COVID-19 INFECTION PREDICTION FROM SYMPTOMS

Mohsen salimy

Abstract— The main objective of the project is to create a classification model [1] , that will learn the pattern available in the dataset about COVID-19 survey in Brazil [2][3] , and later will give probabilities for new observations given to it. The first step of the study was the cleaning of the data and making some transformations [4] after which, 3 different models were created and the dataset was trained on the best one.A very good outcome of the project was obtained, as the model allowed to make high-quality predictions with accuracy varying somewhere between 80 and 90 percent. Therefore, the created model can be a huge helping tool in the world of medicine, as with the help of it, Coronavirus can be diagnosed much more quickly, so more efficient treatments can be given to patients and many lives can be saved.

1 Introduction

COVID – 19 has been a huge disaster for the whole world. Breaking in our lives in the beginning of 2020 it changed all the aspects of it. Economics of all the countries suffered hugely, and the population of the people decreased a lot, because nothing was known about the virus, people were not ready for it which led to more than 4 million deaths[5].
Although vaccination has already started, it is still very important to be able to diagnose COVID-19 in early stages so that some help could be given to people and reduce the spread of the virus as early as possible.
The motivation for this project is to create a machine learning algorithm with the help of a data with information about COVID-19 symptoms and some other economical and regional factors from Brazil [2][3], which will take as an input all that information and will give a probability whether that person is likely to test positive or not. All that will help to diagnose the virus at an early stage and can help governments to define strategies that will reduce the bad impact from the virus on the country and its population.

The project will be executed by the following steps:
1) Translation of dictionaries about the data from Portuguese to English,
2) Exploratory data analysis on the data, to get insights about it and get fully acquainted to select the necessary variables for the model,
3) Creation of 3 different machine learning models –
Random Forest Classifier, K Nearest neighbor, XG boost classifier,
4) Comparison of the results of the created models and final model creation.
5) Overview of the results.

2 Analytical questions and data

The main objective of the analysis is to define a probability for each person in the dataset and to have a complete model that can take as an input necessary information about the person and return a probability of that person being infected with COVID-19.
The first step of the analysis was exploratory data analysis.

Here are some pieces of the information we gathered:

- Most of the people are from federation units Minas Gerais and Sao Paulo.
- Almost 600.000 people are from capitals of federation units, almost 400.000 are from Rest of the RM (Metropolitan Region, excluding the capital), approximately 50.000 people are from Rest of RIDE (Integrated Economic Development Region, excluding the capital), and most of the people, almost 1.600.000 are from Rest of UF (Federal Unit, excluding the metropolitan region and RIDE).
- Most people are in age group of "21-40", and number of men and women taking the survey is almost equal.
- Most of the test results are negative, but there is also solid number of positive tests, therefore this will allow to create a good model.

Also some other analysis will be done in the next step. For example we will determine the missing values in the dataset and will substitute them.

2 ANALYSIS

The first step of the analysis was data cleaning and preprocessing. First, only the rows in which there was a test result from at least one type of test were retained, so that later the model could be trained using this data. After filtering, the data remained with 185,921 rows.

Then, it was the turn to choose only the features that could be useful for creating the models. Therefore, all the variables in the dictionary were covered and the table below mentions the reasons for keeping or eliminating certain variables (all the results are made logically and not mathematically, also we have stored some variables in the same row as they would have the same values in the table). After removing the unnecessary variables, 34 variables (variables with label "Yes" label in the table) were left and 3 more variables where test results are stored.

If test results from the 3 types of tests are not different (meaning one is positive and one is negative), and a results from at least one test is available, then we will give "1" value to positive tests and "2" value to negative tests. A single column for test results was created, thus the variables have 35 columns left in the dataset.

Next step is dealing with missing values. For that, each missing value will be substituted by the mean value of that column. 34 variables are still too much, therefore correlation matrixes were created to find out which variables are more correlated with the target variable and remove unnecessary ones. As there are many variables, 2 correlation matrixes were created to be able to clearly observe the results.

Correlation is counted with the help of Python's Pandas library, which counts Pearson's correlation[6] . There are some variables that have very low correlation with the target variable and they will not help in creating the models, as they are not connected with the target variable, therefore they were removed, as they would not give any help when creating the model. The variables that have correlation less than 0.01 were removed, so there are only 19 features left.

We also solved the problems of outliers and oversampling.

The final data has 178.226 rows, 18 columns representing variables, and one column presenting the label (predictive variable).

Finally, when the final dataset in our hands, we can move on to creating the models to predict the probabilities of each person being infected with COVID-19. Three models were created: Random Forest Classifier[7] ,K Nearest Neighbor[8] and Gradient Boosting Classifier[9]. The three chosen model use different techniques for making predictions and the purpose is to compare the performances of those 3 models and use the best one for the final evaluations.

Before creating the models, an empty dictionary was created, where each models contains information about different metrics about the performance of the model. Some predictions were made on whether the particular person is infected with COVID or not and different metrics of performance based on those predictions were calculated.

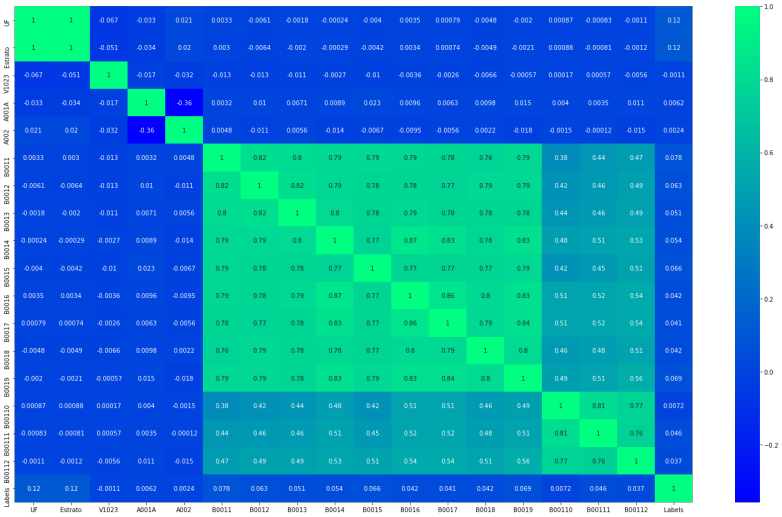

Figure 1 Correlation Matrix

For evaluation of the models, 5 metrics were picked:

- **accuracy score**, which is equal to true predictions divided by the number of all the observations,
- **precision**, which is equal to true predicted negative test cases divided by the number of all negative test guesses,
- **recall (sensitivity)**, which is equal to true predicted negative cases divided by all the number of negative cases,
- **specificity**, which is equal to the number of true positive test guesses divided by the number of positive cases
- **negative** predictive value, which is equal to the number of true positive test guesses divided by the number of all positive test guesses.

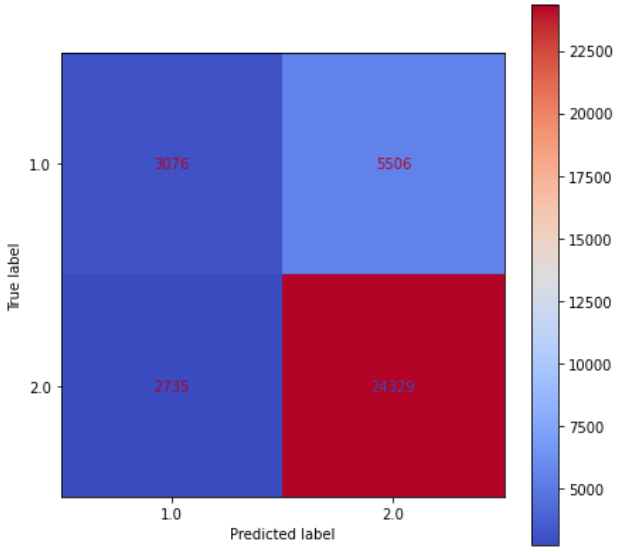

Figure 2 Confusion Matrix of expected results and predicition

Some functions to automate those processes were also created. First, a Random Forest Classifier model was created, and the following results from the model were observed:

```
Accuracy: 0.81
Precision: 0.84
Recall | Sensitivity: 0.93
Specificity: 0.44
Negative predictive value: 0.68
```

The results were double checked by doing cross-validation on the dataset and got the following scores:

```
Scores: [0.79358133 0.83254313 0.83604994
0.82847524 0.8202553]
Random forest average score: 0.82218098705
94614
```

The check passes normally, so Random Forest is a pretty good classifier as it gives results overall higher than 82%.

The next model was K Nearest Neighbor and got the followin g results:

```
Accuracy: 0.77
Precision: 0.82
Recall | Sensitivity: 0.9
Specificity: 0.36
Negative predictive value: 0.53
```

Again, cross-validation was performed to check the results:

```
Scores: [0.75074342 0.79231309 0.78978819
0.77724786 0.77542432]
KNN average score: 0.7771033763870309
```

KNN also performs great, but it is a little bit worse than Random Forest Classifier. The overall score for the KNN is 77%.

The last model created was Gradient Boost Classifier. The following result were obtained:

```
Accuracy: 0.77
Precision: 0.77
Recall | Sensitivity: 0.99
Specificity: 0.09
Negative predictive value: 0.69
```

The accuracy of the predictions:

```
Scores: [0.78283678 0.77393744 0.76978538
0.76675551 0.76821434]
Gradient Boosting Classifier average score
: 0.772305889539717
```

Gradient boosting gave a slightly worse results than KNN, again a little bit higher than 77%.

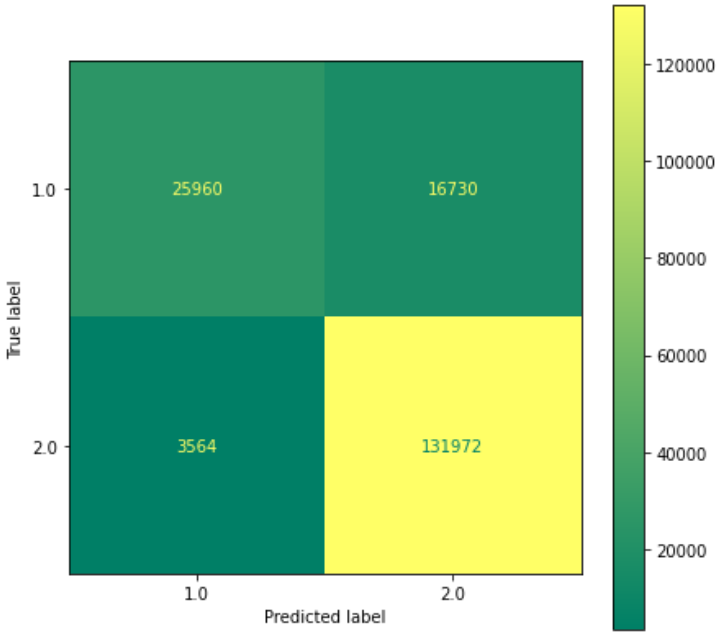

Figure 3 Training the final model

From the gathered results we saw that Random Forest Classifier performed best, so we applied hyperparameter tuning to it and applied it to the whole dataset, having an accuracy of almost 90%, after which we created an additional column in our initial dataset and added the corresponding probability for each survey-taker to it. The visualization below was executed to be informed with there was any problem with oversampling, in this case there was none.

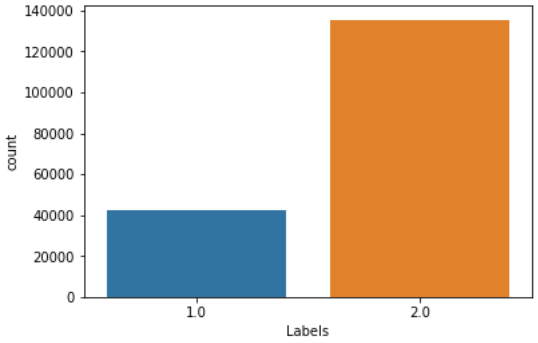

Figure 4 Oversampling

FINDINGS, REFLECTIONS AND FURTHER WORK

Predicting COVID is very complicated task, because the virus is new to everyone, and it is difficult to find the most important variables. Even the doctors not yet have clear image of what are the symptoms of the virus or what pre-symptoms it might have. That is the reason that COVID-19 was such a huge disaster for the whole world and now many health companies are still working on creating efficient vaccines that will help people to stay away from the virus.

As there is not much information about COVID-19 yet, machine learning is the thing that comes on the first place. By using, data scientists might be able to find insights about the data that are not visible by human eye and find the most important variables that can hint us that the person is infected with COVID-19. Therefore, the model created in this project can be a huge help for the whole world to stand even stronger against the Coronavirus.

The initial step of the project was just to create some charts and graphs to get acquainted with the data. Later came filtering some of the features by logic, because they surely don't have any connection with COVID-19. After it, some mathematical tactics were applied to keep only the most important ones. Particularly, correlation of the features was used with the target variable. After feature selection, 3 different models were created, and Random Forest classifier was the one with the best performance. Later hyperparameter tuning was applied to it, and the final model was created, which is a tool that can work with any amount of data. It returns the probability that the person is infected just by getting the necessary information from each row. It can be a powerful tool for preventing a large spread of the virus and detecting COVID in early stages.

The model created had very high accuracy and by doing surveys not only in Brazil but in other countries too, data scientist will be able to achieve even better results for future projects that will lead to an even higher accuracy

REFERENCES

[1] R. GARG, ‘Examples of classification models’. [Online]. Available: https://analyticsindiamag.com/7-types-classification-algorithms/.

[2] IBGE datasets, ‘Dataset used in the project’. [Online]. Available: https://www.ibge.gov.br/estatisticas/downloads-estatisticas.html?caminho=Trabalho%0D%0A_e_Rendimento/Pesquisa_Nacional_por_Amostra_de_Domicilios_PNAD_COVID19/Microdados/Dados%0D%0A.

[3] I. Dict, ‘Dictionaries about the dataset used’. [Online]. Available: https://www.ibge.gov.br/estatisticas/downloads-estatisticas.html?caminho=Trabalho_e%0D%0A_Rendimento/Pesquisa_Nacional_por_Amostra_de_Domicilios_PNAD_COVID19/Microdados/Documentacao%0D%0A.

[4] Deepak jain, ‘Information about data preprocessing’. [Online]. Available: https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/.

[5] ‘Statistics about coronavirus’. [Online]. Available: https://www.worldometers.info/coronavirus/.

[6] ‘Pearson correlation coefficient’. [Online]. Available: https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/.

[7] ‘Random Forest Classifier description’. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForest%0D%0AClassifier.html%0D%0A.

WORD COUNT

Abstract	143/150
Introduction	274/300
Analytical questions and data	219/300
Analysis	878/1000
Findings, reflections and further work	348/600