UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

RUAN LEITÃO NUNES

Predicting the performance of countries in managing the COVID-19 pandemic using WVS data

Work presented in partial fulfillment of the requirements for the degree of Bachelor in Computer Science

Advisor: Prof. Dr. Augusto Couto Barone Dante
Coadvisor: Prof. Ms. Eduardo Gabriel Cortes

Porto Alegre
November 2021

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Reitor: Prof. Carlos André Bulhões
Vice-Reitora: Profª. Patricia Pranke
Pró-Reitora de Graduação: Profª. Cíntia Inês Boll
Diretora do Instituto de Informática: Profª. Carla Maria Dal Sasso Freitas
Coordenador do Curso de Ciência de Computação: Prof. Rodrigo Machado
Bibliotecária-chefe do Instituto de Informática: Beatriz Regina Bastos Haro

“É necessário sempre acreditar que o sonho é possível,
que o céu é o limite e você, truta, é imbatível”
— Racionais MC‘s

ACKNOWLEDGEMENT

Thanks to all people that crossed my path during this long period, each of them contributed in a way or another to the person I turned into.

Firstly, thanks to my sister Lyz, who always stimulated me to be curious about how things work.

Special thanks are to my wife Nana, who encouraged me to not give up. Other special thanks to my friend Flora, who helped me during all courses, and in my life.

Could not forget to thank my child, that is not even born, but already makes me want to be someone they can be proud of.

Another spacial thanks to those who helped me on the last mile providing support and insights on this work: Cristal, Mariana, Valter, Julia and Vinicius.

Finally, I want to thank all professors and the Institute of Informatics for all the learning and experiences provided.

ABSTRACT

With the huge and sad impact that COVID-19 caused in everyone's lives, it is important to reason about the context that led to such several fatalities and the different impacts suffered by different countries. The more information available, the more resources researchers have to propose newer hypotheses and ideas to reduce the future impact of COVID-19 or newer viruses.

This work made use of a data science approach to analyze the relation of the perceived values of the society (through the data from the World Values Survey - WVS) and the impact of the COVID-19 pandemic in a list of countries. The approaches analyzed were the Logistic Regression and the Random Forest, both widely studied in academia and utilized in the market.

An analysis of the performance and configuration of the studied models was provided, identifying prediction results significantly better than the baseline stratified model. The work as well purposed future improvements and possibilities.

Keywords: COVID. WVS. Machine Learning. Data Science.

Predizendo o desempenho de países lidando com a pandemia de COVID-19 usando dados do WVS

RESUMO

Com o enorme e triste impacto que a COVID-19 causou na vida de todos, é importante refletir sobre o contexto que levou a tantas mortes e aos diferentes impactos sofridos pelos diferentes países. Quanto mais informação estiver disponível, mais recursos terão os pesquisadores para propor novas hipóteses e ideias para reduzir o impacto futuro do COVID-19 ou de um novo vírus.

Este trabalho fez uso de uma abordagem de ciência de dados para analisar a relação entre os valores percebidos da sociedade (através dos dados do *World Values Survey* - WVS) e o impacto da pandemia da COVID-19 numa lista de países. As abordagens analisadas foram a Regressão Logística e a Floresta Aleatória, ambas amplamente estudadas no meio acadêmico e utilizadas no mercado.

Foi fornecida uma análise do desempenho e da configuração dos modelos estudados, identificando uma capacidade preditiva significativamente superior ao modelo estratificado utilizado como base. O trabalho também propôs melhorias futuras e possibilidades.

Palavras-chave: COVID, WVS, Aprendizado de Máquina, Ciência de dados.

LIST OF FIGURES

LISTINGS

LIST OF ABBREVIATIONS AND ACRONYMS

WVS	World Values Survey
COVID-19	Coronavirus disease 2019
VT	Variance Threshold
MGI	Mutual Gain of Information
MCC	Matthews correlation coefficient

CONTENTS

1 INTRODUCTION

Based on the impact of the COVID-19 pandemic and the importance of the social and political context on such event, this work aims to make use of data science techniques to get a better understanding of the relation of the social and political contexts of countries, represented by the World Values Survey, and the impact of the COVID-19 in the countries, utilizing available COVID-19 data (as will be described in Section 2.2).

This work presents a data science approach for analyzing the possible correlation between the data available in the World Values Survey and how well each country handled the COVID-19 pandemic. Its research question can be summarized as: Is it possible to predict how well a country would handle the COVID-19 pandemic, utilizing the WVS data?

To deal with this question, diverse techniques of data science and machine learning were used. Some of them are data cleaning, feature imputation, feature selection, logistic regression, random forests, and cross-validation. Each of the utilized techniques is described in more detail in Section 2.3.

In summary, the data of both WVS and COVID datasets were merged and submitted to a pipeline that executes a series of processing steps. The data was first normalized, imputed, and invalid answers cleaned, and then submitted to feature selection and hyperparameter tuning for both logistic regression and random forest.

In the end, these final models are trained with the best-found parameters and evaluated against a baseline stratified classifier, that guesses the resulting class randomly based on the distribution of classes of the dataset. To compare the models, the f1-score and the Matthews Correlation Coefficient were used, as well as comparing the confusion matrix of each test.

The text starts presenting a overview of the related work and theoretical background in Chapter 2. At Chapter 3 the methodology utilized to work with the datasets and to evaluate the results is presented. Chapter 4 describes the obtained results for each trained model during the work. Next, at Chapter 5 the results are summarized and analyzed. Finally, at Chapter 6 future improvements and new ideas that can increase the value of the work are provided.

2 BACKGROUND AND RELATED WORK

This chapter will present the core concepts and theoretical background that serve as the basis of this work. Each section will explore the existing research on the topic and its impact on this work.

2.1 World Value Survey - WVS

The World Value Survey, WVS, is a public opinion survey that aims to identify current and time-series trends in values and the social and political people's perception of their country and the world as a whole. (INGELHART et al., 2014)

The survey uses a common questionnaire, adapted and translated to the native language of each of the almost 100 surveyed countries. The same base questionnaire is used since the first wave of the survey back in 1981. The WVS is the largest cross-national time series investigation of human values, currently interviewing almost 400,000 respondents, covering a large range of economic and cultural national backgrounds.

The WVS helps scientists and policymakers to understand the beliefs, values and motivations of the people and their change over time. This data has been widely used to analyze topics as democratization, patriotism, religion and subjective well-being. (WELZEL, 2021) (VIER, 2020) (NEZLEK, 2021)

For this work, the data from wave 6, which run from 2011 to 2014 was used as it is the most recent data with no overlap to the COVID-19 pandemic.

2.2 COVID-19

The Severe acute respiratory syndrome coronavirus 2, SARS-CoV-2, is a highly transmissible coronavirus that was first spotted in late 2019 and has caused a pandemic of acute respiratory disease, named 'coronavirus disease 2019', COVID-19 for short, which is having a huge impact in public health and global economics. (HU et al., 2020)

There are multiple ways to analyze the impact of the COVID-19 pandemic in society, such as the number of deaths, the variation in the Gross Domestic Product — GPD —, the length of the lockdowns and others.

For the purpose of this work, it was utilized the scores generated by the Lowy

Institute, the methodology described in (LENG; LEMAHIEU,), which generates a score from 0 to 100 per country, based on diverse available metrics, such as deaths per million, cases per million and tests per thousand. The utilized COVID-19 dataset was formed considering the available data up to 9th of January 2021.

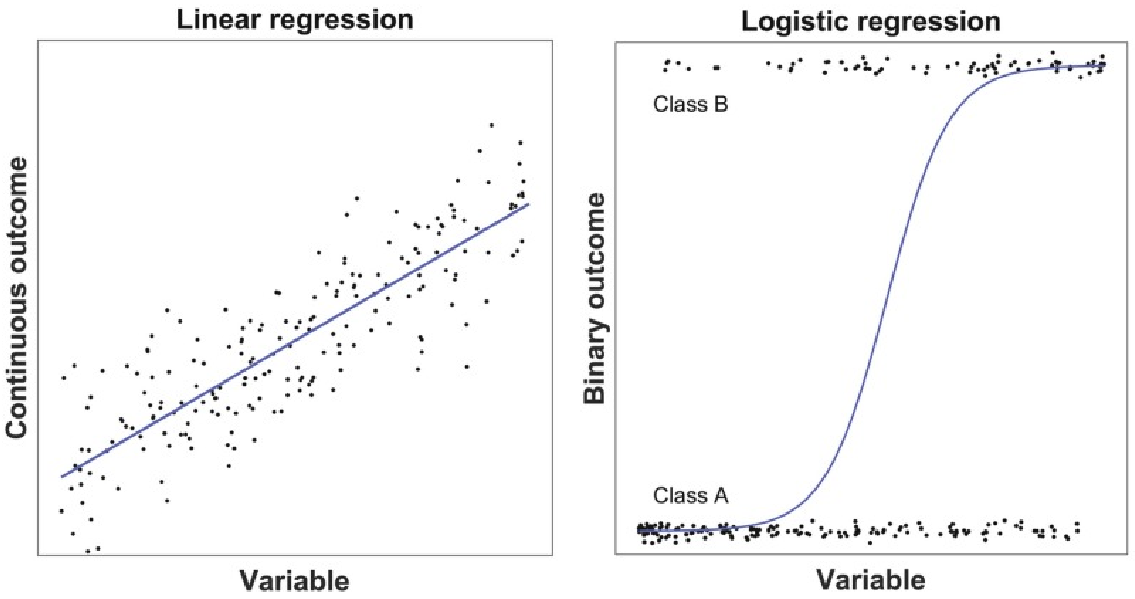
2.3 Machine Learning

2.3.1 Logistic Regression

Logistic regression is a statistic model that uses a logistic function to model the probability of given data to be of a certain class. Similar to linear regression, logistic regression tries to model the output class as a function of the input data, but differently than linear regression, which produces a linear function, it produces a logistic function whose values are bounded between 0 and 1. The predicted class of the functions is defined if the output of the function is closer to 0 or 1. (PENG; LEE; INGERSOLL, 2002)

It is desired to the generated function to have most of the input values being really close to the boundaries (0 or 1) representing one of 2 classes. This characteristic is usually achieved through a Sigmoid function, being asymptotic on 0 and 1 as $x \rightarrow \pm\infty$. (GÉRON, 2019)

Figure 2.1 – Comparison of the function generated by linear and logistic regression



Source: (DANKERS et al., 2018)

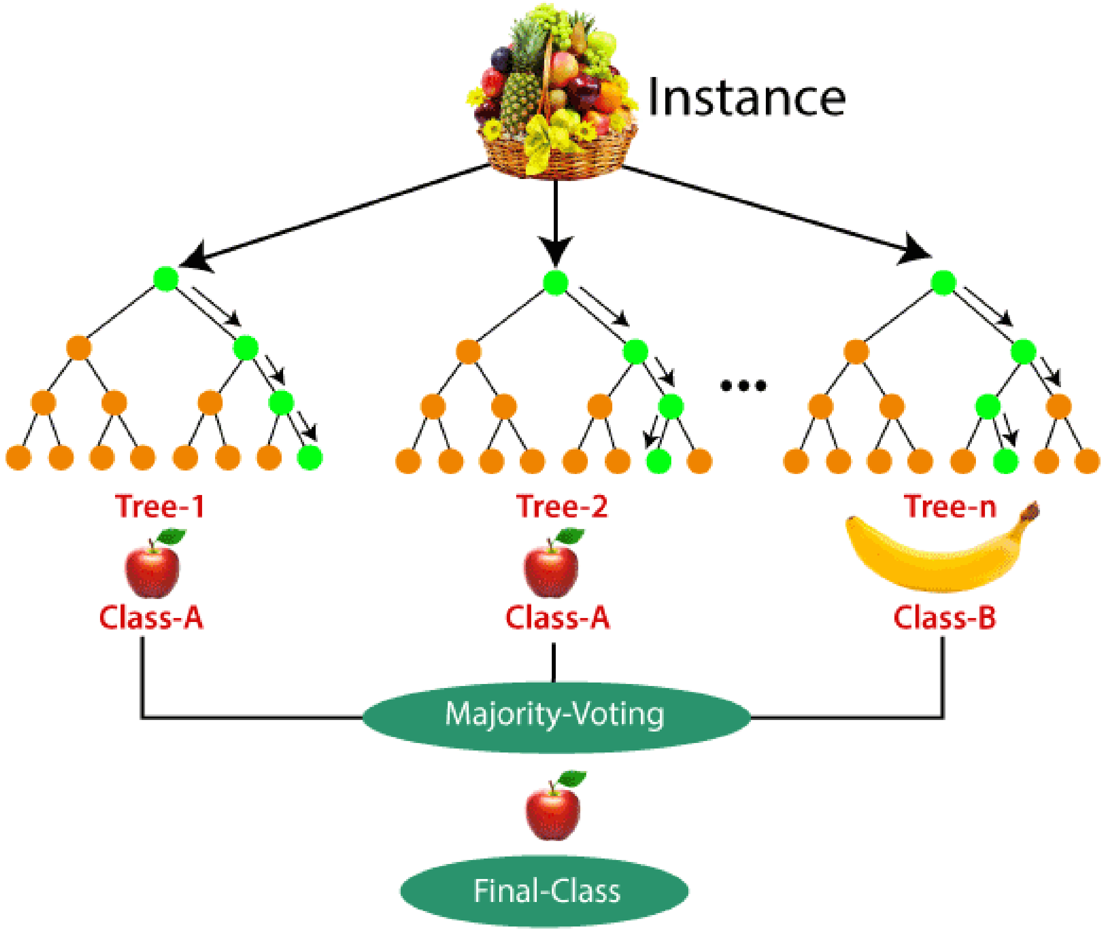
2.3.2 Random Forest

Random forest is an ensemble learning technique that creates diverse instances of Decision trees using a different subset of features to be trained.

A decision tree is a tree-like structure, often a binary one, wherein each node of them checks one condition about the input data: if the condition is true, the algorithm continues the evaluation to one side of the tree, if it is false then if follows to the other side.

The random forest is an ensemble technique — where multiple trained instances of a model vote for a class, and the most voted class wins — that extends the decision tree technique as it combines multiple trees trained with different subsets of the data, improving its accuracy and reducing the probability of over-fitting. (GÉRON, 2019)

Figure 2.2 – Example of Random Forest behavior



Source: (MBAABU, 2020)

2.3.3 Feature Normalization

Feature normalization is a technique of getting features of your dataset to a common range of values, usually between 0 and 1. This technique is valuable for diverse algorithms that use the absolute value of the features to determine their impact, such as K-NearestNeighbours. A feature with a range of values bigger than others can end having a greater than expected weight in the resulting model if not scaled to a common range.

Another option widely utilized, mainly when features have a close to a normal distribution, is the Feature Standardization, where each value is transformed in its deviation from the median. (GÉRON, 2019)

2.3.4 Feature Imputation

Feature imputation is a technique of inferring missing data in a dataset based on the existing data. The most common techniques use the average or the median of the values of the column to fill the missing data.

For this job, the dataset was segmented during the imputation, utilizing the mean of the values of the columns being imputed that belong to the same country of the imputed entry. This strategy was utilized as a form to avoid cross-country information in the dataset.

2.3.5 Feature Selection

Not all variables in a dataset are of equal importance, and some of them are not important at all. These additional variables increase the cost of training the model and can even cause negative effects to the model. Reducing the variable count and keeping only the most relevant and influential can increase the final model quality. (ANDERSEN; BRO, 2010)

There are multiple techniques to choose a subset of variables of a dataset, such as Recursive Feature Elimination, Variance Threshold, and Mutual Gain of Information.

In this work, both Variance Threshold, VT, and Mutual Gain of Information, MGI, were applied. The VT technique consists in removing from the dataset all variables where their variance is below or equals a certain threshold. On the other hand, MGI checks what variables of the dataset provide more information about the target variable, and select the K best variables.

Both of these techniques are independent of the model being used, but their hyper-parameters — the variance threshold for VT and the K for the MGI – were tuned during the Hyper-parameter selection of each learning model.

2.3.6 Hyper-Parameter selection

Each machine learning technique has its own set of hyperparameters — parameters whose control the training process itself — and changes in these parameters will impact the resulting trained model quality.

Hyperparameters can make a model too complex — for example increasing the depth of a Decision Tree —, leading to over-fitting, or too simple — for example, using few features on a Decision Tree —, leading to poor accuracy.

There are diverse techniques for choosing the hyperparameter values for a model, such as Grid Search, Random Search and Genetic Algorithms. For the purpose of this work, the Grid Search was chosen due to its simplicity and reliability, which comes with at the cost of time, as it explores all provided space of hyperparameters. (CLAESEN; MOOR, 2015)

The Grid Search technique receives as input a set of possible values for each hyperparameter being tuned, and runs cross-validation of the model with each possible combination of parameters and selects the best performing one. (BERGSTRA; BENGIO, 2012)

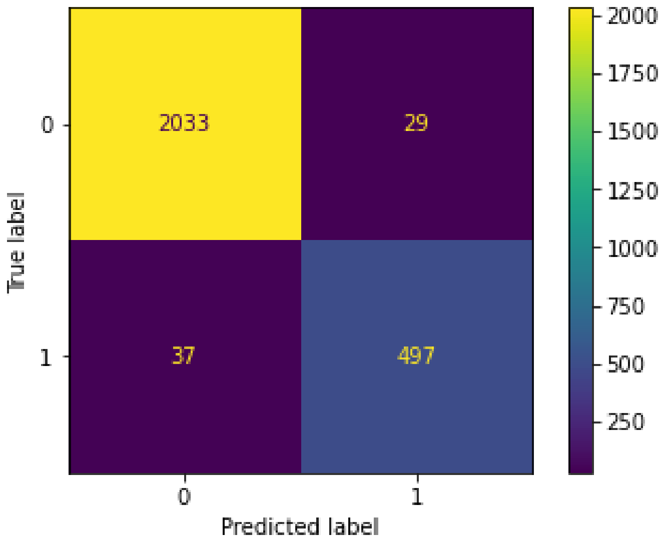
For each learning model tested, both the feature selection and the model hyperparameters were tuned, to capture differences in each model needs in feature selection.

2.3.7 Model Evaluation

For binary classification problems, as it is in this work case, there are 4 possible outcomes for a prediction: a true positive, a true negative, a false positive and a false negative.

The count or proportion of each of these outcomes are often plotted on a confusion matrix such as Figure 2.3, giving an easy visualization of the model quality

Figure 2.3 – Confusion Matrix example



Source: The Author

From this data it is possible to calculate some metrics:

- Accuracy, defined as $\frac{true positives + true negatives}{all predictions}$
- Precision, defined as $\frac{true positives}{true positives + false positives}$
- Recall, defined as $\frac{true positives}{true positives + false negatives}$

When training a model, it is interesting to have a single metric that gives an overall performance of the model under evaluation. Between the existing metrics in the literature, 2 were analyzed in this work, F1-score and Matthews correlation coefficient, MCC.

F1-score is a function of the above metrics, provided by $2 * \frac{precision * recall}{precision + recall}$, where both precision and recall are being considered with the same weight. There are variations of this formula to give more importance for precision or recall depending on your application. (WALSH; RIBEIRO; FRANKLIN, 2017)

MCC is a function of the confusion matrix data, in the form

$$\frac{(TP * TN - FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

This function aims to reduce overly optimistic or statistically imprecise results in imbalanced datasets generated by F1-score. (CHICCO; JURMAN, 2020)

2.4 Related Work

From the bibliography studied during this work, two works that made use of the WVS dataset were found really insightful and worth highlighting here.

(NASCIMENTO; BARONE; CASTRO, 2019) made use of the WVS dataset to identify patterns in social activism in different countries, targeting the usage of machine learning in the social science field. The pipeline utilized in that work served as the base for this work pipeline, as well as the author provided insightful suggestions on the usage of the dataset.

(VIER, 2020) as well, utilized the WVS dataset under the light of social science, but in that work as a tool for helping to understand the data and formulate hypotheses. The approach of this work on using machine learning as a tool to help scientists to formulate hypotheses was at the core of this work, where the focus was on creating models that provide some explanation of its working as well, helping researchers to understand what features are being used and reason on the hypotheses of why they were chosen.

Although both works made use of machine learning techniques in the field of social sciences, utilizing the WVS dataset as the base, none of them combine the WVS data with another source of data, neither are applied to the current COVID-19 pandemic. These two key features differentiate this work from these selected works, expanding on the uses of the WVS data.

3 METHODOLOGY

The methodology of this work consisted of a list of steps as shown in Figure 3.1, briefly described as:

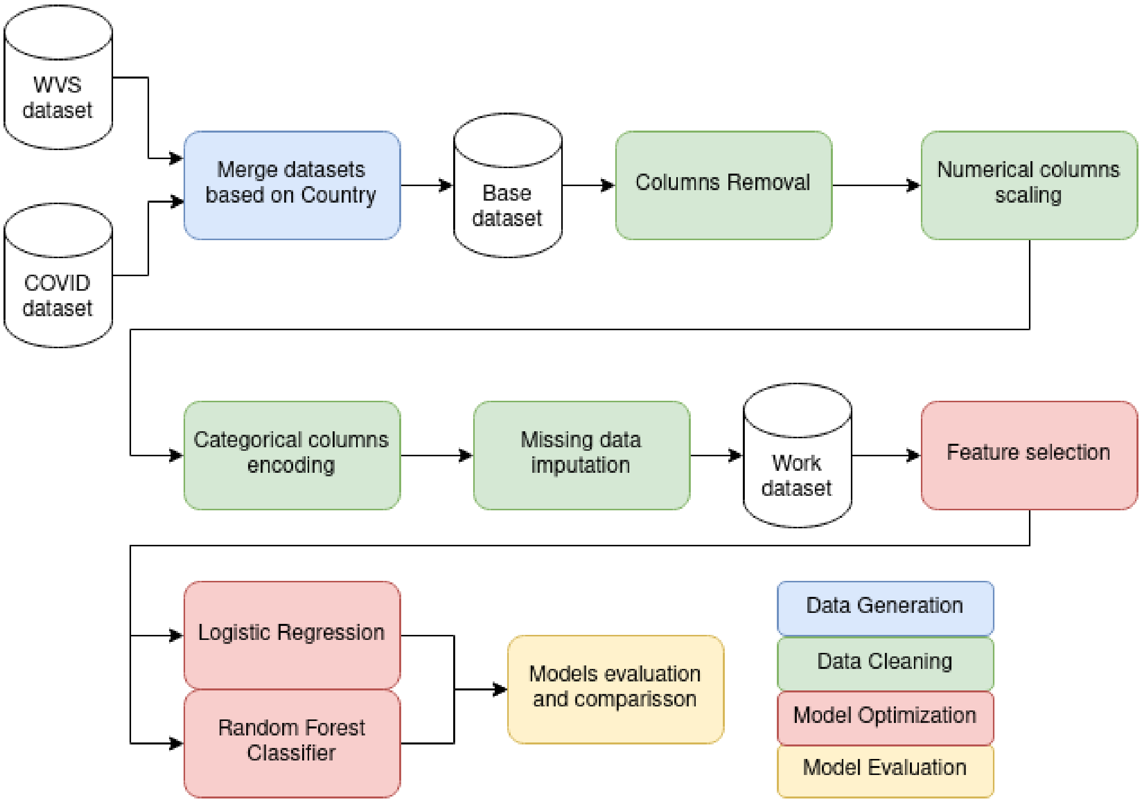
- Data Exploration – this step was not an explicit step as it happened during all the steps of this work. This consists of the hands-on analysis of characteristics of the dataset for gaining information and insights.
- Data Generation and Cleaning – where the 2 utilized datasets were joined together into a novel dataset. After that, some columns were removed (such as country code, interview number or interview year) as they or do not provide information or are too specific and could lead to over-fitting, all invalid responses were removed from the dataset, then the numerical columns were scaled to the range $[0, 1]$, after that the categorical columns of the dataset where encoded utilizing the One-Hot encoding technique.
- Machine Learning Pipeline – during this step the dataset is used as input for different learning models and variations on their hyperparameters, utilizing cross-validation to choose the best combinations of hyperparameters for each model.
- Model evaluation – an individual evaluation of each trained model and how it behaved, as well as some analysis on what could be changed in future works.

3.1 Dataset Generation and Cleaning

This work used 2 unrelated datasets, the WVS dataset and the COVID dataset. The first step was to merge both datasets into a single one, for that a new column was manually added to the COVID dataset with the country code for each line as in the WVS format specification.

In addition to the existing ones, some extra columns were added to the original COVID dataset before merging. these columns were added to turn the regression problem (predicting the country score) into a classification problem (predicting the country class, good or bad). The added columns were binary, where the value zero represents the bad class and the value 1 represents the good class. The tested approaches were: classify as good if the country score is above the average, classify as good if the score is above an arbitrary number, classify as good if the score is above the median, and lastly, classify as

Figure 3.1 – Data Pipeline



Source: The Author

good if the score is above the median plus the standard deviation of the values.

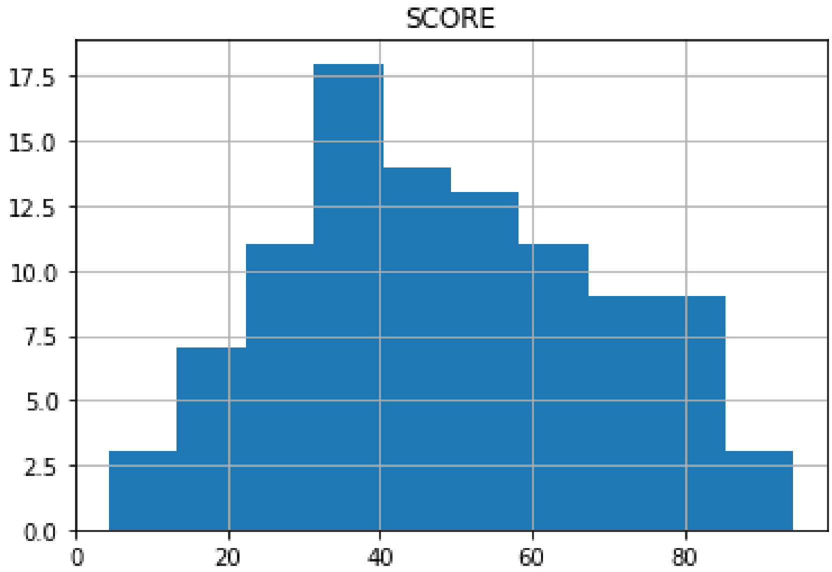
From these 4 columns, only the last one, the median plus the standard deviation, were tested in the final models. Initial tests showed no difference in the behavior of the models. The merged dataset has 52115 lines of class zero, and 13447 of class one. On Figure 3.2 we have a histogram of the scores of the countries.

With both datasets merged, it is necessary to remove some dataset columns, as they would enable the learning algorithms to memorize all classes. Columns such as interview number, country code and the score columns not in use were all removed.

In the merged dataset, there are two types of input data: scalar and categorical. Most of the questions in the survey were structured for the interviewee to answer an integer value from 1 to 5, with some special values for invalid answers that can be checked in the official dataset codebook (INGELHART et al., 2014). For most of the cases, these integer values create a scale, for example, from very important to not important at all, like in question V4. In some other cases, the question options are not in a scale but categories instead, for example in question V57 the interviewee is asked their marital status.

For the scalar type of data, each invalid entry, i.e a negative value entry, was

Figure 3.2 – Country scores histogram



Source: The Author

removed, keeping the cell empty. Next, each column had its values normalized in the range $[0, 1]$ to keep all data in the same range. After this process, the missing values were imputed into the dataset using the average of the values in the same country.

For the categorical type of data, each column was transformed into a set of new columns, using a technique called One-Hot encoding. For each possible answer to the question, a new column is created, then the column corresponding to the answer of the interviewee receives the value 1, and the other the value 0. In this technique, invalid values are automatically handled, as the invalid entry will end with no columns marked as 1.

After all this preparation was done, all columns were formed only by, and all lines that contain any invalid values were removed from the dataset. This work dataset with no invalid values was saved on disk for further usage in the learning phase without the need of executing all steps again. The saved dataset has 12977 lines and 265 columns, while it had 65562 lines before removing lines with any invalid data.

3.2 Machine Learning Pipeline

Based on the work dataset generated in the previous section, some machine learning models were trained, for instance: Logistic regression, Random Forest and Support

Vector Machines.

For each of these models, there were some common steps, the feature selection, and the classifier step itself. For each step there are diverse hyperparameters that can be tuned, the hyperparameters and space of values were chosen based on tests and related literature (GÉRON, 2019).

All the steps for each model were modelled as a pipeline, and this pipeline was given as a parameter altogether to the hyperparameters options to the Grid Search algorithm. An example of the complete Grid Search algorithm configuration can be observed in 3.1 below.

```
GridSearchCV(
    cv=5,
    estimator=Pipeline(steps=[
        (’variance_selector’, VarianceThreshold()),
        (’k_best’, SelectKBest(score_func=mutual_info_regression)),
        (’rf’, RandomForestClassifier())
    ]),
    n_jobs=-1,
    param_grid={
        ’k_best__k’: [16, 32, 64, 128],
        ’rf__criterion’: [’gini’, ’entropy’],
        ’rf__max_depth’: [4, 6, 8],
        ’rf__max_features’: [’sqrt’, ’log2’],
        ’rf__n_estimators’: [10, 50, 100],
        ’variance_selector__threshold’: [0, 0.01, 0.1, 1]
    },
    scoring=make_scorer(f1_score)
)
```

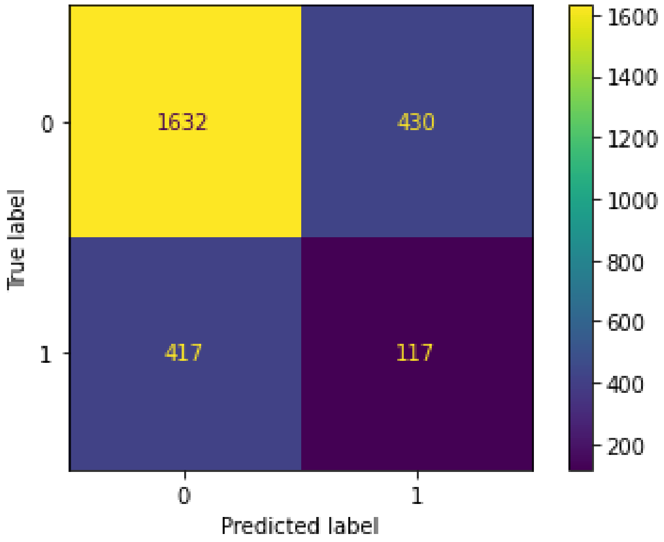
Listing 3.1 – An GridSearch pipeline to tune a RandomForest hyper-parameters

After finding the best hyperparameters with the Grid Search cross-validation, the model is re-trained with the final hyperparameters and all the training data for evaluating its quality.

4 EXPERIMENTS AND RESULTS

From the final trained model, the standard metrics from the confusion matrix were extracted. Other than that, each model was individually analyzed to gather insight about it, as can be seen in the following sections. For determining the success of the model, the results were compared to a random stratified classifier that guesses the class based on the distribution of classes in the training dataset. The random classifier achieved a f1-score of approximately 0.216 and an MCC of approximately 0.010, its confusion matrix can be seen in Figure 4.1.

Figure 4.1 – Confusion Matrix for Dummy Stratified Classifier



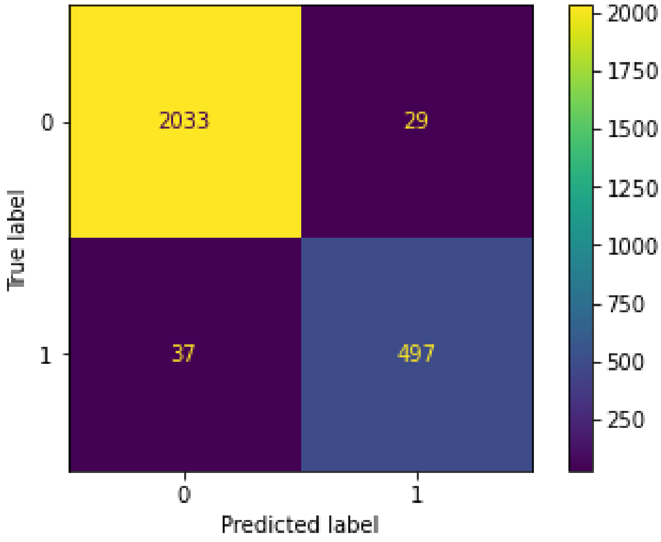
Source: The Author

4.1 Logistic Regression

The first model trained was also the simpler one, logistic regression, but despite the simplicity of the model, the results were quite promising as we can see in the resulting confusion matrix in Figure 4.2.

Given the confusion matrix values, the model achieved a f1-score of approximately 0.938 and an MCC of approximately 0.922, with a good advantage compared to our baseline model. In Figure 4.3 we can see the most important features considered in the model. The most impactful features for the outcome of the model, in absolute values, were:

Figure 4.2 – Confusion Matrix for Logistic Regression



Source: The Author

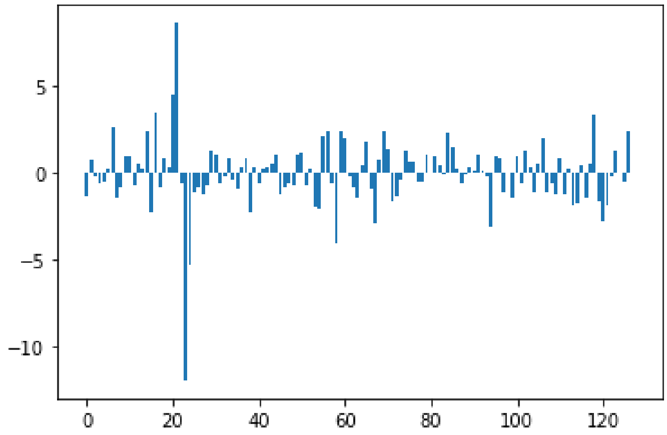
- Feature V27, Active/Inactive membership: Art, music or educational organization, with a coefficient of -11.97197
- Feature V25, Active/Inactive membership: Church or religious organization, with a coefficient of 8.66090
- Feature V28, Active/Inactive membership: Labor Union, with a coefficient of -5.28056
- Feature V24, Most people can be trusted, with a coefficient of 4.47207
- Feature V69, Future changes: Greater respect for authority, with a coefficient of -4.00027

4.2 Random Forest

The next model analyzed was the Random Forest; intuitively this model should work well with the WVS dataset due to the steep nature of the scalar values and the low relation between most of the columns. The intuition is confirmed by looking at the confusion matrix in Figure 4.4.

The Random Forest model performed even better than the Logistic Regression one, achieving a f1-score of approximately 0.995 and an MCC of approximately 0.994, with no false positives and only 5 false negatives. The most important features are inter-

Figure 4.3 – Features importance for Logistic Regression



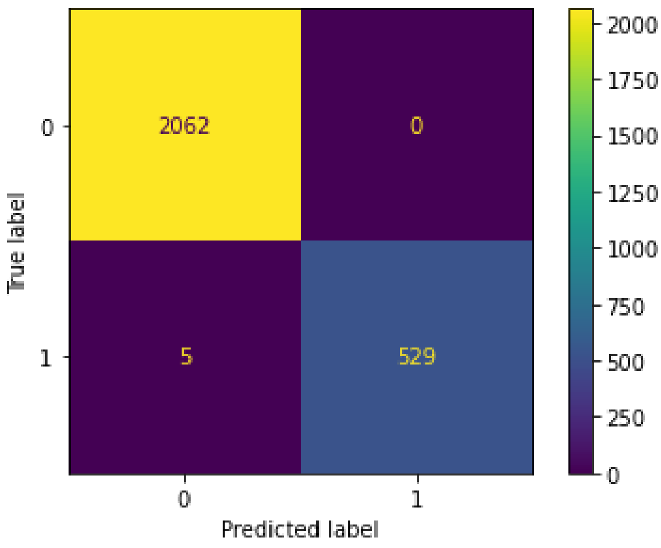
Source: The Author

estingly totally disjoint to the ones in logistic regression:

- Feature V94, Political action recently done: Any other act of protest, with the importance of 0.31
- Feature V93, Political action recently done: Joining strikes, with the importance of 0.2
- Feature V91, Political action recently done: Joining in boycotts, with the importance of 0.17
- Feature V90, Political action recently done: Signing a petition, with the importance of 0.09
- Feature V92, Political action recently done: Attending peaceful demonstrations, with the importance of 0.08

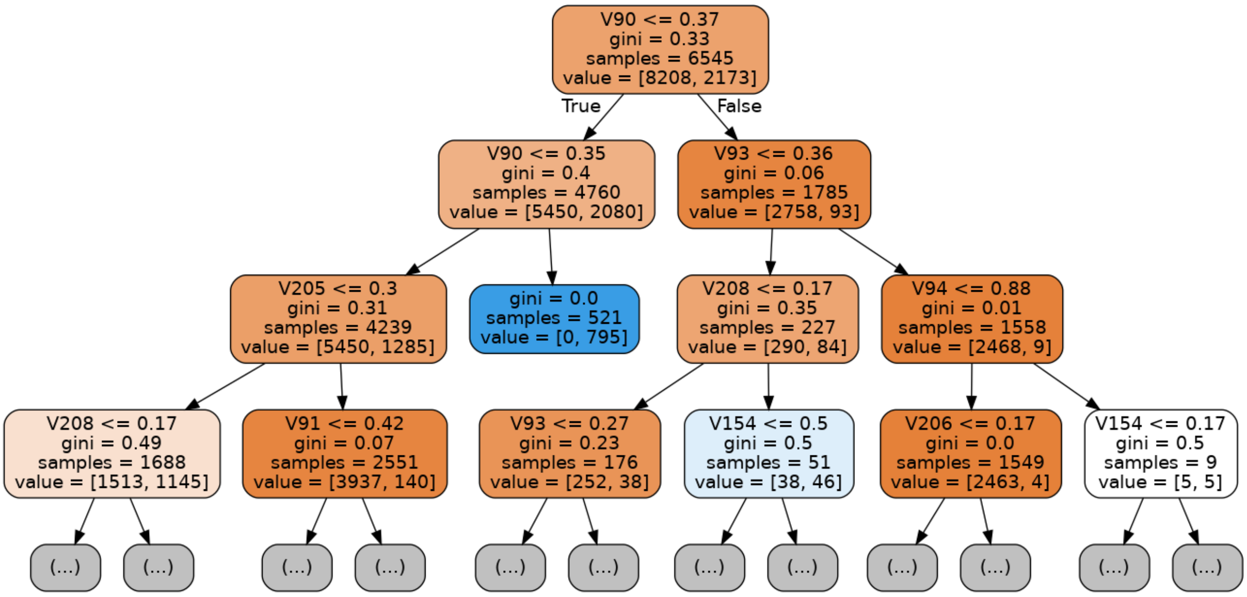
We can observe one of the generated Decision Trees of the Random forest in Figure 4.5 (cropped at depth 3 for legibility).

Figure 4.4 – Confusion Matrix for Random Forest



Source: The Author

Figure 4.5 – Sample Decision Tree from Random Forest



Source: The Author

5 CONCLUSION AND ANALYSIS

This work tested a data science and machine learning approach for using the WVS dataset as a source of information for predicting the performance of countries handling the COVID-19 pandemic. Literature methods were used to implement and validate the developed models.

The results in terms of f1-score and MCC were really high compared to the random classifier, which shows that the models are finding the most interesting features to predict the classes. The performance is not as good when acquiring the score through cross-validation, showing that the models are sensitive to the data being used for training. Even with the bigger variability of the results using cross-validation, the results were always better than the baseline random stratified classifier.

Comparing both trained models it is interesting to notice that both of them seem to have focused on different features and that the features each one focused on are quite similar, membership and activity in groups for the Logistic Regression and political actions for the Random Forest. Both of these groups of question can be interpreted as how politically/community-engaged people of the given country are.

6 FUTURE WORK

During the work new learnings brought insights of things that could have been done differently that could lead to more robust results. This chapter explores some of these possibilities that could be interesting and insightful for the reader.

The first interesting possibility is the use of the MCC metric during the hyperparameter tuning and not only on the model test. as fa-score can be overly optimistic in some cases for unbalanced datasets, it could lead to less optimal hyperparameter choices. Overall the tests showed good results even using the f1-score for hyperparameter tuning.

Consider other models evaluation, as we saw that the models focused on similar groups of questions, it can be interesting to analyze the performance of models that can combine multiple features into new intermediate features such as Neural Networks.

Better model evaluation, other than using the train and test sets, there is additional value in using K-fold cross-validation on the whole dataset to check how sensitive the model is to changes in the training dataset. Based on some evaluation, the standard deviation of the MCC when using cross-validation is high, it is important to investigate the characteristics of the folds that impacted the performance of the model.

Improved hyperparameter search, besides the advantages of simplicity of Grid Search, its cost limits how big is the search space, and the step characteristic of the search makes it easy to miss optimal configuration in between the searched configurations. Genetic algorithms and Random Search are interesting options that could lead to good results, at the expense of some predictability.

During the dataset cleaning some questions needed to be ignored due to their characteristics, for instance, the questions starting with V125 were segmented by country, so given the 17 sub-questions, each interviewee answers only one of them, leading to the other to have an invalid answer. Would be interesting to test different ways to handle these questions, such as merging all sub-questions into a single column.

The WVS dataset is incredibly rich and can be explored in a multitude of ways that were barely touched in this work. Future work could examine the same data utilizing other waves and even multiple waves and the changes between them, providing a temporal dimension to the work.

REFERENCES

ANDERSEN, C. M.; BRO, R. Variable selection in regression-a tutorial. Wiley, v. 24, n. 11-12, p. 728–737, nov. 2010. Available from Internet: <https://doi.org/10.1002/cem.1360>.

BERGSTRA, J.; BENGIO, Y. Random search for hyper-parameter optimization. **Journal of Machine Learning Research**, v. 13, n. 10, p. 281–305, 2012. Available from Internet: <http://jmlr.org/papers/v13/bergstra12a.html>.

CHICCO, D.; JURMAN, G. The advantages of the matthews correlation coefficient (MCC) over f1 score and accuracy in binary classification evaluation. Springer Science and Business Media LLC, v. 21, n. 1, jan. 2020. Available from Internet: <https://doi.org/10.1186/s12864-019-6413-7>.

CLAESEN, M.; MOOR, B. D. **Hyperparameter Search in Machine Learning**. 2015.

DANKERS, F. J. W. M. et al. Prediction modeling methodology. In: . Springer International Publishing, 2018. p. 101–120. Available from Internet: <https://doi.org/10.1007/978-3-319-99713-1_8>.

GÉRON, A. **Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow : concepts, tools, and techniques to build intelligent systems**. Sebastopol, CA: O'Reilly Media, Inc, 2019. ISBN 9781492032649.

HU, B. et al. Characteristics of SARS-CoV-2 and COVID-19. Springer Science and Business Media LLC, v. 19, n. 3, p. 141–154, oct. 2020. Available from Internet: <https://doi.org/10.1038/s41579-020-00459-7>.

INGELHART, R. et al. **World Values Survey Wave 6 (2010-2014)**. World Values Survey Association, 2014. WVS is the largest non–commercial cross p. Available from Internet: <http://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp>.

LENG, A.; LEMAHIEU, H. **Covid Performance Index**. Accessed: 2021-10-30. Available from Internet: <https://interactives.lowyinstitute.org/features/covid-performance/>.

MBAABU, O. 2020. Available from Internet: <https://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp>.

NASCIMENTO, F.; BARONE, D.; CASTRO, H. **Social Activism Analysis: An Application of Machine Learning in the World Values Survey**. 2019.

NEZLEK, J. B. Relationships among belief in god, well-being, and social capital in the 2020 european and world values surveys: Distinguishing interpersonal and ideological prosociality. Springer Science and Business Media LLC, sep. 2021. Available from Internet: <https://doi.org/10.1007/s10943-021-01411-6>.

PENG, C.-Y. J.; LEE, K. L.; INGERSOLL, G. M. An introduction to logistic regression analysis and reporting. Informa UK Limited, v. 96, n. 1, p. 3–14, sep. 2002. Available from Internet: <https://doi.org/10.1080/00220670209598786>.

VIER, T. **O uso da inteligência artificial nas ciências sociais : o caso do patriotismo dos brasileiros**. Thesis (PhD) — Universidade Federal do Rio Grande do Sul, Instituto de Filosofia e Ciências Humanas, Curso de Pós-Graduação em Ciência Política, Porto Alegre, 2020.

WALSH, C. G.; RIBEIRO, J. D.; FRANKLIN, J. C. Predicting risk of suicide attempts over time through machine learning. SAGE Publications, v. 5, n. 3, p. 457–469, abr. 2017. Available from Internet: <https://doi.org/10.1177/2167702617691560>.

WELZEL, C. Why the future is democratic. Project Muse, v. 32, n. 2, p. 132–144, 2021. Available from Internet: <https://doi.org/10.1353/jod.2021.0024>.