and test data had to be scaled using standard scalar. So, for this paper, we took 2,37,600 instances with 12 columns.

The dataset is divided into tables, with data of String and Numeric types. The dataset also includes categorical variables. Since the ML model allows all data passed as input to be in numeric form and the dataset all has numeric datatype, it was helpful for the implementation. But only the country column was removed from the dataset. The null and duplicated values were not found in the dataset.

Each case in the database reflects a person who tested positive for COVID-19 and was collected from various sources. There were initially 3,16,801 instances in this dataset. To secure patients' anonymity, each case is deindividualized and anonymized. The cases are labelled with these:

Country: A list of the countries that the individual has lived.

Age: According to WHO, each person's age group is classified.

Symptoms include: Fever, tiredness, difficulty coughing, dry cough, and sore throat are the five main symptoms of COVID-19, according to the WHO. Pains, Nasal Congestion, Runny Nose, Diarrhoea and Other Symptoms were also included on the dataset.

Severity: The degree of seriousness is divided into Mild, Moderate and Serious Severe

Contact: Has the person approached another COVID-19 patient?

Each variable is listed below and this dataset will be referred to as the Kaggle dataset in the paper. Here the list of columns is stated including datatypes:

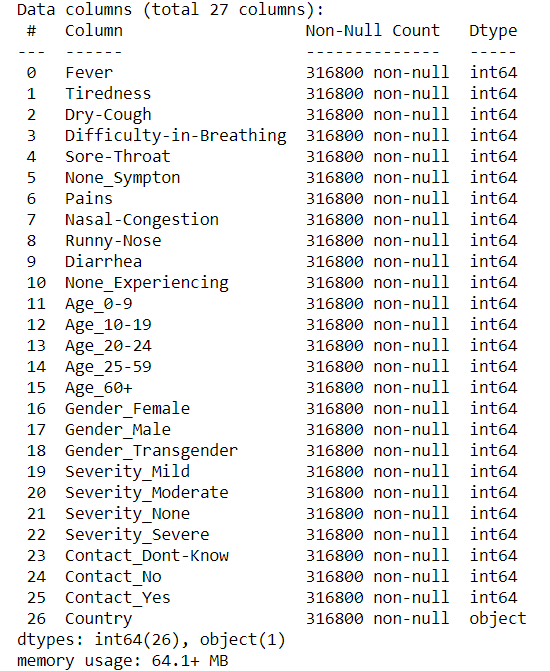


Figure 2-1: The variables of the dataset including datatype [1]

As we can see there are 27 columns and all the datatype have integer values and an object. There is a trade-off with the consistency of this dataset and we hoped to reconcile this situation in the near future. The use of a Kaggle-based dataset was useful due to its large size. Unfortunately, it lacked clear detail about the complications and underlying conditions that each case was dealing with. But this dataset was useful because of its many features, which included various symptoms and ages of the patients infected with covid 19 disease. This dataset was used to train machine learning models in order to reflect the predictive capabilities of datasets of varying quality.

* 1. **Insights and descriptive statistics**

Fever, tiredness, dry cough, sore throat, body pain, nasal congestion, runny nose and diarrhoea were the most frequent symptoms observed in patients whose data were included in this dataset and are depicted in the graph. But there are also data of patients with no symptoms and this was not taken into the consideration. The below figure represents the condition of the patients affected by the forementioned symptoms [1]:

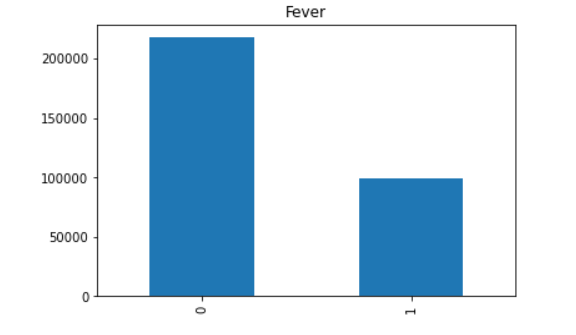


Figure 3-1: Patients with fever symptoms

Here we can see a majority of the people did not had the symptoms of fever.

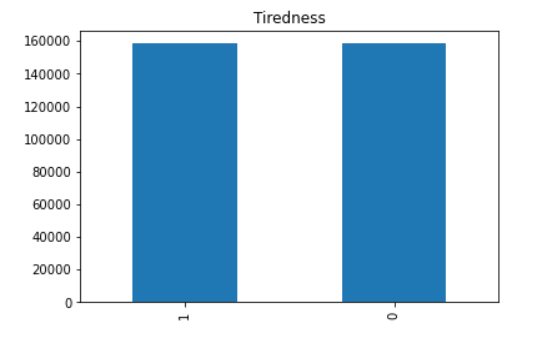


Figure 3-2: Patients with tiredness symptom

Here we can see a half of the people had the symptoms of tiredness and another half did not feel it.

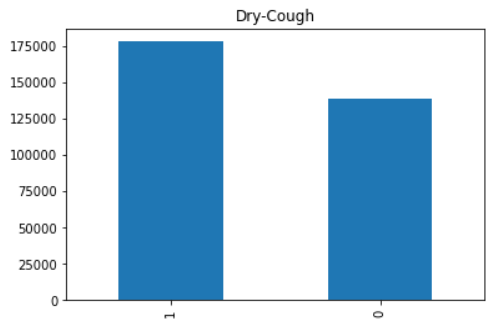


Figure 3-3: Patients with Dry cough symptom

Here we can see a majority of the people had the symptoms of fever.



Figure 3-4: Patients with difficulty in breathing symptom

Here we can see a half of the people had the difficulty on breathing.

.

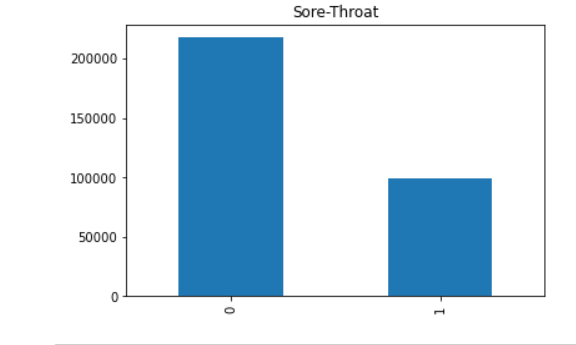


Figure 3-5: Patients with sore throat symptom

Here we can see a majority of the people did not had the symptoms of sore throat.

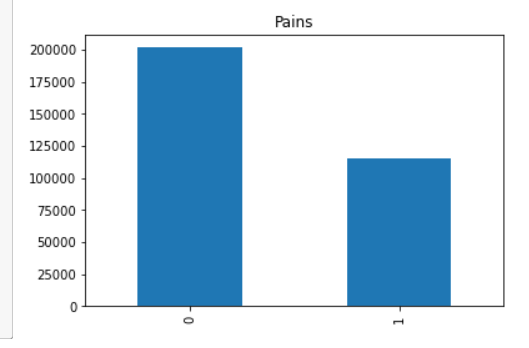


Figure 3-6: Patients with pain symptom

Here we can see a minority of the people had the symptoms of pains in their body.

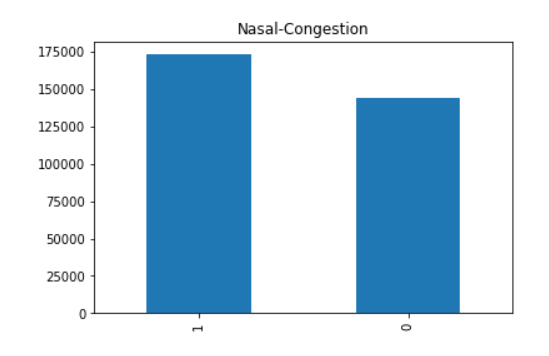


Figure 3-7: Patients with runny nasal congestion symptom

Here we can see a majority of the people had the symptoms of nasal congestion.

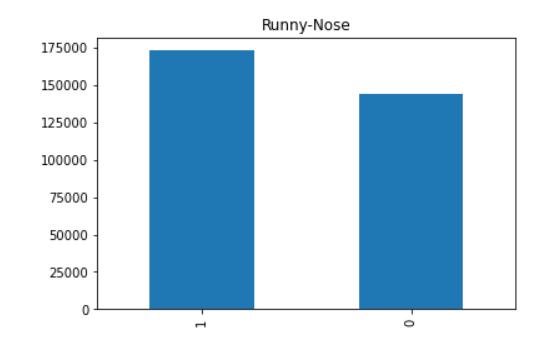
.

Figure 3-8: Patients with runny nose symptom

Here we can see a majority of the people had the symptoms of runny nose.

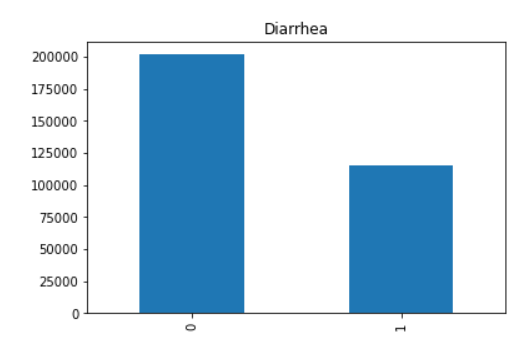


Figure 3-9: Patients with diarrhea symptom

Here we can see a majority of the people did not had the symptoms of diarrhoea.

Correlation between dataset features offers important knowledge about the features and their effect on the goal value. Pearson's temperature chart (figure) depicts the association between the dataset's characteristics, which explicitly shows a comparatively stronger positive correlation with the patients’ symptoms, whether the patient was an age of 0-60 above. This suggests that the patients have higher chance of having affected by the disease. There is also a clear positive association between the symptoms.

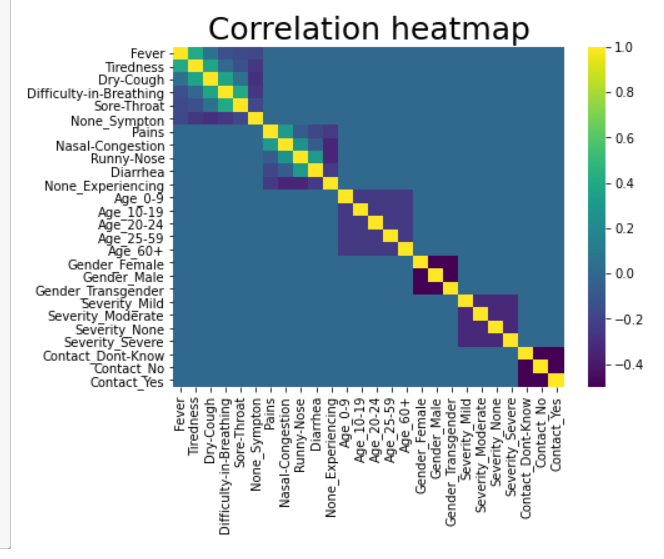


Figure 4-1: Correlation heatmap of the dataset [1]

|  |
| --- |
| **Chapter 4: Modelling** |

* 1. **Chapter Introduction**

In this chapter we will discuss about the model we are going to build by the help of data mining algorithms. But a data model is similar to a specific application written in that language—it utilizes the facilities given by the model in the first sense to address a specific issue.

* 1. **Model selection:**

We have implemented our model by using python language in the Jupyter notebook. Jupyter Notebook is an open-source web application that allows programmers to generate and exchange documents with live code, calculations, visualizations and text. The staff at Project Jupyter are in charge of Jupyter Notebook.

Jupyter Notebooks was a fork of the IPython project, which used to have its own IPython Notebook project [54]. Jupyter gets its name from the three main programming languages it supports: Julia, Python, and R. Jupyter comes with the IPython kernel, which helps you to write Python programs, but there are currently over 100 other kernels available.

Computational notebooks are basically computational computing laboratory notebooks. Rather than pasting DNA gels alongside lab protocols, for example, researchers insert coding, records, and text to record their computational methods. For data scientists, this format may be a source of inspiration for more research. Notebooks are a type of interactive computing, an environment in which users execute code, see what occurs, alter, and repeat in an iterative dialogue between researcher and results. They aren't the only place for such discussions; IPython, the collaborative Python interpreter on which Jupyter's predecessor, IPython Notebook, was developed, is another. Notebooks, on the other hand, encourage users to record such interactions, resulting in more efficient relations between subjects, theories, evidence and results.

The project needs the following packages and libraries: Datetime, Numpy, Pandas, SciPy, Scikit Learn, and Matplotlib. The project was built on the Jupyter notebook platform with the CPU runtime.

We trained several ML classification models using the pre-processed dataset. This research includes the following models: Decision Tree Classifier, Logistic Regression Classifier and Random Forest Classifier. Since the dataset we used can be unbalanced, we can use F1 Score as the primary metric.

The following are the implementation steps:

* Pre-processing of data
* The Random Forest algorithm is being fitted to the Training sample.
* Predicting the outcome of the model
* The result's consistency was checked (Creation of Confusion matrix)
* Visualizing the outcome of the test sample.

These are the steps followed into making the model.

* + 1. **Logistic Regression**:

Logistic regression (LR) is a method for determining the relationship between categorical dependent variables and independent variables [55]. When the dependent variable has two values, such as 0 and 1, yes and no, or true and false, LR is used, and this is known as conditional logistic regression [55]. Multinomial logistic regression is utilized where the dependent variable contains more than two values. To simulate a change of the dependent variables, a statistical model of a series of explanatory variables for LR is used. Mathematically, the LR transformation is written as:

Assume the dependent values are integer values of 1 and 0, with 0 representing a negative value and 1 representing a positive value as a binary variable. As a result, the binary variable's mean would be the proportion of positive values. If p denotes the proportion of observations with a result of 1, then 1 p denotes the likelihood of a result of 0. The odds are described as the ratio p/(1 p), and the LG is the logarithm of the odds, or simply log odds.

* + 1. **Decision Tree:**

Decision tree (DT) is a popular strategy in data mining because of its capacity to manage both categorical and continuous data, simplicity, and comprehensibility. A tree is constructed in the first step by partitioning data into smaller sets until each partition is pure; however, the split form of the data is solely based on the data type. The splits for a numerical attribute C result in the value (C) y, where y is a value in the C domain. Shape the values of (D), B G for breaking a categorical D, where G is a subset of domain (D) [56]. However, the development process of the decision tree is computationally more costly than the pruning phase.

* + 1. **Random forest:**

A random forest is an indicator that is composed of randomized base regression trees. {rn(x, Θm, Dn), m ≥ 1}, where Θ1, Θ2, . . . are i.i.d. outputs of a randomizing variable Θ. The aggregated regression approximation is generated by combining these random trees.

where EΘ is the expectation with respect to the random parameter, conditionally on X, and the data set Dn.  To simplify the notation, we can omit the dependence of the projections in the sample and write, for example, r ¯n(X) instead of r¯n(X) (X, Dn). In reality, the above expectation is calculated using Monte Carlo, that is, by generating M (usually large) random trees and averaging the individual outcomes. A structure of random forest is given below:



Figure 5-1: General process of random forest

The randomizing variable is used to decide how the subsequent three cuts are made when constructing the individual trees, such as the coordinate to break and the location of the split. For this paper we are going to take random forest due to its efficiency in computational time.

* + 1. **Why random forest is selected**

The Random Forest algorithm has the following advantages [57]:

* It avoids overfitting by averaging or integrating the outcomes of multiple decision trees.
* Random forests do better than a single decision tree over a wide variety of data elements.
* The variance of a random forest is lower than that of a single decision tree.
* Random forests are very adaptable and have a high level of precision.
* The random forest algorithm does not require data scaling. It retains good precision even though data is provided without scaling.
* Where a significant majority of the data is unavailable, Random Forest algorithms achieve decent precision.

The Random Forest algorithm has the following drawbacks [57]:

* The biggest drawback of Random Forest algorithms is their complexity.
* Random forests are much more difficult and time-consuming to create than decision trees.
* The Random Forest algorithm necessitates more computing power.
* When we have a huge set of decision trees, it becomes less intuitive.
* In contrast to other algorithms, the prediction process using random forests takes a long time.

In our model we took random forest as a classifier which took less time and had better efficiency than the other classifiers.

**4.3: Proposed method**

The proposed method consists of different steps. Those steps will be followed in order to effectively finding the results of the models. Those results will identify if the person has been infected with covid 19 or not.

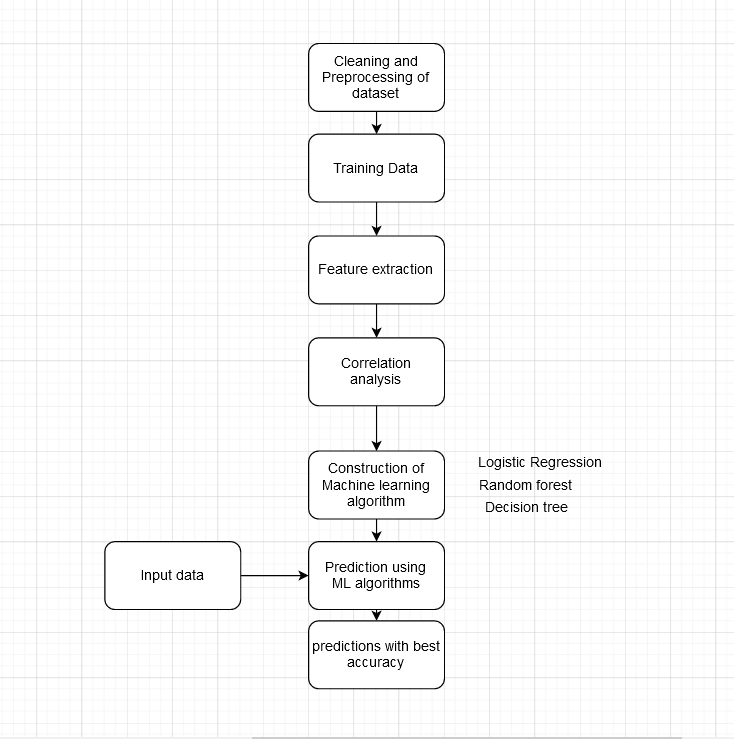


Figure 6-1: Framework of proposed method

Cleaning and pre-processing are mandatory for training the dataset to implement as input in the data mining algorithms. From the dataset features are extracted from the trained data. In this paper we have identified the correlation heatmap and other analysis. Machine learning algorithms such as decision tree, random forest and logistic regression are implemented. From the implementation we will get the accuracy of the proposed model.

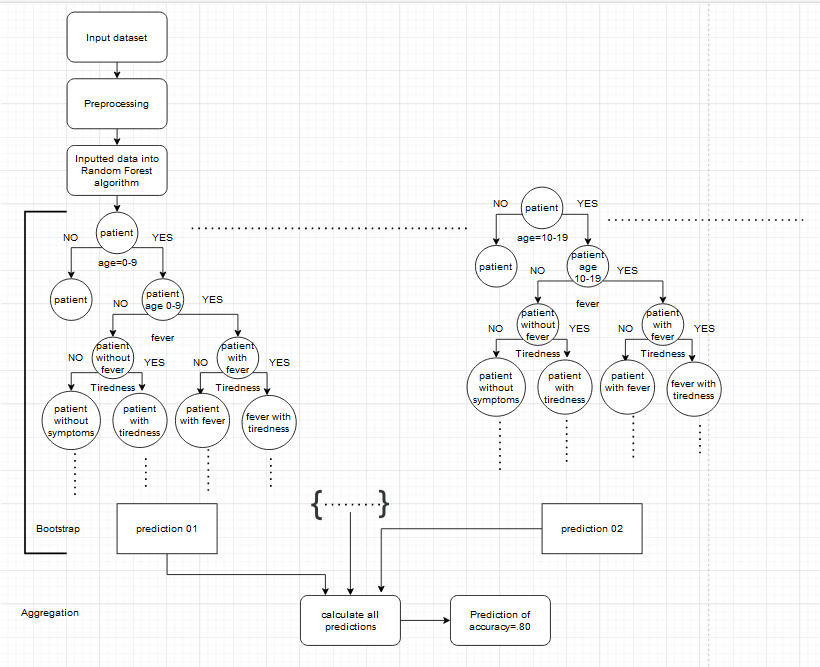


Figure 6-2: Framework of supervised learning

Here is the detailed process of how the model we proposed are come into existence. Random forest algorithm is chosen as supervised learning model in our proposed model. Random forest will split into different decision tree and each decision tree will provide a prediction. After calculating all the predictions, an average or most common predictions will provide the best prediction of the model.

|  |
| --- |
| **Chapter 5: Evaluation** |

**5.1 Evaluation Metrics:**

Here we will evaluate our models that we implemented and compare the performance of the three classifiers. As the cleaned dataset we collected had an issue of having nominal data across the whole dataset, so the accuracy of the all three models had a very similar result. The aim of the following research is to reliably forecast the outcome of a specific patient based on a variety of factors, including ages and so on. Since this is a critical forecast, the accuracy is critical. As a result, for this study's assessment of the model, three evaluation criteria were taken.

In the equations, the following expressions are used: TP, True Positive; TN, True Negative; FP, False Positive; and FN, False Negative.

**Accuracy**:

Given a dataset of (TP + TN) data points, the precision is proportional to the ratio of the classifier's cumulative accurate predictions (TP + TN + FP + FN) to the total data points. Accuracy is a significant metric for evaluating the classification model's efficiency. Accuracy is calculated as shown in the equation below:

**Precision:**

Precision is defined as the ratio of True Positive (TP) samples to the number of TP and FP samples. Precision is another important criterion for determining the number of correctly identified patients in an imbalanced class dataset. Precision is calculated as shown in the equation below:

**Recall:**

The ratio of True Positive (TP) samples to the number of True Positive (TP) and False Negative (FN) samples is equal to recall. Recall is an important criterion for determining the number of correctly identified patients in an imbalanced class dataset out of all patients that may have been correctly expected. Recall is calculated as shown in the equation below:

**F1 score:**

The F1 Score is proportional to the harmonic mean of the Recall and Precision values. The F1 Score achieves the ideal combination between Precision and Recall, offering an accurate assessment of the model's success in classifying COVID-19 patients. F1 score is calculated as shown in the equation below:

**5.2 Performance:**

Here we will evaluate the performance of the three models that are logistic regression, decision tree and random forest. As the dataset had 50/50 percent values which caused a problem for getting the accuracy. All the models had the same accuracy of 66.85%. The figure below shows the results:

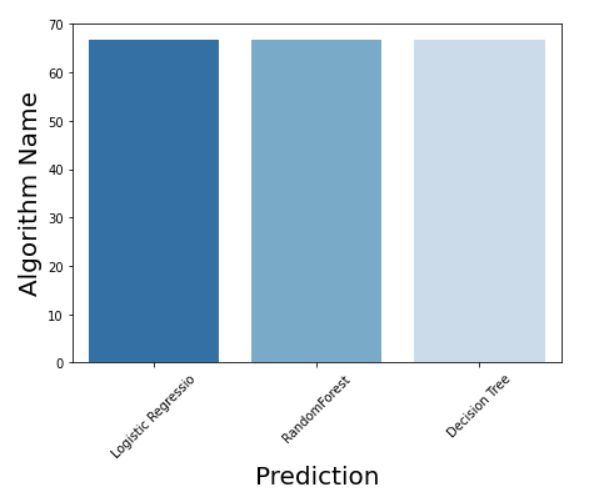


Figure: Predictions of logistic regression, Random Forest and Decision tree[1]

As the figure says, all the algorithm has the same prediction capabilities and this has caused due to the dataset had nominal values.

So, in order to improve our model, we will consider the accuracy of the models in accordance with the age of the patients that are found on the dataset.

Prediction for ages between 0-09:

The accuracy, recall and F1 score are given below:

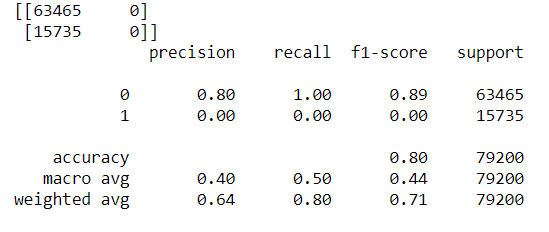


Figure: Accuracy, precision, recall, F1 score of patients aged between 0-9

Random forest algorithms have produced an accuracy of 80.13%,

Prediction for ages between 10-19:

The accuracy, recall and F1 score are given below:

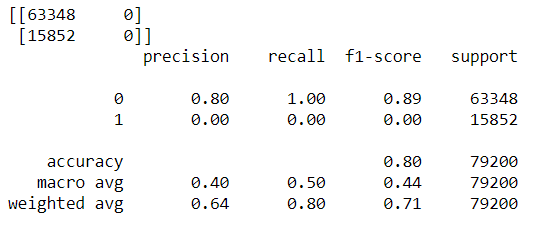


Figure: Accuracy, precision, recall, F1 score of patients aged between 10-19

Even though all the three algorithms have produced an accuracy of 79.98%, we will consider taking only the random forest classifiers value. Because of the efficiency of the random forest and as it took less time than other algorithms, we will consider it as the result of the prediction.

Prediction for ages between 20-24:

The accuracy, recall and F1 score are given below:

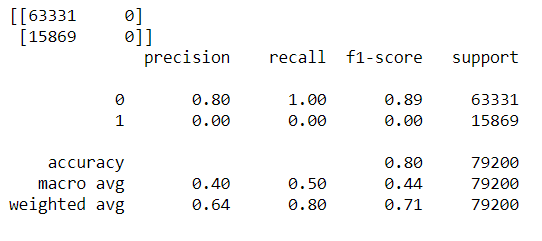


Figure: Accuracy, precision, recall, F1 score of patients aged between 20-24

Even though all the three algorithms have produced an accuracy of 79.96%, we will consider taking only the random forest classifiers value. Because of the efficiency of the random forest and as it took less time than other algorithms, we will consider it as the result of the prediction.

Prediction for ages between 25-59:

The accuracy, recall and F1 score are given below:

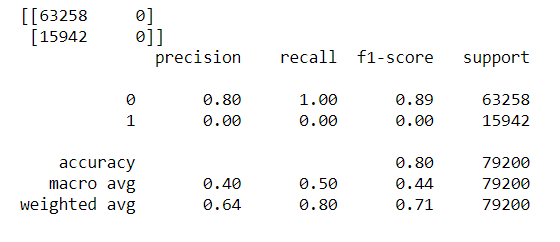


Figure: Accuracy, precision, recall, F1 score of patients aged between 25-59

Even though all the three algorithms have produced an accuracy of 79.87%, we will consider taking only the random forest classifiers value. Because of the efficiency of the random forest and as it took less time than other algorithms, we will consider it as the result of the prediction.

**5.3 Results:**

Here we will discuss about the results of the patients according to the age from 0-59 years. After evaluating all the accuracy of the model, we have found out that the patients who have those symptoms have covid cases at the accuracy of up to 80%. So, it seems that based on ages, the accuracy of the model shows the patients that he/she have been infected by the corona virus.

**5.4 Discussion:**

Based on the model implementation, we have found out that Random Forest algorithm has provided an efficient solution to the problem. As covid is a problematic disease which is quite hard to identify using the clinical test. Sometimes, in hospitals it is quite difficult to get the results of covid if they are positive or negative. To find the disease beforehand, we have used machine learning to get the efficient result and we got a successful accuracy in predicting covid 19.

|  |
| --- |
| **Chapter 6: Conclusion** |

**6.1 Chapter Introduction**

This chapter would bring the whole study project to a close. We would extract our progress still now and what drawbacks we have and how do we scale up this project in future and also will demonstrate a pathway for a robust implementation phase of this project.

**6.2 Success of the thesis**

So far, we have conducted research on base of the open-source dataset we found on Kaggle. The model we created has given an acceptable accuracy rate. The model proposed is focused on age of the patients rather than only the symptoms of the patients. We have successfully created a part of an efficient way to predict the disease.

**6.3 Limitations of the thesis:**

The main limitation of the of the thesis is getting a proper dataset which is available on the open-source platform. As we stated before, this model will efficiently work on the proper dataset. The dataset we found on Kaggle has nominal data which caused the model to provide inadequate information of the process.

**6.4 Future improvement:**

The model we proposed can be improved if the proper dataset can be found on the open-source platform. As COVId-19 is a new disease which recently cause a pandemic. Still the dataset of the COVID-19 is not that available. If a proper medical dataset can be found in the near future, we could use it to get the more efficient result.

**6.5 Conclusion:**

In this research paper, we have discussed about the problems of predicting COVID-19 disease by using clinical method. For which we have had an idea to implement the non-clinical way of using machine learning algorithm to predict the patient whether they have been infected with corona virus or not. The proposed model may not be the best model but this will provide a feasible solution to the present scenario.

**References:**

[1] https://github.com/nafiszawad17/covid-19-pred

[2] <https://covid19.who.int/>

[3] Keeling MJ, Deirdre Hollingsworth T, Read JM (2020) The efficacy of contact tracing for the containment of the 2019 novel coronavirus (covid-19).

[4]<https://www.biorxiv.org/content/10.1101/2020.10.17.343749v1.full#:~:text=Total%20character%20differences%20between%20the,genus%20Betacoronavirus%2C%20subgenus%20Sarbecovirus>).

[5] *Li YC, Bai WZ, Hashikawa T. The neuroinvasive potential of SARS-CoV2 may play a role in the respiratory failure of COVID-19 patients. J Med Virol, Feb 20, 2020. Doi:10.1002/jmv.25728.*

[6] *Lv L, Li G, Chen J, Liang X, Li Y. Comparative genomic analysis revealed specific mutation pattern between human coronavirus SARS-CoV2 and Bat-SARS-CoV Ratg13. bioRxiv,*[*https://doi.org/10.1101/2020.02.27.969006*](https://doi.org/10.1101/2020.02.27.969006)

[7] *Frydman G. Lessons learned: The history, biology and pathophysiology of the current COVID-19 literature. Department of surgery, MGH & Center for Biomedical Engineering Department of Biological Engineering, MIT. March 11, 2020.*

[8] Muhammad LJ, Usman SS. Power of Articial Intelligence to Diagnose and Prevent Further COVID-19 Outbreak: A Short Communication.2020. ArXiv: 2004.12463 [cs.CY] accessed 15th May, 2020

[9] Wölfel R et al. Virological assessment of hospitalized patients with COVID-2019. Nature. 2020.

[10] Stoecklin SB, Rolland P, Silue Y, Mailles A, Campese C, Simondon A, and et al. First cases of coronavirus disease 2019 (COVID-19) in France: surveillance, investigations and control measures, January 2020. Eurosurveillance, 25(6):2000094, 2020.

[11] Coronavirus. World Health Organization: https://www.who.int/healthtopics/coronavirus, 2020.

[12] Rothe C. et al. Transmission of 2019-nCoV infection from an asymptomatic contact in Germany. N. Engl. J. Med. 2020; 382:970–971.

[13] *Radcliffe S. Nearly 40% of those hospitalized for COVID-19 are under 55.*[*https://www.healthline.com/health-news/covid-19-isnt-just-dangerous-for-older-adults*](https://www.healthline.com/health-news/covid-19-isnt-just-dangerous-for-older-adults)

[14] *Guan WJ, Ni ZY, Hu Y, et.al. Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med. 2020. Doi: 10.1056/NEJMoa2002032.*

[15] *Coronavirus kills more men than women. Washingtonpost.com, March 24, 2020.*

[16] *Advice for public/Myth busters. Coronavirus disease (COVID-19) advice for the public: Myth busters. World Health Organization.*[*who.int/emengencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters*](http://www.who.int/emengencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters)*.*

[17] Conditions and Diseases: https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus

[18] *Pan L, Mu M, Ren HG, et al. Clinical characteristics of COVID-19 patients with digestive symptoms in Hubei, China: a descriptive cross-sectional multicenter study. On behalf of Wuhan Medical Treatment Expert Group for COVID-19.*

[19] *Wang GQ. Diagnosis and treatment protococol COVID-19. Released by National Health Commission & State Administration of Traditional Chinese Medicine, March 3, 2020.*

[20] Logistic Regression, NCSS Statistical Software, Chapter 231

[21] Ayon SI, Islam MM, Hossain MR (2020) Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques. IETE J Res 1–20

[22] Islam MM, Iqbal H, Haque MR, Hasan MK (2017) Prediction of breast cancer using support vector machine and K-Nearest neighbors. In: 2017 IEEE Region 10 Humanitarian Technology Conference (R10- HTC). IEEE, pp 226–229.

[23] Mavroforakis ME, Theodoridis S. A geometric approach to Support Vector Machine (SVM) classication. IEEE Transactions on Neural Networks. 2006;17-3,671-682.[1400-1500]

[24] Yahaya BZ et al. An Improved C4.5 Algorithm Using L’ Hospital Rule for Large Dataset, Indian Journal of Science and Technology. 2018; 11:47.

[25] Rong C, Lizhen X. Improved C4.5 Algorithm for the Analysis of Sales. Proceeding of the 6th Web Information System and Application Conference, IEEE. 2009; 173-179

[26] Kohavi R, Quinlan R. Decision Tree Discovery, 1999.

[27] Mehta M, Agrawal R, Risanen J. SLIQ: A Fast Scalable Classier for Data Mining. In: Apers, P, Bouzeghoub M, Ardarin G. Proceedings of the 5th international conference on Extending Database Technology. Berlin: Pringer-Verlag, 1996; 18-32.

[28] Hussain S, et al. Performance Evaluation of Various Data Mining Algorithms on Road Trac Accident Dataset. In: Satapathy S, Joshi A. (eds) Information and Communication Technology for Intelligent Systems. Smart Innovation, Systems and Technologies. 2019; 106.

[29] Muhammad L.J et al. Performance Evaluation of Classication Data Mining Algorithms on Coronary Artery Disease Dataset: IEEE 9th International Conference on Computer and Knowledge Engineering (ICCKE 2019), Ferdowsi University of Mashhad. . 2019.

[30] Gandhi R. Naive Bayes Classier, Towards Data Science. 2018. https://towardsdatascience.com/naive-bayes-classier-81d512f50a7c accessed 25th April, 2020

[31] Haque MR, Islam MM, Iqbal H, Reza MS, Hasan MK (2018) Performance Evaluation of Random Forests and Articial Neural Networks for the Classication of Liver Disorder. In: 2018 International

[32] Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression (PDF). The American Statistician. 1992; 46:3,175–185.

[33] Onel H. Machine Learning Basics with the K-Nearest Neighbors Algorithm, Towards Data Science, 2018 https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm6a6e71d01761 accessed 26th April, 2020

[34] Everitt BS et al. Miscellaneous Clustering Methods in Cluster Analysis. 5th Edition, John Wiley & Sons, Ltd., Chichester, UK. 2011.

*[35] Coronavirus Disease 2019 Testing Basics:* [*https://www.fda.gov/consumers/consumer-updates/coronavirus-disease-2019-testing basics#:~:text=Currently%20there%20are%20two%20types,at%20COVID%2D19%20Diagnostic%20Testing*](https://www.fda.gov/consumers/consumer-updates/coronavirus-disease-2019-testing%20basics#:~:text=Currently%20there%20are%20two%20types,at%20COVID%2D19%20Diagnostic%20Testing)*.*

*[36] Adeel Afzal.* [Molecular diagnostic technologies for COVID-19: Limitations and challenges](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7406419/),*2020 Nov; 26: 149–159. Published online 2020 Aug 6. doi: 10.1016/j.jare.2020.08.002*,*PMCID: PMC7406419*

*[37] Options for the use of rapid antigen tests for COVID-19 in the EU/EEA and the UK ,19 November 2020*

*[38]* *Ismail AA. Serological tests for COVID-19 antibodies: Limitations must be recognized. Ann Clin Biochem. 2020 Jul;57(4):274-276. doi: 10.1177/0004563220927053. Epub 2020 May 14. PMID: 32343598.*

[39] Al-Turaiki, I., Alshahrani, M., & Almutairi, T. (2016). *Building predictive models for MERS-CoV infections using data mining techniques. Journal of Infection and Public Health, 9(6), 744–748.* doi:10.1016/j.jiph.2016.09.007

[40] Jaafar, H.; Ramli, N.H.; Abdul Nasir, A.S.: An improvement to the k-nearest neighbor classifier for ECG database. In: IOP Conference on Series: Materials Science and Engineering, Penang, Malaysia. pp. 1–10 (2018)

[41] Yi, C, et al.: A novel method to improve transfer learning based on Mahalanobis distance. In: 2017 IEEE International Conference on Robotics and Biomimetics (ROBIO), pp. 2279–2283. IEEE (2018)

[42] Fan, H., et al.: Post-fault transient stability assessment based on k-nearest neighbor algorithm with Mahalanobis distance. In: 2018 International Conference on Power System Technology (POWERCON), pp. 4417–4423. IEEE (2018)

[43] D. Ferreira, A. Oliveira, A. Freitas **Applying data mining techniques to improve diagnosis in neonatal jaundice** BMC Medical Informatics and Decision Making, 12 (2012)

[44] B. Venkatalakshmi, M. Shivsankar **Heart disease diagnosis using predictive data mining, International Journal of Innovative Research in Science** Engineering and Technology, 3 (2014), pp. 1873-1877

[45] ] J. Majali, R. Niranjan, V. Phatak, O. Tadakhe **Data mining techniques for diagnosis and prognosis of cancer** IJARCCE, 4 (3) (2015), pp. 613-615

[46] R. Prasad, P. Anjali, S. Adil, and N. Deepa,“Heart Disease Prediction using Logistic Regression

Algorithm using Machine Learning,” Mach. Learn., vol.8, no. 3, p. 4.

[47] Villavicencio, C.N.; Macrohon, J.J.E.; Inbaraj, X.A.; Jeng,J.-H.; Hsieh, J.-G. COVID-19Prediction Applying Supervised Machine Learning Algorithms with Comparative Analysis Using WEKA. Algorithms 2021, 14, 201. <https://doi.org/10.3390/a14070201>

[48] Ahamad, Martuza; Aktar, Sakifa; Rashed-Al-Mahfuz, ; Uddin, Shahadat; LiÃ³, Pietro; Xu, Haoming; Summers, Matthew A.; Quinn, Julian M.W.; Moni, Mohammad Ali (2020). *A Machine Learning Model to Identify Early Stage Symptoms of SARS-Cov-2 Infected Patients. Expert Systems with Applications, (), 113661–.*doi:10.1016/j.eswa.2020.113661

[49] Iwendi, Celestine & Bashir, Ali & Pasupuleti, Naga & Radha, Suja & Chatterjee, Jyotir & Peshkar, Atharva & Mishra, Rishita & Pillai, Sofia & Jo, Ohyun. (2020). COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. Frontiers in Public Health. 8. 10.3389/fpubh.2020.00357.

[50] Khanday, A.M.U.D., Rabani, S.T., Khan, Q.R. et al. Machine learning based approaches for detecting COVID-19 using clinical text data. Int. j. inf. tecnol. 12, 731–739 (2020). <https://doi.org/10.1007/s41870-020-00495-9>

[51] Zhang, J., Jun, T., Frank, J. et al. Prediction of individual COVID-19 diagnosis using baseline demographics and lab data. Sci Rep 11, 13913 (2021). <https://doi.org/10.1038/s41598-021-93126-7>

[52] <https://machinelearningmastery.com/data-learning-and-modeling/>

[53] <https://www.kaggle.com/iamhungundji/covid19-symptoms-checker?select=Cleaned-Data.csv>

[54] <https://realpython.com/jupyter-notebook-introduction/>

[55] Jerry LR, Kaysville U. Logistic Regression. NCSS; 2019

[56] Kohavi R, Quinlan R. Decision Tree Discovery. Handbook of Data Mining and Knowledge Discovery.  University Press; 1999.

[57] https://www.tutorialspoint.com/machine\_learning\_with\_python/machine\_learning\_with\_python\_classification\_algorithms\_random\_forest.htm#:~:text=Advertisements-,Introduction,classification%20as%20well%20as%20regression.&text=Similarly%2C%20random%20forest%20algorithm%20creates,solution%20by%20means%20of%20voting.