CS 4641

12/10/21

Dr. Ritter

Final Project Report

COVID-19 Symptom Predictor Model

By Valentin Stelea, Chihyo Ahn, Yeonhak Kim, Kevin Guernsey

Introduction

For the past two years, our world has changed dramatically. With the emergence of the novel COVID-19 pandemic, we all experienced disruption in our day to day lives. At the forefront of battling this pandemic has been breakthroughs and innovations in the medical community, and a vaccine that would hopefully bring us all back to normal. However, we believe that technology and machine learning can also be at the forefront, and with the powerful capabilities of predictive models such as ID3 decision trees and neural networks, we too, can help bring an end to this pandemic. During this paper, we are going to be analyzing data on common symptoms (five of which are major symptoms of COVID-19), as well as data regarding the countries person X visited, their age, experience with any other symptoms (pains, nasal congestion, runny nose, diarrhea, etc.), severity of symptoms, as well as whether the person contacted some other COVID-19 patient. All of these factors will be taken into consideration, and we will be attempting to create the most accurate COVID-19 predictor we can with the skills we learned throughout this course.

Data

The dataset chosen for our project is adapted from the Isreali government's Ministry of Health's dataset on COVID-19. The original dataset is written in Hebrew, so translation into English was necessary in order for us to interpret the results. Also, upon analysis of the original dataset, it was found that over 90% of the original dataset had a result of negative for COVID-19. Therefore, we had to choose an excerpt from the full dataset based on the time of the year in order to reduce the number of negative entries. A portion of the excerpted dataset is shown below in Figure 1. Each entry in the dataset describes a person in Israel that got tested for

COVID-19 between March 11, 2020 and April 30, 2020, which we excerpted from the original dataset because it was a two month span. Each person either tested negative or positive for COVID-19, and reported any symptoms they had on the day of testing. The symptoms that are included in the Israeli dataset are cough, fever, sore throat, shortness of breath, and headache. Also, the dataset took into account the person's age (whether or not they were older than 60 years old), their gender, and the main reason for testing. The included reasons for testing are: confirmed contact with someone who tested positive, abroad (both going in and out of the country), and all other reasons, which are labeled as 'Other.' All of the attributes in the dataset are binary attributes, except for the test date and the test indication.

Test Date	Cough	Fever	Sore Throat	Shortness of Breath	Headache	Corona Result	Age Over 60?	Gender	Test Indication
4/30/2020	0	0	0	0	0	negative	None	female	Other
4/30/2020	1	0	0	0	0	negative	None	female	Other
4/30/2020	0	1	0	0	0	negative	None	male	Other
4/30/2020	1	0	0	0	0	negative	None	female	Other
4/30/2020	1	0	0	0	0	negative	None	male	Other
4/30/2020	1	0	0	0	0	negative	None	female	Other
4/30/2020	1	1	0	0	0	negative	None	male	Abroad
4/30/2020	0	0	0	0	0	negative	None	female	Other
4/30/2020	0	0	0	0	0	negative	None	male	Other
4/30/2020	0	0	0	0	0	negative	None	male	Contact with confirmed
4/30/2020	1	1	0	0	0	negative	None	female	Other
4/30/2020	0	0	0	0	0	negative	None	female	Other
4/30/2020	0	0	0	0	0	negative	None	female	Other
4/30/2020	0	0	0	0	0	negative	None	female	Other
4/30/2020	0	0	0	0	0	negative	None	male	Other
4/30/2020	1	0	0	0	0	negative	None	male	Other
4/30/2020	0	0	0	0	0	negative	None	female	Abroad
4/30/2020	1	1	0	0	0	negative	None	male	Abroad
4/30/2020	1	0	0	0	0	negative	None	female	Other
4/30/2020	1	0	0	0	0	negative	None	male	Abroad
4/30/2020	1	0	0	0	0	negative	None	male	Other

Figure 1. Excerpt of the Israeli Dataset

Methodology 1: ID3

In this project, our main goal is to predict whether or not a person has COVID-19 based on their symptoms, their age, gender, and test indication. We did this by using two different methods.

The first method is implemented with the Iterative Dichotomizer (ID3) Algorithm, which is a greedy algorithm that is used with decision trees. This algorithm was chosen because the algorithm works best with categorical data. As previously mentioned, almost all of the attributes in the dataset are categorical and distinct, e.g. positive for COVID-19 or negative for COVID-19. Therefore, this algorithm serves as a good model for our dataset.

Before training the data, we had to preprocess the dataset and split it into a training and a test set. To create these subsets, we shuffled the dataset and set 70% of the shuffled dataset as the training set and 30% as the test set, which is also a 1:3 ratio of positive to negative entries. Figure 2 displays the breakdown of both the training set and test set with respect to the positive and negative classes.

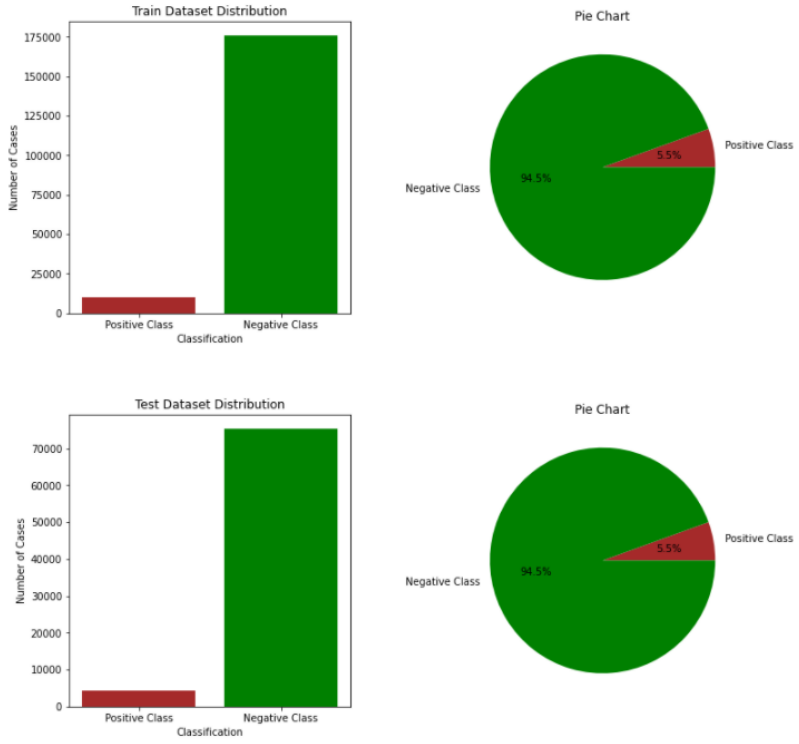

Figure 2. Distributions of Train and Test Datasets

Since 94.5% of both the train and test datasets were in the negative class, we pruned the dataset in order to reduce the amount of entries that are in the negative class. After pruning the dataset, we reshuffled and redistributed the dataset into the train and test sets, for both of which the distributions are shown below in Figure 3.

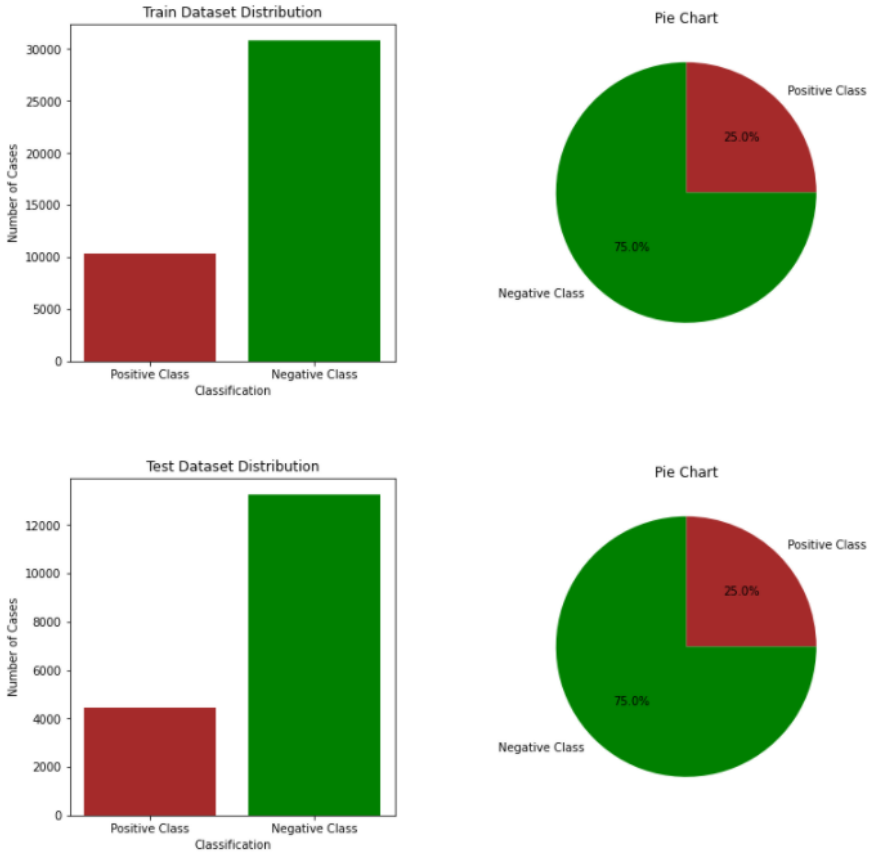

Figure 3. Distributions of Pruned Train and Test Datasets

Upon preprocessing and pruning the train and test datasets, we trained the ID3 model using the train dataset. In our implementation, ID3 maximizes the information gain for all attributes and splits on the attribute that gives the maximum information gain. After training, the model resulted with an 89.09% accuracy. The model then cross-validated with the test dataset, which resulted in an accuracy of 88.65%.

We also re-ran the ID3 model but with a 1:1 ratio dataset, which resulted in training and test accuracies of 50%. One thing that is worth mentioning with the 1:1 dataset is that for both the train and test trials, the accuracy for the positive classes was 0% and 100% for the negative

classes. Therefore, we concluded that for the ID3 model, the 1:3 dataset was more accurate in representing the data.

Methodology 2: Neural Network

For our second method of prediction, we used a neural network. As previously in the first method, the dataset was preprocessed so that each attribute had binary inputs. This time, input features were expanded to 25 binary parameters for faster and uniformly distributed weight training process. Our neural net consists of two fully-connected layers with hidden dimensions of size 100 followed by tanh function. We also set up our training and test datasets so that the ratio of positive class entries to negative class entries is 1:1, compared to the 1:3 ratio from the ID3 methodology. Finally, for our optimizer, we are using Adam, with the hyperparameters set up such that the learning rate is 0.01 and the weight decay is also 0.01. The model then started training based on the training dataset and the weights were updated over 20 epochs. Having cross-validated with the test dataset, the accuracies of the neural net were 83% and 89% for 1:1 and 1:3 datasets, respectively. However, looking at the accuracies for the positive and negative classes separately, we found that the NN was more likely to return a positive result, since the positive accuracy of 92% was greater than the negative accuracy of 76% from 1:1 dataset. Comparing the accuracies of NN and the ID3 models, we can conclude that the neural net performed better than ID3, especially when the data was uniformly distributed. For the 1:3 dataset, both models had a test accuracy of 89%, but with the 1:1 dataset, the neural net was vastly more accurate than the ID3 with an accuracy of 83% compared against an accuracy of 50%.

To further analyze the NN model and the features contributing to the positive and negative results, we calculated feature importance factors using the integrated gradient method. Figure 4 shows each feature's calculated importance towards positive, negative classes and their difference to visualize the relative impact on both results.

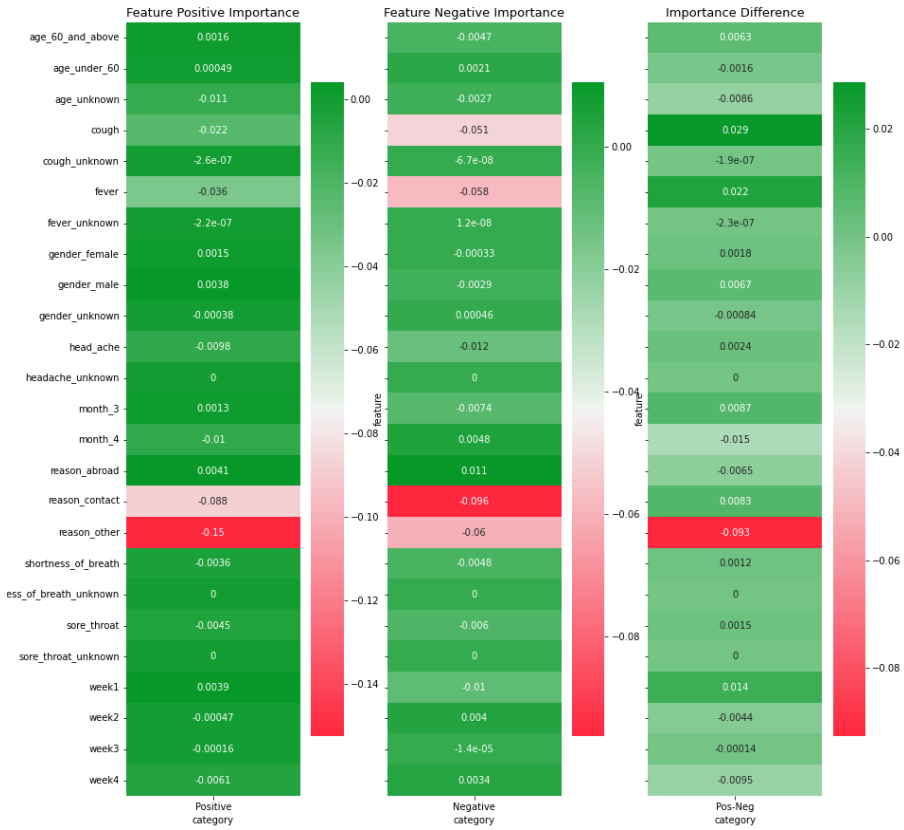

Figure 4. Feature Importances on Positive and Negative Classes

As shown in the third column in Figure 4, the importance difference can be interpreted as the overall contribution to the prediction, where the more positive a feature is, the more impact it has on the final prediction being positive. Therefore, the three most telling or important features for

predicting a positive COVID-19 case are a cough, a fever, and the test being conducted in the first week of the month. On the other hand, the two most negative features are other reasons for getting tested and the test occurring in April. These interpretations match the initial real-world observations of COVID-19 symptoms back in April 2020, when the symptoms that were perceived to be the most common were a cough and a fever. This conclusion also makes sense since the dataset that is used in both models is from April 2020.

Conclusion

Throughout this course, we have constantly been learning about the mathematics, and the reasoning behind many common Machine Learning methodologies. This project was the perfect opportunity for us to use what we learned and apply it to a real-life problem. If we were to redo this project, a possible extension would be to pre-process the dataset more thoroughly, create an even deeper neural network as currently the network only had two fully-connected layers with around 25 parameters in the input. An issue was the dataset was pretty dispersed, and after 2 epochs the training accuracy already maxed out at the highest accuracy, and this showed that increasing training did not create better accuracy. Therefore, creating a deeper network or better pre-processing the data so that the training model can find the tendencies better is what we would do differently to create a more accurate model. Furthermore, implement a more updated dataset that can include some of the more parameters in the data such as vaccination status, vaccination brand, and have a more complicated and in-depth prediction model to account for the most up-to-date COVID-19 and the different Delta and Omicron variants would be a worthwhile addition.

A key observation of the results is that the models tended to yield a positive result rather than a negative in some cases, so you would have some false-positives, but we believed that it is better to have false-positives than false-negatives, so while this is an issue, we did not consider it

a major one. Lastly, while we believe that in the future Machine Learning prediction models will be crucial in helping identify and diagnosing diseases and conditions amongst patients, the result of our models should not be used as a reason for not getting tested for COVID-19, or getting vaccinated, as getting tested on a regular basis and ensuring you are vaccinated is still the most safe and accurate method of protecting yourself, and those around you.

Works Cited

Adamu, Hassan, et al. "Covid-19 Symptoms Detection Using Iterative Dichotomiser ..." *ResearchGate*, Apr. 2021, https://www.researchgate.net/profile/Hassan-Adamu-2/publication/350640627_COVID-19_Symptoms_Detection_Using_Iterative_Dichotomiser_ID3_Algorithm_A_Greedy_Method/links/606b7946299bf1252e2fce89/COVID-19-Symptoms-Detection-Using-Iterative-Dichotomiser-ID3-Algorithm-A-Greedy-Method.pdf.

De Souza , Fernanda Sumika Hojo, et al. "Predicting the Disease Outcome in COVID-19 Positive Patients through Machine Learning: A Retrospective Cohort Study with Brazilian Data." *Frontiers*, Frontiers, 1 Jan. 1AD, https://www.frontiersin.org/articles/10.3389/frai.2021.579931/full#h4.

Hungund, Bilal. "Covid-19 Symptoms Checker." *Kaggle*, Kaggle, 21 Mar. 2020, https://www.kaggle.com/iamhungundji/covid19-symptoms-checker.

Zoabi, Yazeed, et al. "Machine Learning-Based Prediction of COVID-19 Diagnosis Based on Symptoms." *Nature News*, Nature Publishing Group, 4 Jan. 2021, https://www.nature.com/articles/s41746-020-00372-6.