American University of Sharjah

College of Engineering

CMP 466: Machine Learning and Data Mining

Spring 2022

Submitted to

Dr. Salam Ahmad Dhou

Department of Computer Science and Engineering

Predictability of COVID-19 infection based on elementary indicators

---

Date of Submission: 4th May 2022

Name & ID:

Mahira Pathan	**g00085003**
Mariyam Ahmed	**g00085178**
Prem Rajendran	**b00084833**
Smit Vaidya	**b00081577**

1. Abstract:

The COVID-19 pandemic has been the biggest health crisis in recent global history. In addition to causing millions of deaths worldwide, a tremendous strain has been placed on healthcare systems as authorities scramble to limit mortality rates. The distribution of resources during times of surging cases, especially in places with scarce access to them, is an important consideration during this pandemic. This creates a need for better predictive methods to properly strategize resource allocation. In this study, patient symptoms and general demographic data was used in conjunction with machine learning algorithms to create models that predict positive cases. Methods such as decision trees, k-nearest neighbors, naive-bayes, support vector machine were used and compared in this project. Evaluation metrics such as accuracy, f1-score, precision, recall, and area-under-curve indicated that Naive-Bayes and decision tree classifiers performed the best, producing accuracies of 0.9134 and 0.9193 respectively. The results of this project can be improved further if some of the attributes in the dataset were recorded as continuous instead of categorical values and if more advanced methods on machine learning are used.

2. Introduction:

So far, there have been 5.78 million deaths due to COVID-19 worldwide [1]. The global pandemic has resulted in tremendous strain on healthcare systems in every country. With increasing cases comes a greater need for beds, medical equipment, and medication. In places with limited resources, strategic allocation is essential to save lives. In the last few years, the COVID-19 pandemic has spurred research in fields as diverse as medicine and remote working. Among the countless innovations the pandemic has inspired, mining patient data is one avenue that can be further pursued. Such data can be used by health authorities to curtail spread and prioritize access to resources. This project aims to predict the probability of testing positive for COVID-19 given certain parameters and symptoms, which are discussed below.

Fortunately in most cases of COVID-19, a patient shows some common symptoms that can be observed and, as such, diagnosis of the disease is possible even in the early stages of the infection. The most common symptoms are cough, fever, tiredness or lethargy, and a loss of smell and consequently taste. Given the nature of the disease, these symptoms are to be expected as it is a respiratory disease and as such it affects the entire respiratory tract including the lungs, nose, mouth and trachea. If allowed to develop for a significant period of time, it can also cause sore throats, general aches and pains, shortness of breath and irritation across the body including the eyes.

For the purposes of a diagnosis, only the more visible and obvious symptoms of this disease are likely to be examined. As such, most diagnoses look at the behavior of the respiratory tract rather than internal pains that may be caused by the disease as well. Thus, in order to deal with the increase in the spread of the disease and reduce the strain on healthcare workers, scientists have used the aforementioned patient data to create various machine learning models that are fed the observed symptoms and result of a test and asked to predict what a test will say about a potential patient.

This work used several machine-learning algorithms such as decision trees, K-nearest neighbors (KNN), Naive-Bayes, and support vector machines (SVM) to determine the most accurate methodology to predict emerging infections. A similar study using data from the Israeli Ministry of Health found that using basic information about patients like their age, gender, and other symptoms associated with COVID-19 can be used to prioritize testing for COVID-19, when resources are limited [2]. Considering the current circumstances, the results from this project could aid efforts to curtail the spread of COVID-19 and allow for better allocation of medical testing resources and manpower. Even as this pandemic diminishes, the machinery developed in this project can be utilized for similar health epidemics to ease demands on local healthcare systems.

3. Literature Review

Given the impact COVID-19 has had on the world in general, it has become a hotly debated topic in the scientific community. In order to reduce the pressures placed on healthcare workers during this time, the scientific community aimed to introduce methods to alleviate their workload utilizing methods like machine learning. Due to this many studies surrounding this topic have been conducted over the last few years, so we have a few papers to analyze that tackle our specific topic through many different methods.

A paper published by Alazab et al. in May 2020 tackles the prediction and detection of COVID-19 utilizing deep learning techniques[3]. As previously mentioned the gravitas of the topic at hand is the primary motivation presented for this paper as they wish to alleviate the mounting pressure on healthcare workers. In order to predict COVID-19 in patients they looked at chest X-Ray scans to view changes caused by COVID as they are relatively cheap to perform and fairly consistent in terms of any inferences that can be made. Specifically concerning the diagnosis element of the paper, they obtained from a test consisting of 1000 X-Ray sample images an F-Measure rating between 95-99% which indicates that their findings are accurate and their model is usable for practical applications of detecting COVID 19 cases.

A paper published by Wynants et al. and finalized in January 2021 tackles the increasing number of COVID related studies and their efficacy specifically relating to the predictive models used for diagnosing the disease[5]. While this paper may not be entirely related at first glance, it does allow us to gauge the reliability of the papers that we have looked at as well as some of the general commonalities between these papers. They specifically looked at papers that established a multivariate relationship between the symptoms and the disease. Through this process, they looked at 169 studies that established 232 prediction models, 118 of which concern our project of diagnosis while the others look at prognostic models concerning the events after contracting the disease. 75 of these based their model on medical imaging. They state that flu-like symptoms are most frequent attributes used in these models which corroborates the variables within our dataset which also implies commonality between the flu and COVID. The C index for diagnostic models in particular have a massive range of 0.65 to more than 0.99 which implies a varying range of efficacy of the models available, however the authors do state that many studies did not ensure that their testing audience was unbiased.

A paper published by Zoabi, Deri-Rozov and Shomron in the nature magazine is a more straightforward look at the topic and most closely resembles our exploration of the diagnosis of COVID through machine learning [6]. They trained a model on 51,831 tested individuals and 4,769 of these were confirmed to have the disease. They took next week's tests, consisting of 47,401 people with 3,624 positives, and used 8 binary features to train and test their model. In order to build their model they used a gradient-boosting machine model built with decision-tree base-learners and area under the receiving operator characteristics as the performance measure. Missing values within the dataset were filled in using the gradient boosting predictor. To conclude, they state that the framework they provide through this model is usable but would need further development requiring more robust data.

Another paper published by Elaziz et al. in June 2020 also looks at X-Ray scans of various patients and aims to classify them as a COVID patient or a non-COVID patient [7]. To do this, they need to extract features from the image, which is done by using new Fractional Multichannel Exponent Moments. Next, the most important features are selected by using a modified Manta-Ray Foraging Optimization based on differential evolution which then forms their model for testing. Since they have two more datasets available to them, they decide to use both for testing their model which provides an accuracy rate of 96.09% and 98.09% for the two sets. They conclude that this model is worthwhile and achieves not only high accuracy but also low consumption of computational resources.

A study aimed to predict the incidence of COVID-19 infection in a patient based on clinical markers was published by the BMC [8]. By performing a meta-analysis of published literature to identify all clinical variables associated with covid-19, the authors aimed to build a model to predict positive COVID-19 cases. The authors used 151 articles in their analysis, including variables such as symptoms, test results, demographics and scans. Methods such as Spearman correlation, Kruskal-Wallis test and chi-squared test were applied to correlate pairs of variables to each other. Moreover, algorithms such as XGBoost, Ridge, Random Forest and LASSO were used to predict cases. It was found that XGBoost was the best performing model, with age being the biggest predictor of the model's accuracy.

Using a dataset that is publically available and provided by the Mexican Federal Government, this paper published by Quiroz-Juárez et al. tries to identify high-risk patients using medical history, demographic information, and recent medical information [9]. It consists of all confirmed and suspected COVID-19 cases, with over 4 million data points. Their model relied on training multilayer feed-forward neural networks to predict whether a positive patient succumbs to the infection or recovers. The model showed a specificity of 82%, a sensitivity larger than 86%, and an accuracy of almost 84%. To sum, the model shows high promise to identify at-risk individuals.

4. Description of Dataset

The dataset found on dockship.io is a tabular dataset that stores the records of patients who have undergone COVID-19 tests [10]. It has a total of 824,579 records with 9 attributes as follows:

1. Age over 60 - Is Age Over 60 (1 for yes, 0 for no)
2. Gender- Male or Female
3. Cough - No Cough (0), Cough (1)
4. Shortness of breath - No Shortness_of_breath (0), Shortness_of_breath (1)
5. Fever -No Fever (0), Fever (1)
6. Sore throat - No Sore_throat (0), Sore_throat (1)
7. Headache - No Headache (0), Headache (1)
8. Contact with a confirmed individual - Contact_with_confirmed or Other or Abroad
9. **Corona_result** - 1 if they have coronavirus, 0 if they don't [TARGET COLUMN]

Additional attribute (not contributing to the model): Test Date

The objective of the project is to classify the patients as COVID-19 positive (indicated by 1) or COVID-19 negative (indicated by 0). The primary class in the dataset is corona_result which has a total of 84,346 1s and 740,233 0s.

The tests were performed between the 10th of September, 2020 and the 12th of November, 2020. Upon analyzing the data at the basic level, it was found that there are some null values for the features ‘Age_over_60’ and ‘gender’.

In particular there were 2101 null values in the dataset for the gender feature. For the ‘Age_over_60’ feature we noticed that there were three values assigned, 0, 1 and _ which we considered to be null and decided to remove from the dataset ending up with 782,628 values thus showing there 40,940 entries with the value ‘_’.

5. Data Visualization

Since the test date does not affect our topic of research, we decided to drop the column. Thus viewing the count of the features in the current dataset before removing the'_' values in the 'age_60_and_above':

	Cough	Fever	sore_throat	shortness_of_breath	head_ache	age_60_and_above	corona_result
0	802537	802182	817206	822363	809040	679302	740232
1	22042	22397	7373	2216	15539	104337	84347
_	0	0	0	0	0	40940	

Additionally for the features that do not fit in this table, we have gender in which there are 422,583 female entries and 399,895 male entries and for test_indication we have other with 771, 313 entries, contact with confirmed with 52,109 entries and 1,157 entries with Abroad as the value and finally we have corona_result whose split is shown in Figure 5.1.

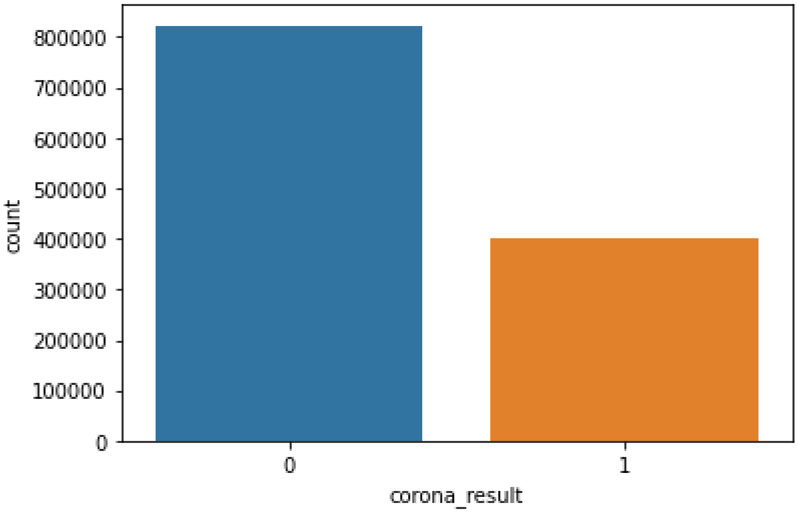

Figure 5.1

All of the features and their spread of values are visualized in the graph below:

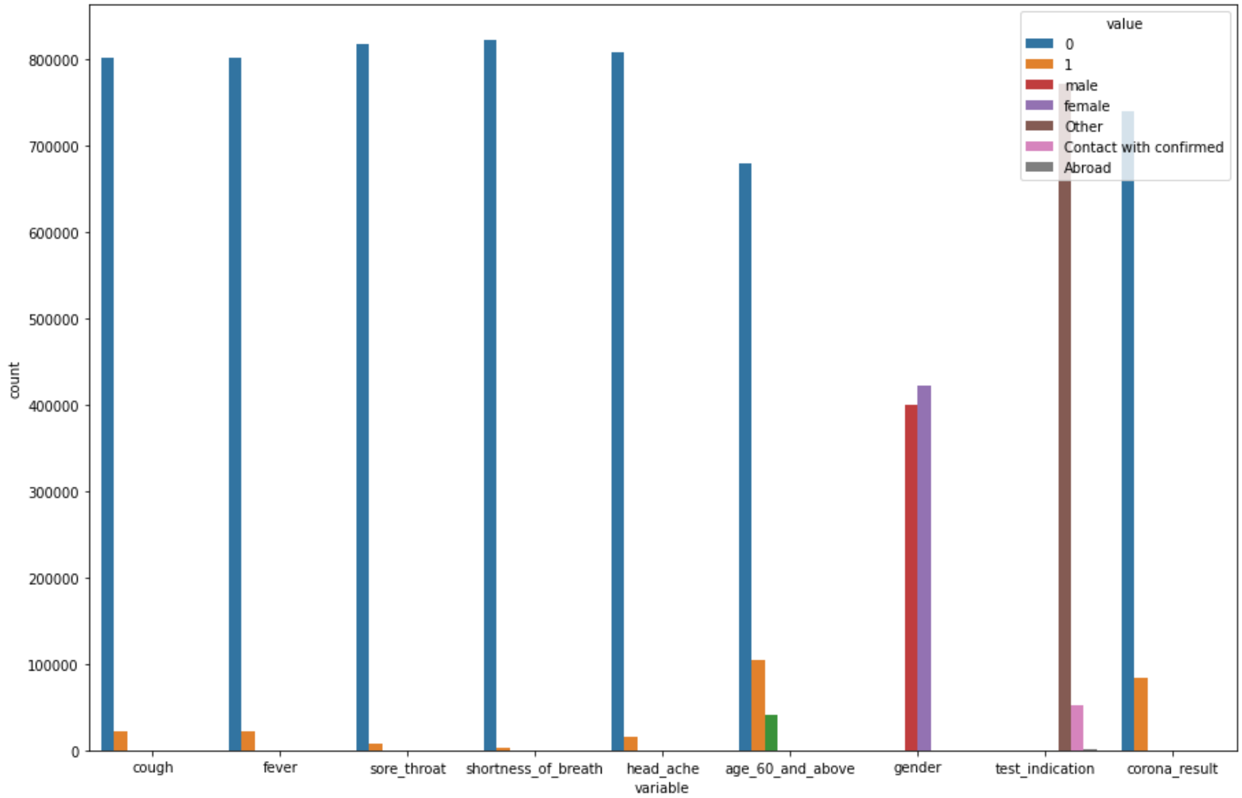

Figure 5.2

We then plot line graphs that indicate the distribution of the records and when they were collected. From our plot, we can see that the most number of tests were conducted between the months of September and November.

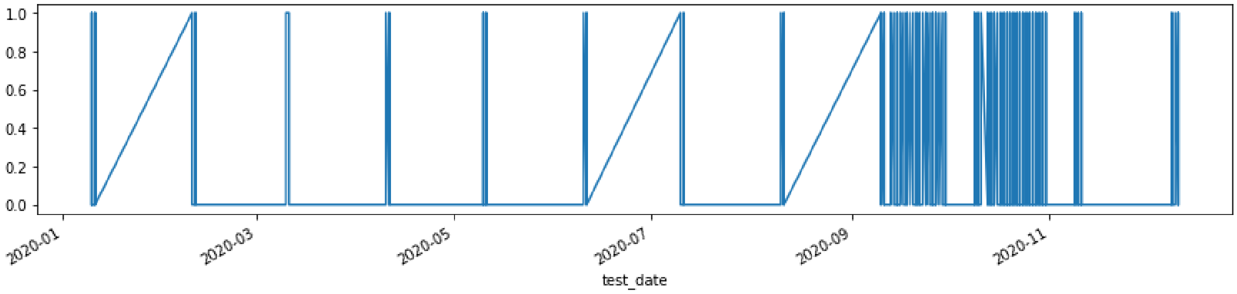

Figure 5.3

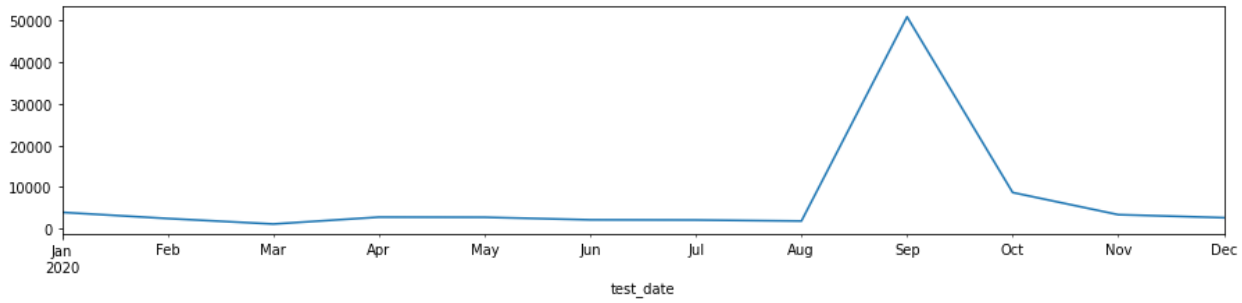

Figure 5.4

6. Data Preprocessing

Since the dataset deals with a medical issue such as COVID-19, it is optimal to discard the data points that contain missing values. The dataset duplicates, if any, should not be dropped as they add information to the dataset. It is a medical dataset so duplicate information is important and cannot be dropped. After dropping the tuples with missing values, we have 822,478 records in our dataset.

Next, we encoded the gender and test_indication attributes using the following metric:

Male = 1, Female = 0
Abroad = 0, , Contact with confirmed = 1, Other = 2

All attributes were converted into integer objects after this, but upon performing this operation we encountered a problem with the attribute ‘age_60_and_above’ as it had rows with the value as ‘_’. We dropped these values, since they are missing values. Upon dropping these tuples, our final dataset had 782,628 records.

7. Methods

The dataset is split into 80% training data and 20% testing data. First, we fit our training data into a decision tree classifier, using the default parameters (using the gini criterion, using a minimum samples split of 1, etc.) and we performed the holdout method 10 times. After obtaining the testing and training datasets, we applied an entropy-based decision tree classifier to compare accuracies between the models.

A gini-based approach was also implemented for the decision tree classifiers that were constrained by the sample size. After obtaining the results for these trees, we recorded the exact value for the minimum sample size at which we believe the model is neither underfitted nor overfitted based on the training and testing accuracies.

Finally, we implemented KNN, gaussian Naive-Bayes, and SVM classifiers. For the KNN classifier, n-neighbors values of 100 and 1000 were compared to determine the best fit using precision, recall, and f-score measures. The Naive Bayes model will allow us to gauge how interrelated the symptoms are allowing us to further refine our methods.

In an attempt to enhance performance, feature selection was applied to the dataset using the SelectKBest and SelectPercentile methods. However, the dataset used in building the various classification models has features stored as categorical variables. Hence, Principle Component Analysis (PCA) cannot be applied to the dataset since the variance cannot be calculated for categorical variables. Additionally, reducing the dimensions of the attributes might actually produce counterintuitive results since the number of attributes is very small. As a result, the dataset was tested feature selection was applied to the dataset.

8. Results and Discussion

8.1 Decision Trees

From our initial assessment of the dataset using the holdout method, we observed the average training accuracy of our decision tree to be 91.91691% and the average testing accuracy to be 91.90575%, which indicates that most of the splits had around the same accuracy. The standard deviation of the training accuracies was around 0.0001 and the testing accuracies were around 0.0005, indicating that the various splits yield similar accuracies. Hence, we chose to randomly split our dataset into training and testing dataset giving us a decision tree as seen in Figure 8.1.

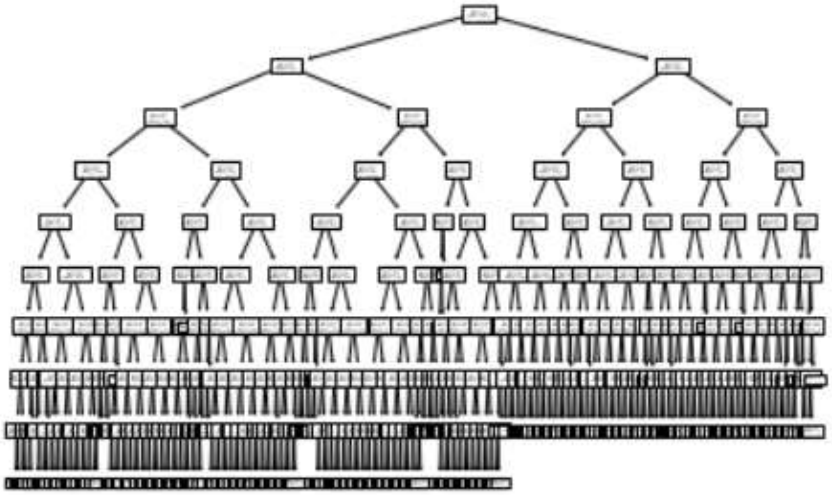

Figure 8.1

The entropy-based decision tree classifier applied on the selected training and testing datasets gave us a training accuracy of 91.89669 and a testing accuracy of 91.98919 resulting in the decision tree as seen in Figure 8.2.

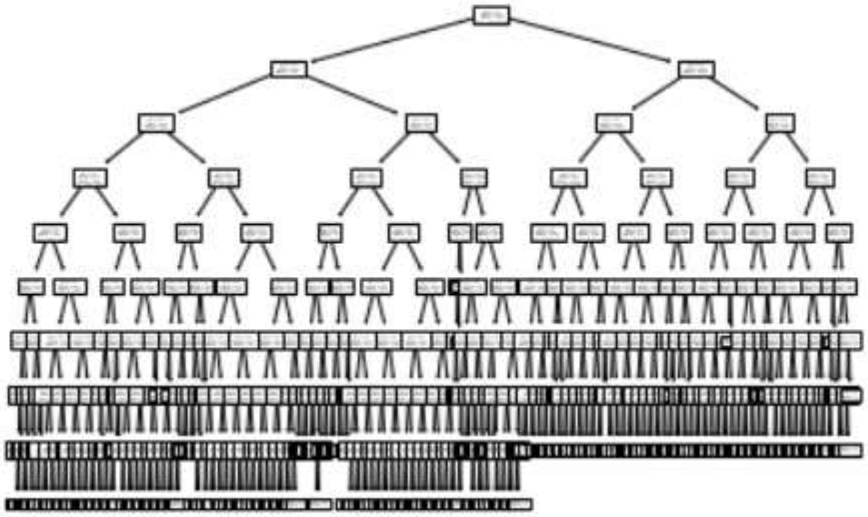

Figure 8.2

Next, a gini-based decision tree classifier was applied to the same dataset resulting in approximately the same training and testing accuracies. Figure 8.3 showcases the gini-based decision-tree obtained. This tree has a training accuracy of 91.89669 and testing accuracy of 91.98791.

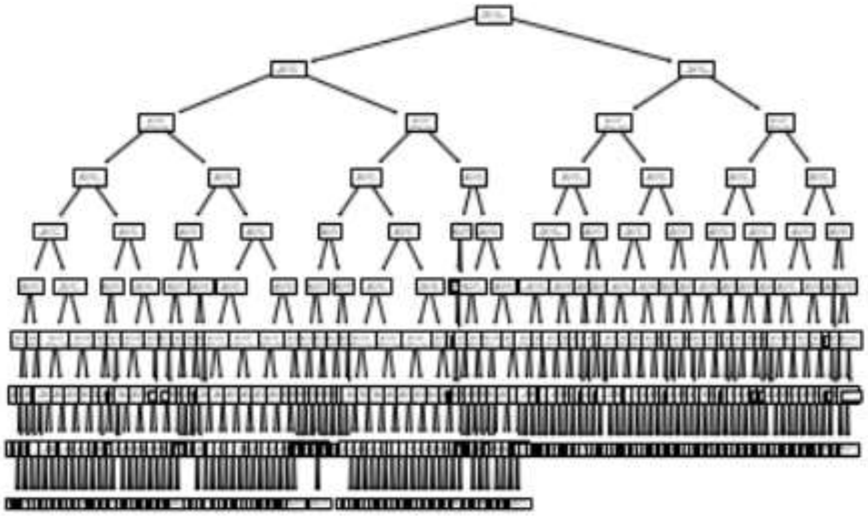

Figure 8.3

Upon splitting the gini-based decision tree for minimum sample split values of 1000, 100, 500, and 600, the model with 600 performed the best based on training and testing accuracy both. The same result was obtained for an entropy-based tree. We tabulated the accuracies for these trees as seen below:

	100	500	600	1000
Training Accuracy	91.92208	91.92192	91.92192	91.91825
Testing Accuracy	91.88186	91.88186	91.88186	91.88186

The tree obtained is as seen in Figure 8.4.

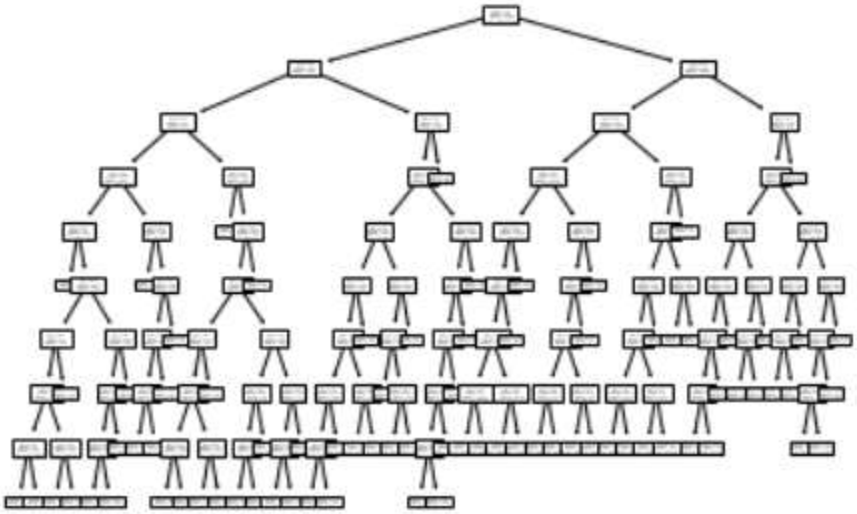

Figure 8.4

We tested the above minimum sample split with the entropy criterion and it yielded the same result as the gini criterion, The tree obtained is as seen in Figure 8.5.

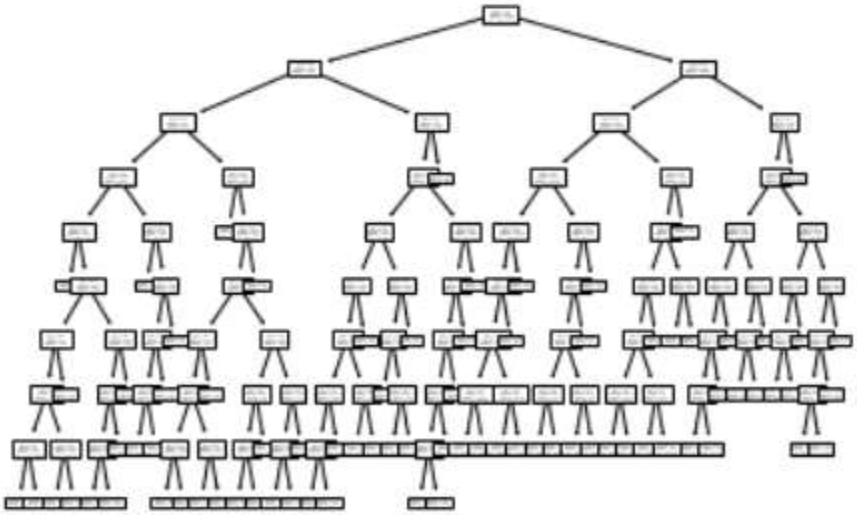

Figure 8.5

As proven above the best decision tree was obtained for a minimum sample split size of 600. Additionally, using the gini and entropy criteria yielded the same results. The overall best accuracy we obtained was 91.91% training accuracy and 91.93% testing accuracy.

For the training accuracy, a plot between the min_samples_split and the corresponding training accuracy yields the graph seen in Figure 8.6.

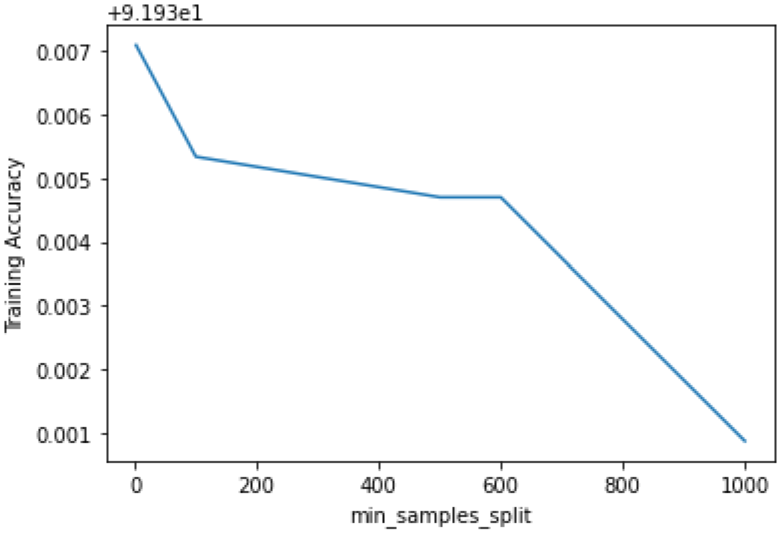

Figure 8.6

For the testing accuracy, a plot between the min_samples_split and the corresponding testing accuracy yields the graph seen in Figure 8.7.

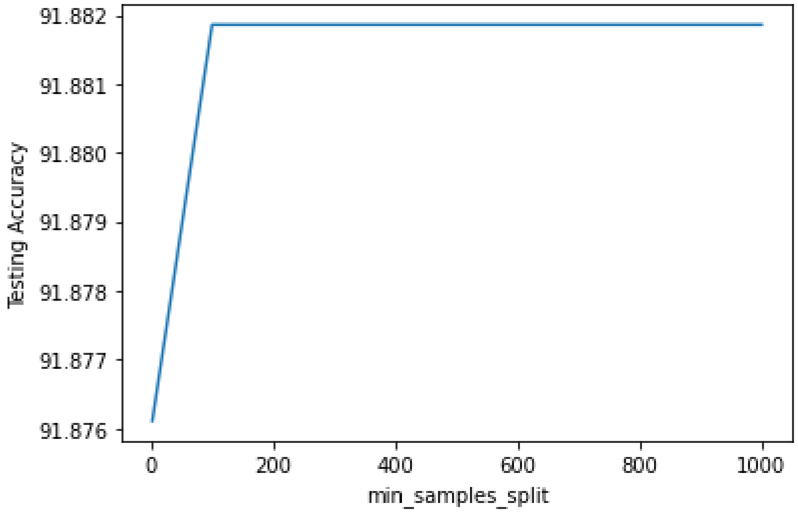

Figure 8.7

Given the Figures 8.6 and 8.7, we can conclude that a minimum sample split size of 600, is optimal. It ensures that the model is not underfitted or overfitted, since a minimum sample split size of 2 leads to an overfitted model, and a minimum sample split size of 1000 and above results in an underfitted one.

8.2 KNN

We implemented the KNN classifier for 100 and 884 nearest neighbors. Based on previous research, the optimal value for nearest-neighbors in the square root of the size of the dataset, which in our case is around 884. Moreover, these values were chosen because a value at the order of something smaller, such as 10 nearest neighbors, would likely produce a highly overfitted model since our dataset contains 100,000 data points.

An accuracy of 0.92 was found for nearest neighbors values of 100.The f1-scores for testing and training were found to be 0.918772 and 0.918435 respectively. To assess these models in more detail, we also tabulated precision, recall, and f-scores for both, as seen below.

NN=100	precision	recall	f1-score	support
COVID positive	0.68	0.42	0.52	16,494
COVID negative	0.93	0.98	0.96	140,032
Macro average	0.81	0.70	0.74	156,526
Weighted	0.91	0.92	0.91	156,526

average				

Further, the confusion matrix was also produced as seen in Figure 8.8

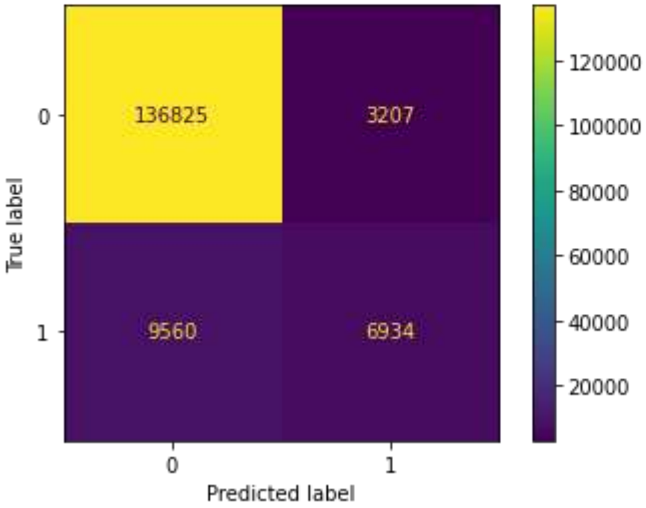

Figure 8.8

Finally, the ROC curve for the model was plotted as seen in Figure 8.9. The area-under-curve was found to be 0.771556, which is a relatively low value.

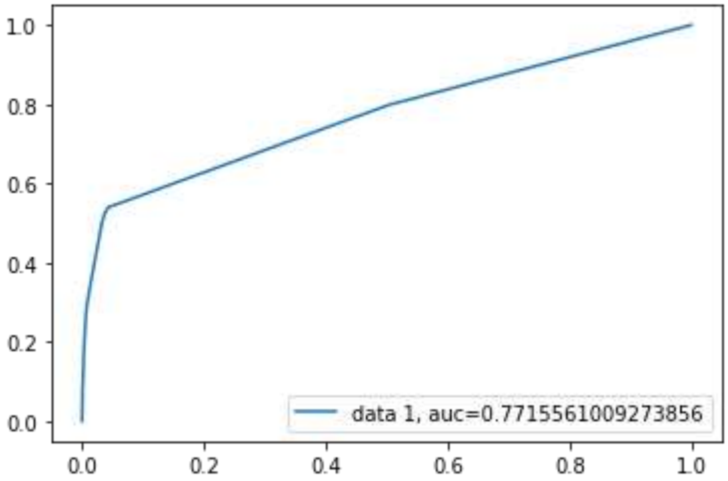

Figure 8.9

As seen above through the low precision and recall rates, this classifier did not perform at a satisfactory level. Results for nearest-neighbors of 884 were similar to nn of 100. This illustrates the limitation of using the KNN classifier on our dataset, which is highly imbalanced. This imbalance leads to a higher likelihood of negative cases, the majority label, being predicted since it is much more abundant than the number of positive cases in the training set.

8.3 Naive Bayes

We also applied a Naive Bayes classification model to our dataset and obtained an f-score of 91.34% and a misclassification rate of 8.68610% which is higher than the values obtained from using other models. This shows that the Naive Bayes model is not fit to use for our dataset. This was to be expected since the model treats all features of the dataset as independent features. However, in the case of the diagnosis of an infectious disease, the likelihood of one symptom being compounded by another is high since they more often than not measure the severity of the infection.

8.4 Support Vector Machines

Finally, we used a Support Vector Machine to classify our dataset. We used a c and gamma value of 1 and obtained a precision of 91.9%. We further tested out other higher values of c and gamma but did not notice any significant increase in the precision and as such chose to stick with the aforementioned c and gamma values. Additionally, we used various kernels such as RBF, linear, and polynomial to test the SVM model which performed as per the results and confusion matrix shown below.

```
0.775 (+/-0.552) for {'C': 1, 'gamma': 0.001, 'kernel': 'rbf'}
0.775 (+/-0.552) for {'C': 1, 'gamma': 0.0001, 'kernel': 'rbf'}
0.775 (+/-0.552) for {'C': 10, 'gamma': 0.001, 'kernel': 'rbf'}
0.775 (+/-0.552) for {'C': 10, 'gamma': 0.0001, 'kernel': 'rbf'}
0.775 (+/-0.552) for {'C': 100, 'gamma': 0.001, 'kernel': 'rbf'}
0.775 (+/-0.552) for {'C': 100, 'gamma': 0.0001, 'kernel': 'rbf'}
0.660 (+/-0.555) for {'C': 1, 'kernel': 'linear'}
0.660 (+/-0.555) for {'C': 10, 'kernel': 'linear'}
0.660 (+/-0.555) for {'C': 50, 'kernel': 'linear'}
0.660 (+/-0.555) for {'C': 100, 'kernel': 'linear'}
0.775 (+/-0.552) for {'C': 1, 'degree': 2, 'kernel': 'poly'}
0.775 (+/-0.552) for {'C': 1, 'degree': 4, 'kernel': 'poly'}
0.775 (+/-0.552) for {'C': 1, 'degree': 6, 'kernel': 'poly'}
0.775 (+/-0.552) for {'C': 1, 'degree': 8, 'kernel': 'poly'}
0.775 (+/-0.552) for {'C': 1, 'degree': 10, 'kernel': 'poly'}
0.775 (+/-0.552) for {'C': 10, 'degree': 2, 'kernel': 'poly'}
0.775 (+/-0.552) for {'C': 10, 'degree': 4, 'kernel': 'poly'}
0.775 (+/-0.552) for {'C': 10, 'degree': 6, 'kernel': 'poly'}
0.775 (+/-0.552) for {'C': 10, 'degree': 8, 'kernel': 'poly'}
0.775 (+/-0.552) for {'C': 10, 'degree': 10, 'kernel': 'poly'}
0.775 (+/-0.552) for {'C': 100, 'degree': 2, 'kernel': 'poly'}
0.775 (+/-0.552) for {'C': 100, 'degree': 4, 'kernel': 'poly'}
0.775 (+/-0.552) for {'C': 100, 'degree': 6, 'kernel': 'poly'}
0.775 (+/-0.552) for {'C': 100, 'degree': 8, 'kernel': 'poly'}
0.775 (+/-0.552) for {'C': 100, 'degree': 10, 'kernel': 'poly'}
```

The confusion matrix for the best SVM model is shown below:

	precision	recall	f1-score	support
0	0.95	1.00	0.98	20
1	0.00	0.00	0.00	1
Accuracy			0.95	21
Macro avg	0.48	0.50	0.49	21
Weighted avg	0.91	0.95	0.93	21

The best SVM model tested on a small sample of the dataset resulted in an excellent f-score of 95%.

8.5 Feature Selection

Feature selection was applied on the dataset using the SelectKBest model having k=5 and applied on an SVC model using RBF as the kernel and gamma = 0.001 resulting in the following confusion matrix. The matrix indicates that applying feature selection did not significantly enhance the performance of the model since the f-score is still in the range of 91%. Feature selection using the SelectPercentile model with percentile = 80 gave similar results. Percentile was selected as 80 since the number of features used for prediction is only seven.

	precision	recall	f1-score	support
0	0.92	0.99	0.95	392265
1	0.74	0.28	0.41	46007
Accuracy			0.91	438272
Macro avg	0.83	0.64	0.68	438272
Weighted avg	0.90	0.91	0.90	438272

```
Confusion matrix:
[[387625   4640]
 [ 33026  12981]]
```

8.6 Literature Review Comparison

As mentioned earlier in this paper, while there are many prospective groups who created a model for COVID-19 detection through symptoms, most of the approaches use image recognition, X-ray analysis and other alternate sources of data by which they define their model. As such, they are not directly comparable to our work. However, due to the increased variance in their possible feature sets, they have generally obtained higher accuracies than our model did. Other papers also tackled aspects that are related to the disease but not directly to our objective.

The closest comparison we have to our approach is the one taken by Yazeed Zoabi et al. in their machine learning-based paper published in the esteemed Nature journal [2]. While the datasets they used may be similar, the approaches we took could not be more different. They first used an ROC curve and a SHapley Additive exPlanations (SHAP) beeswarm plot to take note of the most important features in the dataset. Next, they chose to account for the innate biases present in data collectors and pruned the data to leave only the features that are the most unbiased. They did this by looking at secondary research to take note of the symptoms most ignored. This pruning of the data allowed them to have the data that would make it easier for their gradient boosting algorithm to provide a better model than ours.

However, we did obtain a higher accuracy than one group. Quiroz-Juárez et al. obtained an accuracy of 84%, although their testing set was significantly larger. Although there may have been some groups we bettered, the paper by Yazeed Zoabi et al. proves that our model still has room for improvement. We can fulfill this by using methods that we didn’t use here such as the methods employed by them and further boost the usability of our model.

9. Conclusion

Our model for the given dataset aims to predict whether a patient has COVID-19, based on some elementary indicators. It contains indicators such as cough, fever, gender etc. We pre-processed the data to remove any entries within the set that may not be suitable for creating our model such as removing the missing values, invalid entries, etc. We then visualized the data to view any possible trends or anticipate how the data would affect our model.

We split the data into a training set and testing set, and applied the decision tree classifier (results shown above) among various other models such as Naive Bayes, KNN, etc. Although the SVM models performed well, these results are not perfectly reliable since they were tested on a rather small sample of the dataset. From the above results, we can conclude that a decision tree classifier reached a maximum training accuracy of 91.91%, and testing accuracy of 91.93% in the optimal conditions. In order to further improve accuracy for this particular phenomenon we could, when collecting the data, consider using continuous values rather than binary nominal values for features such as fever or ordinal values for age, which will enable us to test the models based on various threshold values. By increasing the range of possible values and variance in our dataset we could have a model that is less limited and is thus far more accurate when placed in a real-life scenario.

10. References

[1] University of Oxford, “Our world in data: Covid-19 data explorer, ” 2022. [Online]. Available: https://ourworldindata.org/explorers/coronavirus-data-explorer

[2] Zoabi, Y., Deri-Rozov, S. & Shomron, “Machine learning-based prediction of COVID-19 diagnosis based on symptoms” *npj Digital Medicine*. https://doi.org/10.1038/s41746-020-00372-6

[3] Alazab, M., Awajan, A., Mesleh, A., Abraham, A., Jatana, V. and Alhyari, S., 2020. *COVID-19 Prediction and Detection Using Deep Learning*. International Journal of Computer Information Systems and Industrial Management Applications. Available at: <http://www.softcomputing.net/ijcisim_1.pdf> [Accessed 10 April 2022].

[5] L. Wynants, B. V. Calster, G. S. Collins, R. D. Riley, G. Heinze, E. Schuit, M. M. J. Bonten, D. L. Dahly, J. A. Damen, T. P. A. Debray, V. M. T. de Jong, M. D. Vos, P. Dhiman, M. C. Haller, M. O. Harhay, L. Henckaerts, P. Heus, M. Kammer, N. Kreuzberger, A. Lohmann, K. Luijken, J. Ma, G. P. Martin, D. J. McLernon, C. L. A. Navarro, J. B. Reitsma, J. C. Sergeant, C. Shi, N. Skoetz, L. J. M. Smits, K. I. E. Snell, M. Sperrin, R. Spijker, E. W. Steyerberg, T. Takada, I. Tzoulaki, S. M. J. van Kuijk, B. C. T. van Bussel, I. C. C. van der Horst, F. S. van Royen, J. Y. Verbakel, C. Wallisch, J. Wilkinson, R. Wolff, L. Hooft, K. G. M. Moons, and M. van Smeden, “Prediction models for diagnosis and prognosis of covid-19: Systematic Review and Critical Appraisal,” *The BMJ*, 07-Apr-2020. [Online]. Available: https://www.bmj.com/content/369/bmj.m1328.long. [Accessed: 11-Apr-2022].

[6] Y. Zoabi, S. Deri-Rozov, and N. Shomron, “Machine learning-based prediction of COVID-19 diagnosis based on symptoms,” *Nature News*, 04-Jan-2021. [Online]. Available: https://www.nature.com/articles/s41746-020-00372-6. [Accessed: 11-Apr-2022].

[7] M. A. Elaziz, K. M. Hosny, A. Salah, M. M. Darwish, S. Lu, and A. T. Sahlol, “New Machine Learning Method for image-based diagnosis of COVID-19,” *PLOS ONE*, 26-Jun-2020. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0235187. [Accessed: 11-Apr-2022].

[8] W. Tse LI, W. M. Ongkeko, M. Rajasekaran, E. Y. Chang, M. Andrew Yu, S. Z. Kuo, T. K. Honda, A. Gnanasekar, A. Lee, T. Zhang, L. M. Wong, J. Xu, C. O. Honda, L. Apostol, J. C. Tsai, J. Chakladar, G. Castaneda, N. Shende, and J. Ma, “Using machine

learning of clinical data to diagnose COVID -19: a systematic review and meta-analysis," *BMC Medical Informatics and Decision Making*, 29-Sep-2020. [Online]. Available: https://bmcmedinformdecismak.biomedcentral.com/track/pdf/10.1186/s12911-020-01266-z.pdf. [Accessed: 11-Apr-2022].

[9] M. A. Quiroz-Juárez, A. Torres-Gómez, I. Hoyo-Ulloa, R. de J. León-Montiel, and A. B. U'Ren, "Identification of high-risk COVID-19 patients using machine learning," *PLOS ONE*, 20-Sep-2021. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0257234. [Accessed: 11-Apr-2022].

[10] COVID19 Prediction Based On Symptoms, Machine Learning Challenge AUSxUOS, 2022. [Online]. Available: https://dockship.io/challenges/61c59cf04753920d3faa699b/covid19-prediction-based-on-symptoms/overview