Feature Importance Investigation for Estimating Covid-19 Infection by Random Forest Algorithm

André Vinícius Gonçalves[1,5(✉)], Ione Jayce Ceola Schneider[2], Fernanda Vargas Amaral[3], Leandro Pereira Garcia[4], and Gustavo Medeiros de Araújo[5]

[1] Federal Institute of Northern Minas Gerais, Minas Gerais, Brazil
[2] Federal University of Santa Catarina, Araranguá, Brazil
ione.schneider@ufsc.br
[3] University of Malaga, Malaga, Spain
[4] Florianópolis Municipal Health Department, Florianópolis, Brazil
[5] PGCIN, Federal University of Santa Catarina, Florianópolis, Brazil
gustavo.araujo@ufsc.br

Abstract. The present work raises an investigation about the feature importance to estimate the COVID-19 infection, using Machine Learning approach. Our work analyzed 175 features, using the Permutation Importance method, to assess the importance and list the twenty most relevant ones that represent the probability of infection of the disease. Among all features, the most important were: i) the period comprised between the date of notification and symptom onset stand out, ii) the rate of confirmed in the territory of health units in the last 14 days, iii) the rate of discarded and removed from the health territory, iv) the age, v) variables of the traffic flow and vi) symptoms features as fever, cough and sore throat. The model was validated and reached an accuracy average of 78.19%, whereas the sensitivity and specificity achieved 83.05% and the 75.50% respectively in the infection estimate. Therefore, the proposed investigation represents an alternative to guide authorities in understanding aspects related to the disease.

Keywords: Feature importance · Feature engineering · Machine learning · Prediction model · COVID-19

1 Introduction

In December 2019, a new coronavirus, called SARS-CoV-2, was recognized in the city of Wuhan, China, and spread quickly to other countries in the world. In January 2020, the World Health Organization declared a Public Health Emergency of International Importance, and in March, the COVID-19 pandemic. At

E. Bisset Álvarez (Ed.): DIONE 2021, LNICST 378, pp. 272–285, 2021.
https://doi.org/10.1007/978-3-030-77417-2_20

the beginning of October 2020 there are already more than 33 million confirmed cases and more than 1 million deaths from the disease [14].

When infecting the human body, there is a period of latency, followed by an infectious period. During this period, the infected person can transmit to others through coughing and sneezing. The virus mainly affects the respiratory tract and the first symptoms appear after the incubation period. The main symptoms include fever, cough and fatigue, which appear on average after 11 days of contamination. Other symptoms, such as mucus production, headache, hemoptysis, diarrhea, dyspnoea, lymphopenia can also appear. The main clinical diagnosis is pneumonia [2,13,20,24]. Furthermore, the risk of symptomatic infection increases with age. Thus, older individuals are more likely to have symptomatic infection and worse outcomes [2].

Laboratory diagnosis is an important tool for diagnosis, as well as for follow-up, evaluation and evolution of the case. The recommended diagnostic test is the real-time polymerase chain reaction (RT-PCR) of nasal and oropharyngeal swab samples. Other serological tests can be used to detect immune responses, such as class M (IgM) and class G (IgG). However, it is important to use resources rationally in conducting diagnostic tests [26].

Towards the rational use of the infection spread of detection capabilities, artificial intelligence techniques have been used to predict the diagnosis of COVID-19. The algorithms are managing to predict the stage of COVID-19 by means of several features such as age, comorbidities, symptoms, diagnosis and outcome [7].

In order to create a model, we developed a investigation to assess the main features that can determine Covid-19 infection. To conduct our research, we collected and analysed data from the public health system in the capital of Santa Catarina, a state in southern Brazil. The set of features are composed of several variables from symptoms to demographic data.

Furthermore, we modeled a machine learning algorithm to estimate the infection of an individual. In our work, we conduct several experiments with 175 features to label the 20th most important features that represent the high Covid-19 infection likelihood.

1.1 Contributions

Among the contributions of our work, we can highlight:

1. The verification of the high importance of the features of confirmed, discarded, and removed by region of health, as well as the features of symptoms (fever, cough, and sore throat), all along the time of the notification date.
2. An intensive feature importance investigation results in findings that also highlighted the importance of traffic load, which reflects the people's isolation level.
3. The accuracy of the predictive model with an average of 78.19% of correctness in determining whether the individual is infected with Covid-19.

The remainder of this paper is structured as follows: In Sect. 2, we describe the more relevant related works on the effort to determine the Covid-19 infection; Sect. 3 introduces the methodology applied to feature engineering; Sect. 4 detail the experimental assessments; Sect. 5 outline the discussion about results and finally, in Sect. 6, we present our final remarks and future work.

2 Related Work

COVID-19 had a significant impact on the life and economy of several countries [10]. In addition to collapsing economies, the moral values of nations have been strongly affected by the pandemic [21]. All the impact, economic and social, motivated the Pan American Health Organization to seek to better understand the signs and symptoms of Sars-cov-2, in order to disseminate this knowledge. From now, the challenge of the pandemic is to find the best model that elucidates the initial growth trajectory and the epidemiological characteristics of the new coronavirus [19]. In this sense, the application of Forecasting models has been useful to deal with the dynamic behavior of this virus [22].

Sars-cov-2 is a respiratory virus transmitted through droplets of saliva, sneezing or by close contact. In their study, [25] described 69 cases of COVID-19 in China, where it was identified that 15% of individuals had fever, cough and dyspnoea. However, a survey conducted in the United States, showed that 50% of patients affected by this virus did not have a fever, however cough and dyspnea were reported by 88% of people with the virus [3]. Still, in other studies, reports of symptoms were difficult to measure objectively, such as anosmia (loss of smell), hyposmia (decreased smell) and ageusia (loss of taste) [11].

In addition, infected individuals may never develop symptoms, others may have mild symptoms or develop moderate to severe Sars-cov-2 disease [15]. In order to understand the symptoms that best represent the pandemic, researchers around the world try to understand the behavior of the virus [11]. A group of researchers from Spain found five patterns of skin infection that may be associated with COVID-19. These patterns were repeated in patients with varying demographic characteristics, in different periods and with different severities of the disease. Among these patterns are maculopapular rashes (47% of cases), vesicles or pustules (19% of cases), hives (19% of cases) and other vesicular rashes (9% of cases) and livedo or necrosis (in 6% of cases) [8].

A preliminary analysis by the World Health Organization (WHO) shows that in relation to gender, there is a relatively uniform distribution of infections between women and men (47% versus 51 respectively), however, it seems that men have a higher rate mortality rate (58%) in relation to women [15]. Nevertheless, due to the need to know the outbreak of COVID-19, some studies are being carried out considering exogenous factors such as the social environment, climatic variables, pollution and population density [22]. Other studies point to the role of room temperature in the survival and transmission of viruses. According to the WHO, several environmental factors can influence the spread of communicable diseases that can cause epidemics. The underlying theory is

that the number of cases and the spread of previous infectious viruses demonstrate seasonal patterns, affected by the climate, and therefore Covid-19 is likely to be similar in this respect [12].

Therefore, the prediction of a pandemic can be made based on several parameters, such as the impact of environmental factors, incubation period, impact of quarantine, age, sex and many more. The difficulty in predicting the number of cases of a pandemic is the fact that the number of cases to be studied does not match the total infected population. [17]. Considering the importance of knowing this difficult epidemiological scenario in a short-term horizon [22], Forecasting models have had a positive impact to mitigate the pandemic [21].

Forecasting techniques assess past situations, which allows for better predictions about the situation that will occur in the future [21]. These models allow managers to develop strategic planning and carry out decision making in the most assertive manner possible [12]. Since in addition to the concern with public health, the danger of the pandemic is also with the supply of food, medicine, and other supply chains needed by the population. Statistical analyzes such as forecasting allow governments not to focus only on underlying decision-making methods such as personal judgment [17].

To understand the nature of the coronavirus and predict its spread, the analysis models must be trained on a large volume of the data set. The ideal amount of the data set plays an extremely important role in training the task and affects the performance of the proposed algorithms [12]. The forecasting of COVID-19 cases is a challenging task, since Forecasting models are impacted by the effect of a small data set [22].

3 Methodology

The main goal of our work is to analyze which features most contribute to the diagnosis of suspected cases of COVID-19 using the classification technique with Machine Learning. The methodology steps that can be seen in Fig. 1 are: 1) data selection and extraction, 2) data pre-processing and feature engineering, 3) hyperparameterization and feature selection and 4) model validation. Each step is detailed further next.

3.1 Data Selection and Extraction

In the first step, the database used has been set corresponding to 1,930 reported cases of COVID-19 in the period between 02/15/2020 and 05/25/2020. The database was extracted from the Health Department of Florianópolis, capital of the State of Santa Catarina in southern Brazil and is available[1] to be analysed.

According to the [9], the database comes from three sources: 1) anonymized data on suspected and confirmed cases resident in Florianópolis; 2) demographic data of the 49 health regions that make up the municipality; and 3) data on the mobility represented by the traffic flow in the municipality.

[1] https://github.com/avgandre/covid_florianopolis/tree/master/dados/Dione.

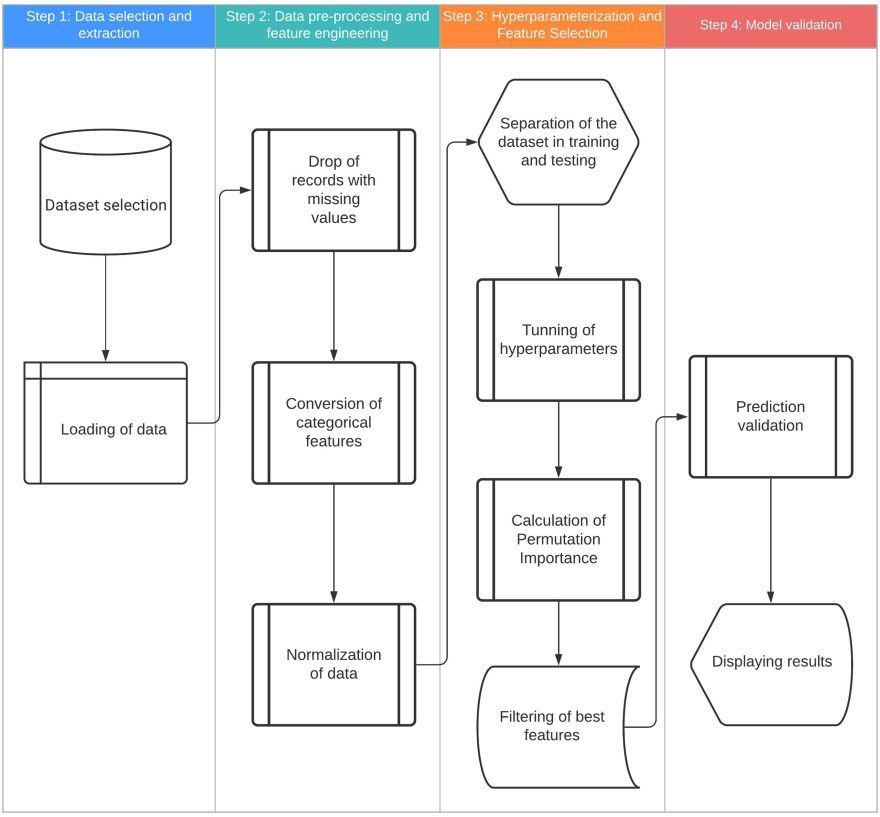

Fig. 1. Methodology flow

The database contains individual data on the diagnosis (confirmed or discarded), sex, age (in years), and age groups (under 10 years old, 10 years old under 20 years old, 20 years old under 40 years old) years old, 40 years old to under 60 years old, 60 years old to under 80 years old, 80 years old or more), skin color (white and not), date of onset of symptoms, in addition to the following clinical data of symptoms of the disease: pain throat, dyspnoea, fever and cough.

There is also data on health regions in the city of Florianópolis. There are 49 territories and 104 sub-territories that correspond to regional divisions of the city.

Furthermore, the database contains the following demographic data for health territories the total number of inhabitants and by sex; the number of persons aged 1 year, 2 years and so on up to 100 years or more; the number of people by skin color (white, black, yellow, Brazilian, indigenous and ignored); the number of people by years of schooling (from 1 to 17 years completed or more, in addition to literate, non-literate, literate through youth and adult literacy programs and with uninformed schooling); the total income per household, the

Table 1. Conversion of categorical features

Categorical feature	Factors	Method	New features
Race/Color	White; Yellow; Black; Parda; Unknown	One-hot-encoding	Race_Col0 to Race_Col4
Age range	<10; 10 <= *years* <20; 20 to 80 step 20; >80	One-hot-encoding	Age_range0 to Age_range5
Screening method	3 methods	One-hot-encoding	Screening_method0 to Screening_method2
Fever	Yes; No	One-hot-encoding	Fever0 to Fever1
Cough	Yes; No	One-hot-encoding	Cough0 to Cough1
Sore throat	Yes; No	One-hot-encoding	Sore_throat0 to Sore_throat1
Dyspnea	Yes; No	One-hot-encoding	Dyspnea0 to Dyspnea1
Health territory	48 regions	FeatureHasher	Health_territory0 to Health_territory19
Health subTerritory	104 subregions	FeatureHasher	Health_subterritory0 to Health_subterritory19

average income of the households, the total income of the heads of households, the average income of the heads of households, the total income per person and the average income per person, the proportion of males, persons with 60 years of age or more, of people with non-white skin and of people with 10 years or less of education, as possible indicators of vulnerability.

Regarding mobility features, the database provides data on the average daily traffic on four major avenues in the city. The time window for calculating the average considers it starts on the day of symptom detection until the thirteenth day before, that is, it is a window delayed in time.

3.2 Data Pre-processing and Feature Engineering

Initially, all records with the value 'Missing' in the attributes of symptoms (Sore Throat, Dyspnea, Fever and Cough) and Diagnosis were removed. Then, the categorical attributes were converted to numerical ones, using the One-hot-encoding technique for Race/Color, Age group, Screening Method and symptoms, and Feature Hashing [23] for Territory and Subterritory. The Table 1 has the conversion process result detailed:

Another procedure performed was the creation of new attributes. As suggested by [9], the number of infected people (with a positive diagnosis and up to 14 days after the onset of symptoms) in each health territory was calculated.

Moreover, according to the principle of the SIR model of epidemiology [4], it was proposed to include the number of people discarded (with a negative diagnosis) and the number of people removed (with a positive diagnosis and more than 14 days after the onset of symptoms).

Furthermore, it was included the rate of people infected by the number of inhabitants of their respective health territory, as well as the rate of discarded and removed rate. Finally, the data were normalized, transforming them to values within the range [0, 1] and, thus, establishing a common scale.

3.3 Hyperparameterization and Feature Selection

The database was divided into training and test basis, 70% for training and 30% for testing. As there is an imbalance in the amount of data between the discarded and confirmed cases, the first being a larger amount, the sample was balanced using the Undersampling technique.

In the training stage, cross-validation was adopted as a way to assess the model's generalization. According to [18], the technique consists of dividing the database into k folds, one of which is selected at a time to be the test set and the other k−1 are used as a training set. The test is repeated until each of the k folds is used as a test set. In the end, the accuracy is given by the average of the accuracy obtained for each of the k folds.

Hyperparametrization was performed using a random combination of parameters with 10 iterations in each tuning process. Accuracy was chosen as the maximization score.

After defining the parameters of the algorithm, the feature selection was performed considering the values of permutation importance as a criterion for assessing the degree of importance [1]. The criterion used was to select only those features with a value greater than zero. In this way, the features with values above this threshold remained in the model and the rest were removed from the database.

3.4 Model Validation

In the last step of the process, with the algorithm trained and configured with the best parameters that fit the model, the algorithm was validated with the test base to assess its prediction capacity.

Steps 3 and 4 were repeated 100 times and the results for each were stored. Then the data were used to calculate the mean and stantard devation of evaluation metrics and permutation feature importances.

The experiments were initially tested with three algorithms: Random Forest and Support Vector Machine (SVM). The equipment used to carry out the experiments had: i) Intel (R) Xeon (R) Gold 6126 CPU @ 2.60 GHz CPU with

12 CPUs, ii) 32.0 GB of RAM, iii) 250 GB of hard disk and iv) Linux Ubuntu 16.04. The entire implementation was developed in the Python programming language, version 3.7.

4 Experiment Assessments

We carried out experiments to analyze the evaluation metrics that measure the accuracy, in addition to ascertaining the features that had the most contribution to the performance.

The specific parameters of the Random Forest and SVM are presented in Table 2 and Table 3, as well as the possible value ranges. Through them, the best configuration is adjusted by means of a random search of hyperparameters.

Table 2. Random forest hyperparameters

Parameter	Value
criterion	[entropy, gini]
n_estimators	[5...100]
max_depth	[None, 1...5]
min_samples_split	[2...5]
min_samples_leaf	[1...5]
min_weight_fraction_leaf	[0.0...0.5]
max_features	['auto', 0.1...0.5]
Bootstrap	False, True

Table 3. SVM hyperparameters

Parameter	Value
C	[0.025...1.0]
degree	[3...25]
gamma	['scale', 'auto', 0.1...2.0]
shrinking	[True, False]
probability	[False, True]
decision_function_shape	[ovo, ovr]

And the parameters described in Table 4 relate to the general settings of the environment.

In the experiments, the metrics used in the analysis of the proposed model to assess performance were accuracy, sensitivity and specificity. The data samples were obtained by running the algorithm repeatedly and they are presented below in the form of average and standard deviation.

Table 4. General settings

Parameter	Value
Execution amount	100
Folds	10
Training/Test	70/30
RandomizedSearchCV interactions	10
Features selection threshold	Above zero (>0)

Among the two algorithms used, Random Forest was the one with the highest efficiency, as shown in Table 5. Therefore, it was extensively explored in this study.

Table 5. Metrics from accuracy score

	Random forest		SVM	
Metric	Training (M ± SD)	Test (M ± SD)	Training (M ± SD)	Test (M ± SD)
Accuracy	0,80682 ± 0,02279	0,79755 ± 0,02198	0,76733 ± 0,02285	0,75328 ± 0,02542
Sensibility	0,83802 ± 0,03916	0,84141 ± 0,04320	0,80397 ± 0,05217	0,80015 ± 0,05508
Specificity	0,77561 ± 0,02779	0,77332 ± 0,03545	0,73069 ± 0,05110	0,72740 ± 0,04980

The Fig. 2 shows the comparison between the evaluation metrics of the two algorithms: Random Forest and SVM. The data presented are only from the Test Set.

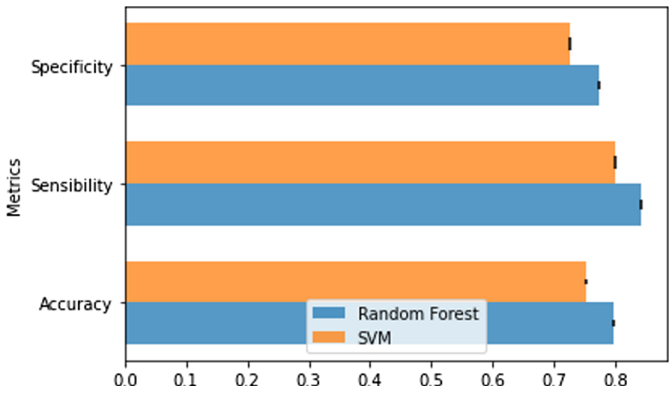

Fig. 2. Metrics from test set

The Random Forest algorithm performed better compared to SVM. It had an accuracy of 0.80682 ± 0.02279 (mean ± standard deviation) on the training set and 0.79755 ± 0.02198 on the test set. The sensitivity was 0.83802 ± 0.03916 and 0.84141 ± 0.04320 in each of the two bases, respectively. The specificity was 0.77561 ± 0.02779 in the training base and 0.77332 ± 0.03545 in the test base.

The main features selected by Random Forest with their respective Permutation Importance percentages are shown in Table 6. The results presented below are in the form of average and standard deviation for the set of executions.

Table 6. Features permutation importance of accuracy score

Position	Feature	Permutation importance (M ± SD)
1[a]	Notification date	0,05412 ± 0,03538
2[a]	Confirmed rate territory 14 days	0,03016 ± 0,01456
3[a]	Fever1	0,02783 ± 0,01964
4[a]	Fever0	0,02638 ± 0,01992
5[a]	Cough0	0,01575 ± 0,01256
6[a]	Cough1	0,01505 ± 0,01299
7[a]	Confirmed territory 14 days	0,01100 ± 0,00677
8[a]	Symptoms start date	0,00454 ± 0,00488
9[a]	Discarded rate territory	0,00247 ± 0,00303
10[a]	Sore throat 1	0,00166 ± 0,00226
11[a]	Removed rate territory	0,00140 ± 0,00195
12[a]	Sore throat 0	0,00140 ± 0,00202
13[a]	Age	0,00123 ± 0,00165
14[a]	Discarded territory	0,00115 ± 0,00161
15[a]	Age above 91	0,00108 ± 0,00253
16[a]	Removed territory	0,00105 ± 0,00164
17[a]	traffic mean	0,00098 ± 0,00142
18[a]	traffic mean lag2	0,00088 ± 0,00125
19[a]	traffic mean lag1	0,00084 ± 0,00125
20[a]	Screening method 2	0,00081 ± 0,00122

For a better visual understanding of the features importances, the Fig. 3 is shown with the values of each variable.

Lastly, the response time of the algorithm had an average result of 17.49 and standard deviation of 4.87 s, considering the training step that involved the hyperparameter tuning process and feature selection, in addition to the test step that consisted of the model validation.

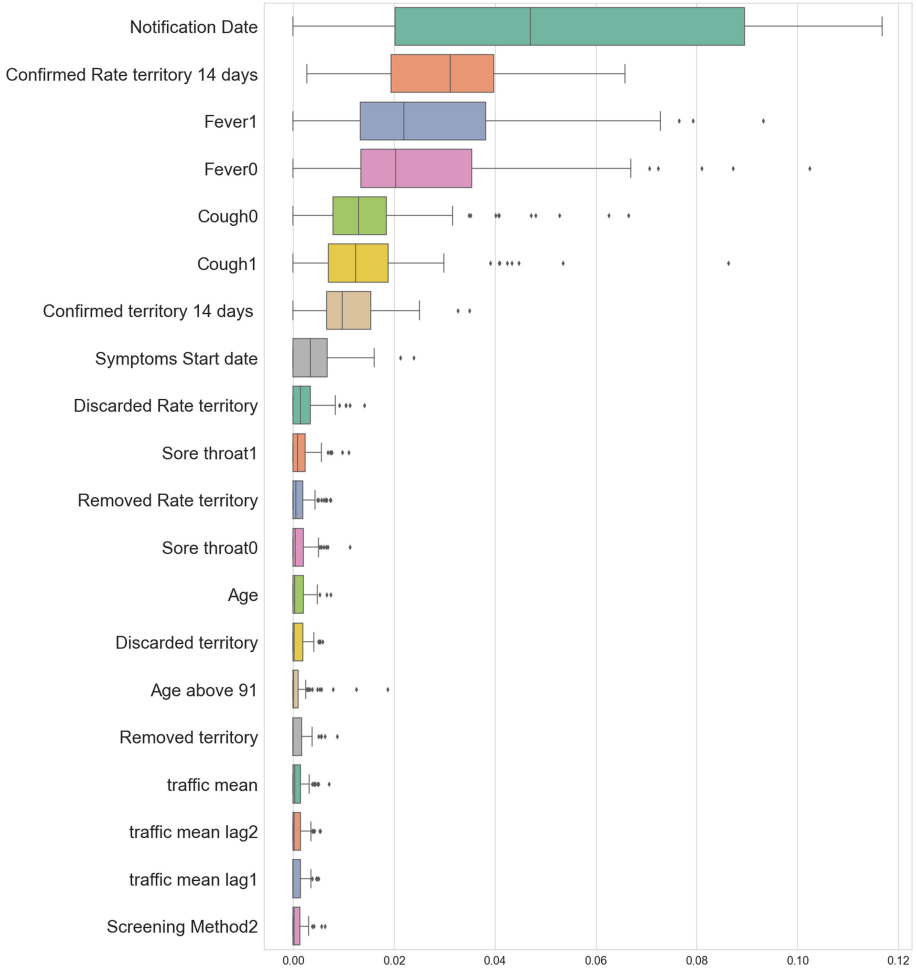

Fig. 3. Features permutation importance of accuracy score

5 Discussion

After nearly a year of its discovery, COVID-19 is a disease that arouses much interest because of its great impact on humankind. Wherefore, this work proposed to investigate the relevance of a set of variables in the diagnostic prediction of the disease.

The investigation started with the acquisition of a preliminary database with 175 features, which after going through pre-processing, increased to 228 due to the techniques of coding categorical variables. Then, the model was processed and analyzed by the Permutation Feature Importance method to assess the impact of each feature on the accuracy metric.

The most important feature was the Notification Date. All symptom features appeared among the twenty most significant, with the exception of dyspnea. This fact corroborates with the researches that investigate the symptoms and indicate fever, cough and sore throat among some of the most common ones [5,6,16].

Mobility features also meaningful. There were fourteen features to represent traffic on the city's four major avenues. Thus, four of them are among the most important features, listed in Table 6

The feature engineering process carried out in the pre-processing step resulted in the creation of new health territory variables (Confirmed_territory_14days, Removed_territory and Discarded_territory, Confirmed_rate_territory_14days, the Rate_Discarded_Territory and the Rate_Removed_Territory) contributed significantly to the model performance.

However, the Confirmed_territory_14days, the Rate_Discarded_Territory and the Rate_Removed_Territory features were even more expressive and were among the ten most important. In this way, it is noticed that the factor "health region" was very important in the correct classification of the algorithm's diagnosis.

Finally, the results of the model were somewhat satisfactory. Coping with the research developed in [9], there is an improvement of approximately 15% in the accuracy measured in the final test stage. This matter may be associated with the addition of new variables that were not present in the previous work, including symptoms and those confirmed, discarded and removed related to the health territory.

6 Final Remarks and Future Work

The present work shows a investigation about the feature importance in a prediction diagnostic model for cases of COVID-19, using the classification technique with Machine Learning.

These classification approaches are fundamental for monitoring the number of virus reproductions and for making decisions in the face of the pandemic. The advantage of them is to produce quick responses and relatively low cost compared to laboratory diagnosis.

The methodology section emphasized the hyperparametrization and feature selection techniques, as the research aimed to investigate two aspects: the features that best contributed to the performance of the model and the results of the hit rates in the validation of the test step.

The Random Forest performed better compared to SVM. So it was explored in more detail. In the first investigation step, the Permutation Importance method was used to assess the impact of the features on the results. Among the 228 variables that make up the database, the most relevant are: the date of notification and onset of symptoms, the rate of confirmed in the territory of the last 14 days, the rate of discarded and removed from the territory, age, flow variables traffic, in addition to the attributes of fever, cough and sore throat.

Regarding the second stage of the investigation, the metrics showed consistent results. Accuracy had a mean of 78.19%, whereas sensitivity reached 83.05% and specificity 75.50% of cases.

Therefore, the research conducted has shown that there is a feasible alternative in the process of underdiagnosis COVID-19 disease, considering the most relevant characteristics for the determination of infection.

In near future, we pretend to extend the model including new features, such as climatic conditions, as [22] suggests that could have an impact on people behavior. Another possibility is to reduce the granularity of the health territory to sub-territories in the calculation of confirmed, discarded and removed and, later, to analyze the behavior of the algorithm.

References

1. Altmann, A., Toloşi, L., Sander, O., Lengauer, T.: Permutation importance: a corrected feature importance measure. Bioinformatics **26**(10), 1340–1347 (2010)
2. Ashour, H.M., Elkhatib, W.F., Rahman, M., Elshabrawy, H.A., et al.: Insights into the recent 2019 novel coronavirus (SARS-CoV-2) in light of past human coronavirus outbreaks. Pathogens **9**(3), 186 (2020)
3. Bhatraju, P.K., Ghassemieh, B.J., Nichols, M., Kim, R., Jerome, K.R., Nalla, A.K., Greninger, A.L., Pipavath, S., Wurfel, M.M., Evans, L., et al.: Covid-19 in critically ill patients in the seattle region-case series. New Engl. J. Med. **382**(21), 2012–2022 (2020)
4. Brauer, F.: The Kermack-McKendrick epidemic model revisited. Math. Biosci. **198**(2), 119–131 (2005)
5. Burke, R.M., et al.: Symptom profiles of a convenience sample of patients with covid-19–United States, January–April 2020. Morb. Mortal. Wkly Rep. **69**(28), 904 (2020)
6. Carfì, A., Bernabei, R., Landi, F., et al.: Persistent symptoms in patients after acute covid-19. JAMA **324**(6), 603–605 (2020)
7. Chisari, E., Krueger, C.A., Barnes, C.L., Van Onsem, S., Walter, W.L., Parvizi, J.: Prevention of infection and disruption of the pathogen transfer chain in elective surgery. J. Arthroplasty **35**(7), S28–S31 (2020)
8. Galván Casas, C., et al.: Classification of the cutaneous manifestations of covid-19: a rapid prospective nationwide consensus study in Spain with 375 cases. Br. J. Dermatol. **183**(1), 71–77 (2020)
9. Garcia, L.P., et al.: Estimating underdiagnosis of covid-19 with nowcasting and machine learning: experience from Brazil. medRxiv (2020)
10. He, S., Tang, S., Rong, L.: A discrete stochastic model of the covid-19 outbreak: forecast and control. Math. Biosci. Eng. **17**, 2792–2804 (2020)
11. Iser, B.P.M., Sliva, I., Raymundo, V.T., Poleto, M.B., Schuelter-Trevisol, F., Bobinski, F.: Definição de caso suspeito da covid-19: uma revisão narrativa dos sinais e sintomas mais frequentes entre os casos confirmados. Epidemiologia e Serviços de Saúde **29** (2020)
12. Malki, Z., Atlam, E.S., Hassanien, A.E., Dagnew, G., Elhosseini, M.A., Gad, I.: Association between weather data and covid-19 pandemic predicting mortality rate: machine learning approaches. Chaos, Solitons Fractals **138**, 110137 (2020)
13. Meo, S., et al.: Novel coronavirus 2019-nCoV: prevalence, biological and clinical characteristics comparison with SARS-CoV and MERS-CoV. Eur. Rev. Med. Pharmacol. Sci. **24**(4), 2012–2019 (2020)
14. World Health Organization: WHO coronavirus disease (COVID-19) dashboard (2020). https://covid19.who.int/

15. Organization, W.H., et al.: Diagnostic testing for sars-cov-2: interim guidance, 11 September 2020. World Health Organization, Technical report (2020)
16. Pan, L., et al.: Clinical characteristics of covid-19 patients with digestive symptoms in Hubei, china: a descriptive, cross-sectional, multicenter study. Am. J. Gastroenterol. **115** (2020)
17. Petropoulos, F., Makridakis, S.: Forecasting the novel coronavirus covid-19. PLoS ONE **15**(3), e0231236 (2020)
18. Rodriguez, J.D., Perez, A., Lozano, J.A.: Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE Trans. Pattern Anal. Mach. Intell. **32**(3), 569–575 (2009)
19. Roosa, K., et al.: Real-time forecasts of the covid-19 epidemic in china from February 5th to February 24th, 2020. Infect. Dis. Model. **5**, 256–263 (2020)
20. Russell, C.D., Millar, J.E., Baillie, J.K.: Clinical evidence does not support corticosteroid treatment for 2019-NCoV lung injury. Lancet **395**(10223), 473–475 (2020)
21. Shinde, G.R., Kalamkar, A.B., Mahalle, P.N., Dey, N., Chaki, J., Hassanien, A.E.: Forecasting models for coronavirus disease (covid-19): a survey of the state-of-the-art. SN Comput. Sci. **1**(4), 1–15 (2020)
22. da Silva, R.G., Ribeiro, M.H.D.M., Mariani, V.C., dos Santos Coelho, L.: Forecasting Brazilian and American covid-19 cases based on artificial intelligence coupled with climatic exogenous variables. Chaos, Solitons Fractals **139**, 110027 (2020)
23. Soheily-Khah, S., Wu, Y.: A novel feature engineering framework in digital advertising platform **10**, 21 (2019). https://doi.org/10.5121/ijaia.2019.10403
24. Vannabouathong, C., et al.: Novel coronavirus covid-19: current evidence and evolving strategies. J. Bone Joint Surg. Am. **102**(9), 734 (2020)
25. Wang, Z., Yang, B., Li, Q., Wen, L., Zhang, R.: Clinical features of 69 cases with coronavirus disease 2019 in Wuhan, china. Clin. Infect. Dis. **71**(15), 769–777 (2020)
26. Xavier, A.R., Silva, J.S., Almeida, J.P.C., Conceição, J.F.F., Lacerda, G.S., Kanaan, S.: Covid-19: clinical and laboratory manifestations in novel coronavirus infection. Jornal Brasileiro de Patologia e Medicina Laboratorial **56** (2020)