

École polytechnique de Louvain

Predicting armed conflicts: a machine learning approach

Author: **Corentin HAVELANGE**
Supervisors: **Raphaël JUNGERS, Valérie ROSOUX**
Reader: **Jean-Charles DELVENNE**
Academic year 2020–2021
Master [120] in Mathematical Engineering

Abstract

Political conflicts may not be the first cause of death worldwide, they constitute nevertheless a major source of destruction and prevent many countries from developing themselves and providing stability, health, and security to their citizens. Assessing the risk of conflicts in every country with evidence-based predictions is now a duty for the governments and international organisations. In this thesis, we address the problem of conflict prediction through a machine learning approach. We aim at predicting whether there will be a conflict which corresponds to a quarrel with at least 25 casualties within the next four years, either at national or subnational scale. The prediction is based on social, demographic, economic, geographic and political national indicators measured on a yearly basis. First, we determine which variables are significant using analysis of variance. Then, using mutual information, we measure the strength of the relationship between pairs of input variables and between input variables and the outcome. Using mutual information in the context of conflict prediction is a contribution of this work. After this, we use logistic regression to classify the instances in two categories : future peace or future conflict. Then, with a random forest model, we obtain better performance than previously reported for the dataset that we use. In the last section, we address the problem of the interpretability of the results using SHAP values, a method able to generate an explanation model such that it relates the predicted probability of conflict and the input variables by a linear function. We show that the lack of democracy is not a determining variable for our algorithms whereas it is frequently used to justify international interventions by foreign countries. In addition, our observations suggest that the best predictors for future onset are the current intensity of conflict, the amount of structural constraints and the government's effectiveness when the fight is for national power. When the conflict is at subnational level, the best predictors are the current intensity of conflict, the population size and the level of repression.

Acknowledgements

Firstly, I want to thank my two thesis supervisors for accompanying me through this work. I am very grateful to Professor Raphaël Jungers for his enthusiasm for this unconventional topic, his time and his technical expertise and help. I also want to express my sincere gratitude to Professor Valérie Rosoux for accepting to supervise my project, for her time, her expertise in the field, her kindness and encouragements, and her valuable advice.

I would like to thank Professor Jean-Charles Delvenne for accepting being my reader.

A special thanks to Benjamin Chiêm for his time, his guidance, technical help, many suggestions and proofreading of my thesis. Another thanks to Brieuc Pinon for the technical advices and proofreading.

I would like to sincerely thank my family for their support and especially my sister Andine for the proofreading.

Table of Contents

1	Problem definition and historical overview	5
1.1	Historical overview	5
1.2	Informal description of the problem	8
1.3	Mathematical description	9
1.4	Dataset description	11
1.4.1	Variables description	14
1.4.2	Data discussion	16
2	ANOVA and mutual information	21
2.1	Variables significance: analysis of variance	21
2.1.1	Analysis of variance: theory	22
	ANOVA suitability for this situation	25
2.1.2	Results of ANOVA and assumptions test	27
2.2	Mutual information	29
2.2.1	Mutual information: theory	30
2.2.2	How to apply mutual information to our problem	32
2.2.3	Pairwise mutual information	33
2.2.4	Mutual information between input variables and outcome	34
2.3	ANOVA and mutual information: comparison and discussion	38
2.4	Conclusion	38
3	Classification approach	39
3.1	Performance metrics: how to assess $\hat{f}$ and $\hat{p}$	40
3.1.1	The confusion matrix and basic metrics	40
3.1.2	The ROC AUC	42
3.1.3	The precision-recall curve AUC	44
3.1.4	The Brier score	45

3.1.5	Cohen’s Kappa	46
3.1.6	Final choice of metrics	47
3.2	Methodology	47
3.2.1	Variables consideration	47
3.2.2	Procedure	48
3.3	Logistic regression	51
3.3.1	Logistic regression: theory	52
3.3.2	Strength and weaknesses of logistic regression	57
3.3.3	Analysis of the appropriateness of logistic regression for our analysis	57
3.3.4	Results obtained with logistic regression	58
	Experiment 1: insignificant variables dropped and ℓ_2 –regularisation	58
	Experiment 2: insignificant variables dropped, no regularisation	60
	Experiment 3: all variables, ℓ_1 –regularisation	60
	Experiment 4: all variables, no regularisation	60
	Overall comparison of the models and conclusions	63
3.3.5	Coefficients of the logistic regression	64
3.4	Random forest	67
3.4.1	Random forest: theory	67
	Decision tree	67
	From decision tree to random forest	71
	Strengths and weaknesses of random forest algorithm	71
	Suitability of random forest for the conflict prediction problem	71
	Hyperparameters tuning	72
3.4.2	Results obtained with random forest	74
3.5	Comparison of the models and conclusion	75
4	SHAP or how to make the black box white	79
4.1	Method: SHAP values	81
4.1.1	Coalitional game and Shapley values	81
4.1.2	SHAP or how to determine features importance using Shapley Value	83
4.1.3	SHAP : global interpretation	88
4.1.4	SHAP interaction values	89
4.1.5	SHAP: strenghts and weaknesses	89
4.1.6	Conclusion : is SHAP appropriate for our problem ?	90
4.2	Results	90

4.3	SHAP to improve model performances	98
4.4	SHAP to understand model failures	99
4.5	Chapter Conclusion	99
5	Final discussion and future work	101
5.1	General discussion and limitations	101
5.2	Future work	104
5.3	Conclusion	105
A	Variables origins and details	108
B	Algorithms of Chapter 2	111
C	Additional figures for Chapter 4	113

List of Figures

1.1	Simplified tree for interstate conflicts resolution found in [24]. At each node of the three, there are several branches to follow. Some nodes are related to internal characteristics of the conflicts, others concern the type of resolution that is undertaken. At each leave, one finds the probability of success based on past conflicts. The goal is to determine which type of mediation attempt has the highest chance to be successful.	7
1.2	World map with countries considered in the GCRI dataset highlighted in blue. As explained previously, the European Union countries are not part of it. Furthermore, we see that the other absent countries are very small states. . .	17
2.1	The five most significant variables according to ANOVA and their p -values on a log-scale. The exact interpretation of the p -value is the following : it is the probability to observe the sample $x_{.j}$ given that H_0 is verified, meaning that the mean of the values for positive instance is the same as the mean of the values for the negative instance for the variable given on the y -axis. Let us recall that, the smaller the p -value, the most significant the variable is. . . .	27
2.2	Schematic representation of mutual information and entropy in form of a Venn diagram extracted from [90]	32
2.3	Mutual information matrix between the input variables. Entry k, l is equal to $I(x_{.k}, x_{.l})$ or intuitively to the amount of information about variable k that we can retrieve by knowing variable l (or conversely since mutual information is symmetric).	35
2.4	The five variables with highest mutual information with $\mathbf{y}$. On the ordinate, the name of the considered variable and on the abscissa, the mutual information $I(x_{.k}, \mathbf{y})$. The values are ordered following NP scores. It is not more important than SN, but one order had to be chosen.	36

3.1	Illustration of ROC which is the trade-off between the TPR, the proportion of identified relevant instances and the FPR, the proportion of irrelevant instances that are classified as positives. In the Figure on the left, we observe that it is possible to achieve a higher TPR for a same FPR than in the Figure on the right. In the Figure on the right, it is the score obtained by tossing a coin for determining if a phenomenon will happen (if the phenomenon has a probability of 0.5 to happen).	43
3.2	Illustrations of precision recall curves. These represent the trade-off between sensitivity, the proportion of relevant instances, and the specificity, the identified instances that are relevant. In the Figure on the left, we observe that it is possible, for the same sensitivity, to achieve a higher specificity than for the classifier depicted in the Figure on the left. Based on this criterion, the model for which the performance is given in the Figure on the left is a better classifier.	44
3.3	5-fold test scheme with each line representing an independent run (coming from [59]). Each row represents a train and test phase for the model such that there are 5 different models. The blue boxes correspond to the part of samples used for the training phase and the red boxes correspond to the part of the samples used for the test phase.	49
3.4	Illustration of a one-dimensional logistic regression (the green and red dots are positive and negative instances respectively). On the ordinate, we have the assessed probability. On the abscissa, we have the feature value. An instance is classified as positive if the assessed probability is greater or equal to the threshold t which is represented by the dashed line. Otherwise, it is classified as negative.	53
3.5	Median precision-recall curves of the logistic regression classifier, NP case. The specificity is given on the ordinate-axis and the sensitivity on the abscissa. The red curve shows the median performance on the test sets and the dashed black curve shows the median performance for the train sets. The shaded area corresponds to the interquartile range of the precision-recall curves on the test sets.	61

3.6	Precision-recall curves of the logistic regression classifier, SN case. The specificity is given on the ordinate-axis and the sensitivity on the abscissa. The red curve shows the median performance on the test sets and the dashed black curve shows the median performance for the train sets. The shedded area corresponds to the interquartile range of the precision-recall curves on the test sets.	62
3.7	The five biggest mean coefficients of logistic regression for experiments 3 and 4 in absolute value. The black lines correspond to their standard deviations. The variables' names are given on the y -axis and the value of their coefficient on the x -axis.	66
3.8	Theoretical scheme of a decision tree. At each node, there is a condition of the form $x_{ij} < \theta$. If this condition is met, we take the left branch, otherwise, we take the right branch. The leaves correspond to the final answer, either a class index (0 or 1) or a probability of belonging to $\mathcal{C}_1$	69
3.9	Conceptual example : decision tree	70
3.10	PR curves obtained with random forest classifier. The red line corresponds to the median of the test curves and the grey zone to their interquartile ranges. The dashed black line corresponds to the median train curve.	75
3.11	Comparison between the median PR curves for test sets obtained by Halkia et al. in [35] with logistic regression, by our implementation of logistic regression and by our implementation of random forest algorithm. The shedded area corresponds to the interquartile range. LR stands for logistic regression and RF for random forest.	77
4.1	One of the trees of the conceptual example	80
4.2	Decision plot for Angola in 2002, NP. On the left, the feature names are displayed with the SHAP values in brackets. The line starts from the expected value for the classifier which is 0.183. It then follows a trajectory according to the sum of the SHAP values displayed on the y -axis. The features' values (not rescaled) are given in brackets next to the line.	92
4.3	Summary plot for all the dataset instances (NP). It depicts the global explanation model. For each line, a dot corresponds to an instance and the colour to the feature value. The dot abscissa corresponds to the SHAP value for the feature displayed on the ordinate.	93

4.4	Summary plot for all the dataset instances (SN). For each line, a dot corresponds to an instance and the colour to the feature value. The dot abscissa corresponds to the SHAP value for the feature displayed on the ordinate. . .	94
4.5	Dependence plots between SHAP and features values for the five most important variables (NP). It is a representation of the pairs $(x_{ij}, \phi_j^{(i)})$	95
4.6	Dependence plots between SHAP and features values for the five most important variables (SN). It is a representation of the pairs $(x_{ij}, \phi_j^{(i)})$	96
4.7	Interaction plots for GDP Per Capita* and Max Value: NP and SN. The dots represent the pairs $(x_{ij}, \phi_j^{(i)})$ for $1 \leq i \leq n$. On the abscissa, we have the GDP Per Capita* value and we have the SHAP value on the ordinate. The colour of the dots is related to the value of the current intensity of conflict for each instance. This figure allows us to understand interaction effects between these two variables.	98
C.1	Summary plot for NP test set. For each line, a dot corresponds to an instance and the colour to the feature value. The dot abscissa corresponds to the SHAP value for the feature displayed on the ordinate.	114
C.2	Decision plot for misclassified instances (NP). The line starts from the expected value for the classifier which is 0.183. It then follows a trajectory according to the sum of the SHAP values displayed on the y -axis.	115
C.3	Summary plot for SN test set. For each line, a dot corresponds to an instance and the colour to the feature value. The dot abscissa corresponds to the SHAP value for the feature displayed on the ordinate.	116
C.4	Decision plot for misclassified instances (SN). The line starts from the expected value for the classifier which is 0.28. It then follows a trajectory according to the sum of the SHAP values displayed on the y -axis.	117

List of Tables

1.1	The two types of classification. National power fight (NP) opposes a non-state actor against the government, while subnational level quarrel (SN) opposes two non-state organisations.	9
1.2	Prevalence in the dataset. The prevalence is the proportion of positive instances which are the conflict situations. We observe that the prevalence is higher for subnational quarrels than for national power fights.	12
1.3	The variables categories. The variables are divided into 5 key categories and 10 concepts.	13
1.4	Transnational ethnic bonds scores (adapted from [34])	14
1.5	Regime type scores (taken from [14]). As explained, the scores are related to the instability resulting from the indicated regime type and that is how they are ranked between 0 and 10.	16
1.6	Descriptive statistics of GCRI dataset variables. It is also specified whether the variable is discrete or continuous. There are no missing values as can be observed in the first column.	19
2.1	Toy dataset for conceptual example with 5 country-years and only two variables	21
2.2	p -values obtained for the Shapiro-Wilk, Brown-Forsythe and ANOVA tests for each variable (group 0 being the negative instances and group 1 the positive ones). The p -values highlighted in bold are the one bigger than 0.01. For these, we cannot reject H_0 at 0.01 significance level.	28
2.3	Mutual information between $\mathbf{y}$ and x_j for each j . It is ordered according to the NP scores. It is not more important than the SN scores, but we had to choose one of the two to increase readability.	37
3.1	Toy dataset for conceptual example	39
3.2	Confusion matrix	40
3.3	Abbreviations used for the classifier results	41

3.4	Assessment of the model with ROC AUC score by Hosmer & Lemeshow [41]	42
3.5	Toy dataset and predicted probability $\hat{p}$ for conceptual example	45
3.6	One interpretation of Cohen's κ by Koch & Landis [52]	47
3.7	Median performance metrics for the logistic regression classifier for each experiment (the numbers in brackets are the interquartile ranges except for the confusion matrix where it corresponds to the percentages). The best scores are highlighted in bold writing.	63
3.8	Mean coefficients of the logistic regression obtained with ℓ_1 -regularisation or no regularisation, with all variables kept both for NP and SN. (the numbers in the brackets correspond to the standard deviations)	68
3.9	Toy dataset for conceptual example	69
3.10	Random forest hyperparameters choice	74
3.11	Median results for random forest classification (the numbers in brackets are the interquartile ranges except for the confusion matrices where it corresponds to the percentages).	76
4.1	Toy dataset and predicted probability $\hat{p}$ for our conceptual example	80
4.2	Toy dataset and predicted probability $\hat{p}$ for our conceptual example	84
A.1	Data origin of each variable	109
A.2	Years of data for countries in the dataset	110

Abbreviations list

GCRI	Global Conflict Risk Index
NP	National Power (conflict)
SN	Subnational (conflict)
GDP	Gross Domestic Product
ML	Machine Learning
ANOVA	ANalysis Of VAriance
BF	Brown-Forsythe (test)
MI	Mutual Information
TP	True Positive
FP	False Positive
FN	False Negative
TN	True Negative
TPR	True Positive Rate
FPR	False Positive Rate
PPV	Positive Predictive Value
ROC (curve)	Receiver Operator Characteristic (curve)
PR (curve)	Precision Recall (curve)
BS	Brier Score
LR	Logistic Regression
RF	Random Forest
GNN	Graph Neural Network

Introduction

Political conflicts may not be the first cause of death worldwide, they constitute nevertheless a major source of destruction and prevent many countries from developing themselves and providing stability, health, and security to their citizens. According to Strand et al., armed conflict has still cost 53 000 lives in 2018 [81]. That is only the top of the iceberg since instabilities cause many supplementary deaths due to the decrease in public health and infrastructure. The end of armed clashes does not mean that the conflict is over. Armed forces must be demobilised as well as population minds who carry grievances and traumas. These who want to initiate a reconciliation process need to consider the transformation of relation between former opponents over several generations [71]. Taking care of armed communities takes tens of years and constitutes a substantial amount of work since it has to be done individually [72]. Furthermore, the cost of wars in terms of military investment, destructed infrastructure and additional health service needed is considerable. All the budget that is devoted to it does not go into social security, better healthcare or scientific research. Besides, these consequences are not limited to the states where the conflicts take place. It impacts citizens who left their country at war in order to live in diasporas all over the world, though in a different measure.

Assessing the risk of conflicts in every country is now a duty for the governments and international organisations. From there, it becomes possible to build an efficient foreign policy in the hope that it prevents most of the conflicts to happen. However, predicting peace or unrest is challenging. Numerous cognitive biases prevent humans to forecast effectively the imminence of conflict since it generates very diverse emotions depending on the living experience of individuals. It is often driven by past trauma, passions or interest in international intervention [67]. Governments and international organisations need reliable evidence-based conflict risk assessments. Because of the complexity of conflicts, it is hard for a single human to process all the elements that come into play.

In this thesis, we propose to address the problem of conflict prediction through a machine learning approach. Machine learning is a sub-field of artificial intelligence inspired by human learning. More particularly, we apply supervised learning in this thesis. An algorithm needs to output the right outcome based on an input, the variables. When the outcome is continuous, we talk about regression, when it is ordinal, as in our case, we talk about classification. The algorithm learns how to produce the right output by being trained with many examples

and adapting its parameters depending on the errors and success. In this thesis, we will often refer to the concept as machine learning which includes supervised learning. What we aim at is to train an algorithm to recognise a situation of upcoming conflict based on yearly national indicators. More precisely, we want to classify the instances in two categories : future peace and future conflict. The input indicators are geographic, economic, political, demographic and social. We believe that the machine learning approach can be part of an early warning system. It would output probabilities of future conflicts and highlight high risks. Specialists of the field could then investigate the worrisome cases since the algorithm results need human check. We believe that, in such complex phenomena, the advantage of the computer is that it can grasp complex patterns that a human brain cannot process. The machine may not understand the complexity of hate, social anger or revenge, but it is not driven by past trauma or interest in intervening somewhere.

We suppose that we could face some difficulties because of the interdisciplinary nature of this work. In a world where everybody becomes more and more specialised in his or her specific domain due to the increasing complexity of all things, it is enlightening to work with experts coming from different fields.

This thesis is structured in the following way. In Chapter 1, we provide an historical overview, we define the problem, both informally and mathematically. Then we take a closer look at the dataset and the variables.

In Chapter 2, we use the ANOVA method to determine which input variables are significant with respect to the outcome. Then we analyse the link between the input variables but also the link between them and the outcome based on mutual information, a concept coming from information theory. This has not been applied yet to the conflict prevention field. At last, we compare the results of both methods.

In Chapter 3, we realise the classification task to determine the outcome - conflict or peace - based on the input variables defined in Chapter 2. We begin with logistic regression, a machine learning technique frequently used in the conflict prevention field and as a starting point in M.L in general. Secondly, we use random forest models. Thirdly, we compare our results for both methods and the performances obtained by the GCRI team.

In Chapter 4, we use SHAP, an interpretable machine learning tool based on game theory, more precisely on Shapley values. This allows us to analyse the input variable's importance in the predictions coming from random forest model, both globally and locally. This is our contribution since it has not been applied yet in the conflict prevention field. This allows us to transform black box models into white box models and still benefit from better perform-

ances

In Chapter 5, we discuss and compare the main results from Chapters 2, 3 and 4. After this, we draw the limitations of our work, provide ideas for future work and a general conclusion.

Chapter 1

Problem definition and historical overview

In this chapter, we provide a historical overview of conflict prediction and prevention. After this, we define, both informally and mathematically, the problem that we want to address in this thesis. Then, we present the dataset and the variables.

1.1 Historical overview

Machine learning has been applied for political conflicts prediction in previous reports. It was already a field of interest in the 60' [39]. As explained above, we apply supervised learning which is a sub-field of machine learning where algorithms predict an output based on input variables. The main principle is the learning : the algorithm is fed with training data on which it can "learn" how to make good predictions. One can determine how we apply ML to political conflicts according to two axes. The first one is the action taken about the conflict: either trying to forecast it or finding the best way to solve it. The second axis is the type of conflict taken into consideration. It begins with very local violence, at town scale, and it can evolve to regional disputes. After this, one can consider civil wars and, even further, interstate conflicts. The scale is continuous and one can think about any level of resolution. A third axis that we could imagine is the type of machine learning models that are applied. We shall cover here the combinations that we encountered in the literature.

There have already been attempts to address the problem of interstate conflicts. This has been tried mainly from 1990 to 2010 by many scientists, among which Robert Trappl, head of the Austrian Research Institute for Artificial Intelligence. This institution was a major contributor to the field, as was the Political Science department from the university of Heidelberg and Canterbury in New Zealand with Jacob Berkovitch [87]. Trappl et al. wanted to predict whether a particular attempt of resolution of a conflict would fail. They fed the decision trees algorithm with national indicators, but also variables directly linked to the conflict. A simplified example is given in Figure 1.1. However, they obtained poor results with this technique. Only 59.1% of the cases were correctly predicted. The difference with

the default value of always predicting failure, which is 55.3% is only 3.8 % [86].

Others tried to forecast interstate conflicts. In 2013, Gleditsch & Ward proposed a model based on regression. The variables are among others whether there have already been militarised disputes between both countries, whether there is a river, territorial or maritime claims, whether there have been attempts of peaceful settlement or the democracy score [30]. However, the number of omitted disputes made the model barely usable. Several others have tried but with limited success [9]. Gleditsch said in 2017 that "Despite this promise, however, prediction remains highly controversial in academic conflict research. Relatively few conflict experts have attempted explicit forecasting of conflicts. Furthermore, no system of early warning has established itself as a reliable tool for policy-making, although major efforts are currently under way." [9]. It seems now, probably due to the decrease of interstate quarrels, that the interest of the scientific community has decreased.

At other resolution levels, there are promising projects. The ViEWS project predicts quarrels both at national and regional scales for the African continent accurately [38]. It assesses probabilities for month $t + 1$ until $t + 36$ by averaging the output of several models. This project is supervised by the University of Uppsala which is an important actor in the peace-building community. Still at national scale, the Global Risk Conflict Index has been developed based on logistic regression as the majority of the actual models [35]. It is fed with national numerical indicators and outputs a probability of conflict. At regional scale, Philip Schrodtt developed an early warning system based on the automated analysis of English-language news [87].

Very local violence can also be predicted with variable accuracy. One of the problems is the availability of very local and precisely collected data. In many unstable countries, this is impossible nowadays [6]. The methods encountered in this field are often the same as for national prediction [6]. The results obtained were disappointing. The place is often correctly guessed, but the time of arising is seldom predicted with accuracy. The implication of non-easily measurable variables like social grievance or deterioration of communal thrust in the formation of quarrels makes them difficult to forecast [6].

Next to the logistic regression models, some researchers use neural networks [5]. Others have implemented random forest models [65]. However, interpretable models are preferred to black box ones at the expense of performance [35]. Indeed, policymakers do not like to use a single output without any explanations.

The development of machine learning models is possible thanks to many conflict databases. We present the major ones but others exist. The scope of the database presentation is not exhaustive.

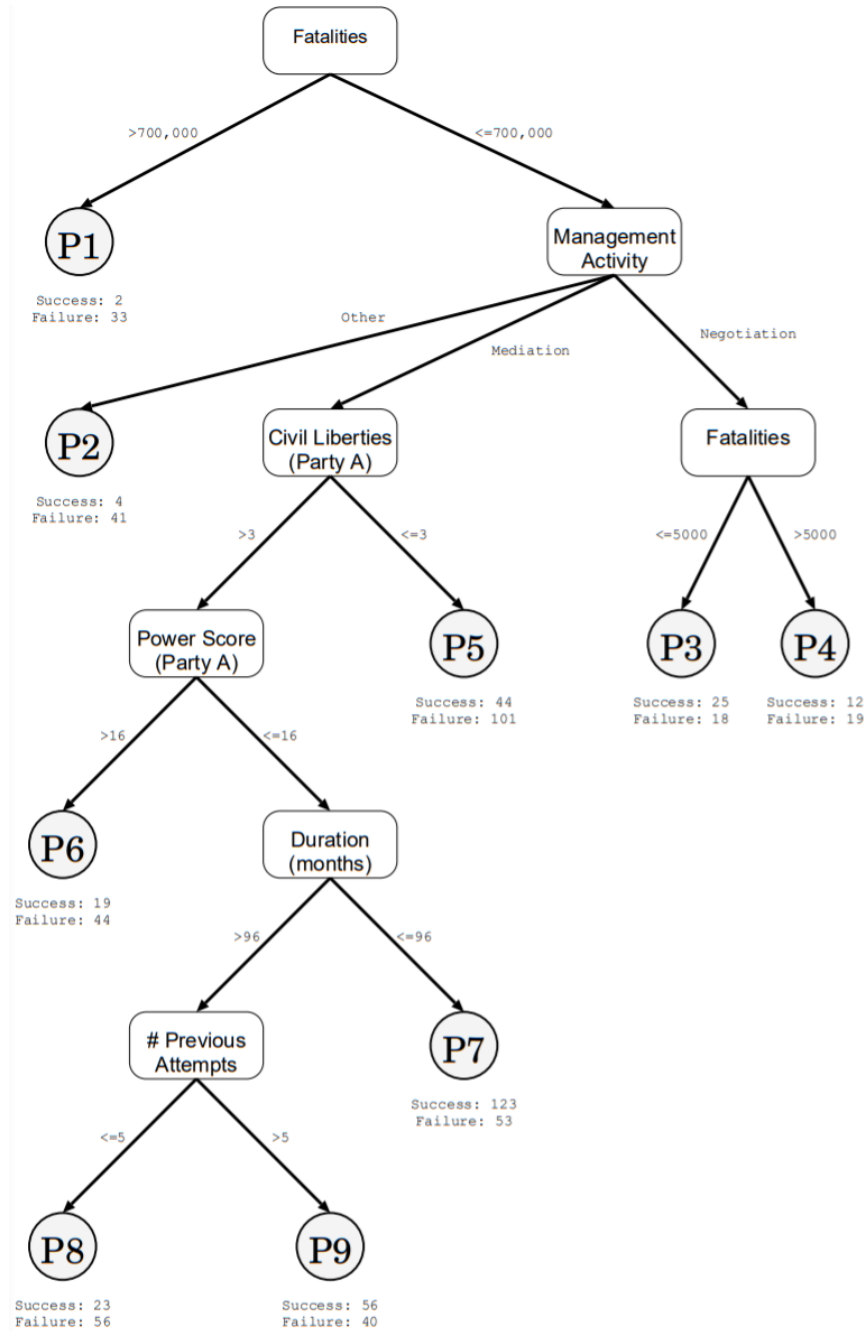


Figure 1.1: Simplified tree for interstate conflicts resolution found in [24]. At each node of the tree, there are several branches to follow. Some nodes are related to internal characteristics of the conflicts, others concern the type of resolution that is undertaken. At each leaf, one finds the probability of success based on past conflicts. The goal is to determine which type of mediation attempt has the highest chance to be successful.

- UCDP (for Uppsala Conflict Data Program): many kinds of datasets are available. It lists various conflicts from very local ones to civil wars [84]. It seems that this is one of the most updated and exhaustive data collections. It is used in the GCRI project and of course for the ViEWS project [89].
- CONFMAN (for conflict management): a database focused on conflict mediation attempts, developed by the team of Jacob Berkovitch from the University of Canterbury, New Zealand. It considers 241 conflicts and about 920 mediation attempts and gives in addition several dispute attributes. This data has been widely used [87].
- PRIO (for Peace Research Institute Oslo) disposes of several datasets, many of which are based on spatial data [82]. It is directed by Henrik Urdal.

Governments and international organisations are also interested in building such models for efficient policymaking based on early warning systems [35]. Among them, we count national agencies but also international organisations such as the World Bank. Generally, their model is not publicly released. There is little cooperation with the academic world. The GCRI is an exception since it originates from the European Union, but is developed with the help of many scientists [35].

1.2 Informal description of the problem

As explained previously, conflict and peace are complex and intricate phenomena. In order to shape the policy of governments and international organisations, it is necessary to have a vision about the future peace or unrest in the rest of the world. It is even more than necessary, it is crucial. Omitting a conflict or taking the decision of intervening or not can change the life of hundreds of thousands of people for several generations.

We start from the Global Conflict Risk Index (GCRI). It is a project supported by the European Union. It aims at predicting whether a conflict will outbreak in the next four years in any state outside the EU using machine learning. This is done with political, geographical, economic, social, and demographic data at national scale. As a consequence, this tool is not developed to predict the precise time and place of conflicts. The main goal is to provide probabilities for experts and political instances (e.g. the European Union) for their long-term foreign policy. Other predictors working more locally and on a smaller timescale will be more helpful to dispatch international aid or food stocks. We will not cover these short-term and local prediction model. Our goal is to test the limits of the GCRI tool and develop it further.

We will compare different machine learning algorithms based on rigorous criteria and discuss which one is the most adapted. We start with standard classification methods for two distinct tasks (see Table 1.1). National power fight (NP) opposes a non-state actor against the government, while subnational level quarrel (SN) opposes two non-state organisations.

	Violent conflict
National power (NP)	Yes/No
Subnational level (SN)	Yes/No

Table 1.1: The two types of classification. National power fight (NP) opposes a non-state actor against the government, while subnational level quarrel (SN) opposes two non-state organisations.

We deliberately chose to consider two classification tasks. Although NP and SN unrest are sometimes related, there is no absolute relation between them. The absence/presence of one does not allow us to draw conclusions about the absence/presence of the other.

The threshold to determine that a situation is a conflict is based on the number of civil and military casualties that must be greater or equal to 25. This criterion is based on the absolute number of casualties and not on a relative number, for example related to the total population. Otherwise, populous countries would have a lower score for the same magnitude of conflict.

International organisations tend to prefer interpretable models to black boxes. There is a need to motivate intervention decision with not only an algorithm probability output but also with concrete elements. But this need for interpretability results in a decrease for performance until here. We want to apply state-of-the-art techniques in order to make high-performing but black box algorithms explainable. More particularly, we want to highlight the most important variables in the prediction and possibly explain how the final output has been computed by the model.

1.3 Mathematical description

We can define the problem as follows: a sample point x_i belongs either to class $\mathcal{C}_0$ or to class $\mathcal{C}_1$. y_i is the index of the class x_i belongs to. In all the following analyses there are two classes (binary classification). Classification opposes to regression where one seeks to predict a real number, depending on the input. In this chapter, we use the the following notations (adapted from [45]):

- n , the number of data points
- p , the number of variables (also called features, we will use both terms equivalently)
- K , the number of classes (in this case, two)
- x_{ij} , the j^{th} variable value for the i^{th} observation with $1 \leq i \leq n$ and $1 \leq j \leq p$
- $x_i = (x_{i1} \ x_{i2} \ \dots \ x_{ip})$

- $x_{\cdot j} = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}$
- $\mathbf{X} = [x_{ij}]_{1 \leq i \leq n, 1 \leq j \leq p}$
- y_i is the class indicator for data point x_i with $1 \leq i \leq n$. It takes the following values:
$$y_i = \begin{cases} 0 & \text{if } x_i \in \mathcal{C}_0 \\ 1 & \text{if } x_i \in \mathcal{C}_1 \end{cases} \quad (1.1)$$

It is the value we want to predict for x_i . If y_i is equal to 0, it is called a negative instance. Otherwise, it is a positive instance.

- $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$

We assume that there is a function $f : \mathbb{R}^p \rightarrow \{0, 1\} : x_i \rightarrow f(x_i)$ such that

$$y_i = f(x_i) + \epsilon \quad (1.2)$$

with $\epsilon \in \mathbb{R}$ being the random error term, assumed to be independent of x_i and with zero-mean. This hypothesis will be discussed in section 2.1.1. This function is unknown, and the goal is to find $\hat{f} : \mathbb{R}^p \rightarrow \{0, 1\} : x_i \rightarrow \hat{f}(x_i)$ such that

$$\hat{y}_i = \hat{f}(x_i) \quad (1.3)$$

provides the best estimate for $\mathbf{y}$ based on several criteria that we will introduce in section 3.1. Regarding the conflict problem, x_i is the vector of the indicators for one country on one determined year. y_i represents whether there will be a conflict outbreak in the next four years ($y_i = 1$) or not ($y_i = 0$).

Most of the time, we rather consider a probability of belonging to class $\mathcal{C}_1$:

$$\hat{p} : \mathbb{R}^p \rightarrow [0, 1] : x_i \rightarrow \hat{p}(x_i) \quad (1.4)$$

Since there are only two classes, one can deduce the probability of belonging to $\mathcal{C}_0$ by taking the probability of not belonging to $\mathcal{C}_1$ which is :

$$\mathbb{P}[x_i \in \mathcal{C}_0] = 1 - \hat{p}(x_i) \quad (1.5)$$

We can compute $\hat{f}$ based on $\hat{p}$ by fixing a discriminating threshold t :

$$\hat{f}(x_i) = \begin{cases} 0 & \text{if } \hat{p}(x_i) < t \\ 1 & \text{if } \hat{p}(x_i) \geq t \end{cases} \quad (1.6)$$

Most of the time, this threshold is equal to 0.5. It can possibly be decreased if we want the classifier to be more sensitive. It will typically result in predicting more often that one sample belongs to class $\mathcal{C}_1$ even if it is not, while classifying less often positive samples as negative ones. Threshold t can also be raised, and it results in the opposite phenomenon.

Additionally, we suppose that not all the variables have the same importance to determine the output. Some of them might contain more information and maybe some others are irrelevant. Therefore, we want to test for statistical significance.

Furthermore, the final users want to understand why the model has produced a determined output for a specific instance. For a model with clear concise expression like a linear function with few variables, it is not a problem. However, if $\hat{f}$ is too cumbersome and has no concise clear expression like e.g. for a deep neural network, it is rather complicated to understand how the model works and to explain a prediction. We cannot know how each feature value has contributed to the final result if $\hat{f}$ takes several pages to be written down. That is why we want to derive an explanation model. This is a second model g , derived from $\hat{f}$ that is interpretable. This implies that it can be expressed concisely and provides a clear relation between the prediction and each feature. It can be local or global. This means specific to each instance or not. Of course, it must be as close as possible to $\hat{f}$.

1.4 Dataset description

The dataset used in this thesis also comes from the Global Conflict Risk Index project.

It includes 39493 country-years from 1989 until 2014. A country-year corresponds to the annual measured indicators for one determined year in a particular country. Since we aim at predicting conflict outbreaks in the coming 4 years, the country-years from 2011 until 2014 have not been used during the learning phase, which brings the number of usable country-years n down to 2940. Country-years between 2011 and 2014 are useful to determine whether 2010 is a positive instance or not. Some countries do not cover the full range (see Table A.2 in Appendix A).

Now, the GCRI monitors 191 countries [35], but we only had access to a former dataset from 2017 which limits us to 138. These 138 countries are shown in Figure 1.2. Since this is a European Union project for its foreign policy, no EU-state is in the dataset. It seems surprising that, although they still face risks of conflict, Cyprus and Ireland are not part of this dataset, even if they are in the EU [2, 53]. The other missing states are mainly very small countries. We decided not to adapt the dataset by including them because of the amount of

	Violent conflict
National power	592 (20.14%)
Subnational level	894 (30.41%)

Table 1.2: Prevalence in the dataset. The prevalence is the proportion of positive instances which are the conflict situations. We observe that the prevalence is higher for subnational quarrels than for national power fights.

work of data collection, cleaning, imputation and adaptations.

The prevalence of conflict is given in Table 1.2. As can be noted from that table, the dataset is imbalanced, which constitutes a challenge for the machine learning algorithms.

Correlation between the variables and conflict outbreaks has been strongly established in [34], [35], [36] and more generally in the literature around conflict prediction and prevention. One of the main advantages of these indicators is that they tend to remain stable and change little from year to year [34]. An important variable in conflict prediction are past quarrels. Indicators describing these more clearly should probably add precision to the classifier. For example, we could consider the type of past conflict (interstate war, colonial war, genocide and civil war) and how it has been resolved. The necessity of categorising the type of war for better handling future conflicts has been established in [70]. One could also note the absence of population density in the variables, although some current examples like Rwanda show that it can be determinant as showed in [54].

The team that has built the dataset provides a thorough explanation of the treatment of the variables in [34] and [36] which we do not describe again here. It includes replacing missing values by these found in other historical records or by taking regional means from similar countries [34]. Population size, Homicide Rate, Infant Mortality, GDP Per Capita, Economic Openness and Oil production are log-transformed and all the variables have been rescaled from 0 to 10. Table 1.3 provides more information about the variables' categories. The indicators are divided into five risk areas and ten concepts. Appendix A provides details about the sources of the variables and the precise name of the datasets they come from. Some indicators' names can be misleading for the reader since their score is proportional to the inverse of what their name means. Years Since HVC, Empowerment Rights, Food Security, GDP Per Capita and Government Effectiveness are the variables for which this applies. From now on, they will be followed by a star * to remind the reader of it.

Risk area	Concept	Type of variable
Political	Regime type	Regime type
		Lack Of Democracy
	Regime performance	Government Effectiveness* Level Of Repression Empowerment Rights*
Security	Current conflict situation	Recent Internal Conflict Neighbours with highly violent conflict
	History of conflict	Years Since Highly Violent Conflict*
Social	Social cohesion and diversity	Ethnic Power Change (National power) Ethnic Compilation (Subnational) Transnational Ethnic Bonds
		Corruption
	Public security and health	Homicide Rate Infant Mortality
Economy	Development and distribution	GDP Per Capita* Income Inequality Economic openness
		Food Security* Unemployment
	Provision and employment	
Geographical and demographic	Geographic challenge	Water Stress Oil Production Structural Constraints
		Population Size
	Demographics	Youth Bulge

Table 1.3: The variables categories. The variables are divided into 5 key categories and 10 concepts.

Score	Meaning
0	The group has no close kindred across an international border
3.333	The group has close kindred across a border which does not adjoin its regional base
6.667	The group has close kindred in one country which adjoins its regional base
10	The group has close kindred in more than one country which adjoins its regional base

Table 1.4: Transnational ethnic bonds scores (adapted from [34])

1.4.1 Variables description

1. Years since HVC*: describes the number of years since last high violent conflict. A high violent conflict is defined as a quarrel where there are at least 500 casualties. A value of 10 corresponds to very recent HVC while a value of 0 indicates that the last HVC happened many years ago.
2. Max value: National And Subnational Intensity which is referred in table 1.3 as Recent Internal Conflict: indicates the max value between the NP and SN intensity value for the current year. This scale depends on the number of occurring conflicts and of casualties. It is thoroughly described in [34]
3. Neighbours With Highly Violent Conflict: gives the higher current score for conflicts intensity, either SN or NP, in countries sharing a common boarder.
4. Corruption: indicates a measure of corruption, adapted from the World Bank indicators [85]
5. Transnational Ethnic bonds: quantifies the ethnic dispersion over the boarder following Table 1.4.
6. Economic Openness: weighted sum of foreign investment indicators and export of goods and services indices.
7. Empowerment rights*: measures the government's respect for the following human rights: foreign movement, domestic movement, freedom of speech, freedom of assembly and association, workers' rights, electoral self-determination, and freedom of religion. A value of 0 indicates advanced rights while a score of 10 indicates many restrictions.
8. Ethnic Power Change: depicts the transition of power status among ethnic groups
9. Ethnic Compilation: the maximal amount of power detained by an ethnic group
10. Food Security*: this variable describes food insecurity such that a score of 10 indicates no security at all and a score of 0 full security. It is based on several criteria. Among

them are price volatility, hunger indices, domestic food price level index and the type of diet energy supply.

11. Oil Production: it corresponds to the fuel export as a proportion of the merchandise export. A log transformation is applied, and the result is linearly rescaled from 0 to 10.
12. GDP Per Capita*: it is the gross domestic product divided by the number of inhabitants. A value of 10 means a very low GDP per capita and conversely for a value of 0.
13. Government Effectiveness*: measures the quality of public and civil services, policy formulation and credibility of government as perceived by the population. A value of 0 corresponds to a fully (perceived) effectiveness and conversely for a value of 10.
14. Homicide Rate: depicts the number of homicides per 100 000 persons (armed-conflict casualties are excluded).
15. Income inequality: is derived from the Gini coefficient of net incomes. Gini coefficient is a measure of dispersion introduced by Italian statistician Corrado Gini to depict income inequalities in [28]. It is defined as

$$\frac{1}{2N^2\bar{r}} \sum_{i=1}^N \sum_{j=1}^N |r_i - r_j| \quad (1.7)$$

with $\bar{r}$ being the mean revenue, N , the total number of individuals and r_i , the revenue of individual i . It must not be confounded with Gini impurity that will be introduced in section 3.4.1.

16. Infant Mortality: describes the number of babies deceased before the age of 5 per 1000 inhabitants.
17. Population Size: no additional description required
18. Lack Of Democracy: it corresponds to an index of autocratic features institutionalisation minus an index of democratic features institutionalisation. A value of 0 corresponds to a completely democratic state while a value of 10 corresponds to a totally autocratic state.
19. Regime Type: describes the regime type with indices found in Table 1.5. The score corresponds to the stability resulting from the considered type of regime. As a matter of fact, democracies and dictatorships tend to show stability, contrarily to anocracies¹ [93].
20. Level of repression: measures the political violence and terror. This is related the suppression of oppositional movements, the use of force against separatist movements, restricting the work of the press or other elements related to repression.

¹regime that mixes democratic and autocratic features

Type	Score	Definition
Full autocracy	1.07	Systems that combine an absence of effective competition for chief executive with repressed or suppressed political participation
Partial autocracy	3.98	Hold competitive elections for national office but repress or tightly control participation or allow subtraction political participation but fail to subject the office of chief executive to truly competitive elections
Partial democracy without factionalism	3.91	Systems in which the chief executive is chosen through competitive elections and political competition is not effectively repressed, but either elections are not fully free and fair, or political participation is not fully open and well institutionalised
Partial democracy with factionalism	10.00	Factionalism is a pattern of sharply polarised and uncompromising competition between blocs pursuing parochial interest at the national level
Full democracy	1.00	Systems that combine free and fair elections with open and well-institutionalised political participation
Transition	4.49	Regimes in transition, no assessment of the current level

Table 1.5: Regime type scores (taken from [14]). As explained, the scores are related to the instability resulting from the indicated regime type and that is how they are ranked between 0 and 10.

21. Structural Constraints: assesses the “quality of governance during processes of transition”. A score of 10 represents a poor quality of governance during transition processes and conversely for a score of 0. This score is related to the constraints that are not due to current political leadership and cannot be changed easily. These constraints are for example extreme poverty, lack of an educated labour force, a disadvantageous geographical location, severe infrastructural deficiencies, natural disasters or pandemics [75].
22. Unemployment: depicts the proportion of unemployed persons in the total labour force.
23. Water Stress: it is the mean of the risk of water shortage for agricultural, domestic, and industrial use [34]. These scores are not time-series : they are given for each country and stay the same throughout the years [34].
24. Youth Bulge: represents the number of inhabitants aged between 15 and 24 divided by the number of inhabitants older than 24.

1.4.2 Data discussion

The data comes from institutions such as the University of Uppsala, the World Bank or ETH Zurich. They are mainly international organisations or academic institutions. We do not pretend to be qualified to criticise the work of scientists working there. However, as data scientist, we have to question the pertinence and reliability of our data. Here, we believe that we do not need to worry since it is broadly accepted in the peace-building community that these institutions are reliable. Additionally, we presume that the European Union cares

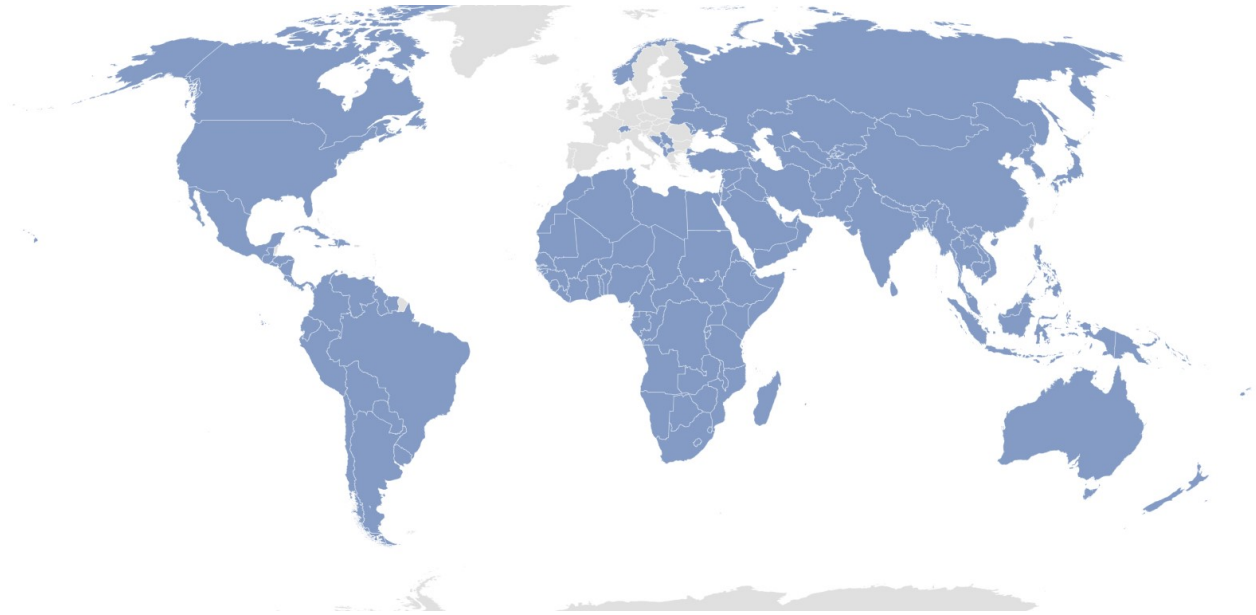


Figure 1.2: World map with countries considered in the GCRI dataset highlighted in blue. As explained previously, the European Union countries are not part of it. Furthermore, we see that the other absent countries are very small states.

for the data quality since it has a big interest in having reliable predictions. It also disposes of great technical and financial means to achieve this.

Even if the dataset has been built for a machine learning purpose, the datasets that have been used to constitute it are not, most of the time. Many of them are mostly used for statistical purposes. GCRI researchers have done an important imputation work for missing values, mostly based on other historical records or regional means [35]. We could fear that Ethnic Power Change, Ethnic Compilation and the Regime type are categorical nominal values that might be a source of problem. However, they have been ordered in a way that makes sense. For example, the scale of Regime Type refers to the stability resulting from the type of regime (see Table 1.5). The link between the variables (linear or not) is not assessed and problems might come from that. As an example, the Level Of Repression and the Regime Type might be strongly correlated. This might affect the classifier performances since handling more variables than necessary tends to decrease these performances. Additionally, the interpretation of the model could be misleading. It is possible that a model takes the information present in a variable via another variable that is correlated. Therefore, we will analyse the links between variables. Additionally, we are not sure that all variables are important for the prediction. That is why we will test their statistical significance.

In Table 1.6, the variable descriptive statistics are given. The count, mean, standard deviation, minimum, the 25, 50 and 75 percentiles, and the maximum of the variable are provided. It is also specified whether it is continuous or discrete. The first observation is that there

are no missing values, thanks to the GCRI team imputing work. As stated in a previous section, the dataset includes 2940 country-years samples. Eleven variables among the 24 are discrete and attention must be paid to this since it can bring difficulties in the computation. Note also that the Ethnic Power Change is used only for NP conflict, whereas the Ethnic Compilation value is used only for SN level classification. Except for Ethnic Power change, the standard deviations are mainly from 2 to 4, which seems high given the interval from 0 to 10. Most of the variables reach the 0-10 bounds.

Variable	count	mean	std	min	25%	50%	75%	max	Discrete/ Continuous
Years Since HVC*	2940	1.27	2.97	0.0	0.0	0.0	0.0	10.0	Discrete
Max Value: National and subnational intensity	2940	1.92	3.36	0.0	0.0	0.0	5.0	10.0	Discrete
Value of the intensity in neighbouring countries in HVC	2940	4.88	4.24	0.0	0.0	5.0	10.0	10.0	Discrete
Corruption	2940	5.9	2.06	0.0	5.18	6.19	7.31	10.0	Continuous
Transnational Ethnic bonds	2940	4.97	4.06	0.0	0.0	3.33	10.0	10.0	Discrete
Economic Openness	2940	4.39	1.67	0.0	3.27	4.34	5.49	10.0	Continuous
Empowerments Rights*	2940	4.75	2.9	0.0	2.14	5.0	7.14	10.0	Discrete
Ethnic Power Change	2940	0.16	0.97	0.0	0.0	0.0	0.0	10.0	Discrete
Ethnic Compilation	2940	3.63	2.21	1.0	1.0	5.0	5.0	10.0	Discrete
Food Security*	2940	5.43	2.12	0.0	3.72	5.75	7.13	10.0	Continuous
Oil Production	2940	3.62	3.7	0.0	0.0	2.48	6.91	10.0	Continuous
GDP Per Capita*	2940	5.26	1.93	0.15	3.98	5.29	6.86	10.0	Continuous
Government Effectiveness*	2940	5.73	1.75	0.0	4.96	5.96	6.86	9.94	Continuous
Homicide Rate	2940	4.51	2.56	0.0	2.63	4.86	6.15	10.0	Continuous
Income Inequality	2940	4.42	1.55	0.0	3.31	4.36	5.46	10.0	Continuous
Infant Mortality	2940	5.92	2.16	0.08	4.43	6.14	7.74	10.0	Continuous
Population Size	2940	4.1	1.87	0.0	2.87	4.03	5.25	9.98	Continuous
Lack of Democracy	2940	4.35	3.31	0.0	1.0	3.5	8.0	10.0	Discrete
Regime Type	2940	4.39	3.03	1.0	1.07	3.91	3.98	10.0	Discrete
Level Of Repression	2940	5.01	2.77	0.0	2.5	5.0	7.5	10.0	Discrete
Structural Constraints	2940	6.12	2.4	1.0	5.0	6.0	8.0	10.0	Discrete
Unemployment	2940	5.24	2.02	0.0	3.97	5.34	6.51	10.0	Continuous
Water Stress	2940	4.46	2.29	0.0	2.78	3.87	6.42	10.0	Continuous
Youth bulge	2940	6.21	2.53	0.0	4.56	6.93	8.21	10.0	Continuous

Table 1.6: Descriptive statistics of GCRI dataset variables. It is also specified whether the variable is discrete or continuous. There are no missing values as can be observed in the first column.

Chapter 2

ANOVA and mutual information

In this chapter, we measure the strength of the links between each input feature and the outcome, through two techniques. Additionally, we compute the links between each pair of input variables. This allows us to determine the important features for the classification task.

2.1 Variables significance: analysis of variance

Suppose that we are not sure about the relevance of all our variables. And we imagine that a variable that is relevant would have a different value depending on the outcome, conflict or not. Let us introduce a small conceptual example.

Example 1

In this small example, we imagine that we only dispose of two variables, Corruption and Unemployment. For the sake of simplicity, we only consider NP fights. Part of a fictitious dataset is given in Table 3.1.

	Country	Year	Corruption	Unemployment	Outcome (NP) y
1	Canada	2008	0.16	4.93	0
2	Yemen	2007	6.69	7.43	1
3	Angola	2001	8.69	5.3	1
4	Angola	2002	7.76	5.2	0
5	RDC	2002	8.06	5.73	1
...	...	...	...	...	...

Table 2.1: Toy dataset for conceptual example with 5 country-years and only two variables

We want to know whether Corruption may vary when the outcome y changes. Therefore, we check whether the mean of Corruption values for countries with conflict, μ_1

and the mean of Corruption values for countries without conflict, μ_0 show significant difference (here, we only consider the values that are printed in Table 3.1):

$$\mu_1 = \frac{1}{3}(6.69 + 8.69 + 8.06) = 7.81 \quad (2.1)$$

$$\mu_0 = \frac{1}{2}(0.16 + 7.76) = 3.96 \quad (2.2)$$

We need to use a technique that shows the magnitude of the difference between the means.

2.1.1 Analysis of variance: theory

To test for statistical significance, the proposed method is the analysis of variance (ANOVA). In its simplest form, it is a model to verify whether two means of two samples come from the same population. It has been formally introduced in 1919 by Ronald Fisher in [19]. Our goal is to apply this technique to determine for each variable whether the groups of positive data points (such that $y_i = 1$) has a different mean from the negatives ones. Let us define $x_i^{(1)}$ as the x_i 's such that $y_i = 1$ and $x_i^{(0)}$ as the x_i 's such that $y_i = 0$. Additionally, suppose that n_1 is the number of positive instances, n_0 the number of negative instances, $\bar{x}_j^{(k)} = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{ij}^{(k)}$ as the mean of category k and the overall mean $\bar{x}_j = \frac{1}{n_0+n_1} \sum_{k=0}^1 \sum_{i=1}^{n_k} x_{ij}^{(k)}$.

Model assumptions

1. Normality of the realisations
2. The independence of the realisations
3. Homoscedasticity of the realisations: $\mathbb{V}[x_{\cdot j}^{(0)}] = \mathbb{V}[x_{\cdot j}^{(1)}] = \sigma^2 \quad 1 \leq j \leq p$
 where $x_{\cdot j}^{(k)}$ is the vector containing all the value of variable j for all instances. $\mathbb{V}[x_{\cdot j}^{(k)}]$ is the variance over all elements of vector $x_{\cdot j}^{(k)}$.

We want to test for each j (with $1 \leq j \leq p$), in other words, for each variable:

$$H_0 : \bar{x}_j^{(0)} = \bar{x}_j^{(1)} \quad (2.3)$$

$$H_a : \bar{x}_j^{(0)} \neq \bar{x}_j^{(1)} \quad (2.4)$$

If we cannot reject H_0 , we cannot conclude anything and variable j is statistically insignificant. Conversely, if H_0 is rejected and H_a cannot be rejected, then the variable j is statistically significant. The informal interpretation is : the chance is high that variable j 's values come from two populations with different mean depending on whether instance i is positive or negative. It is then considered to be significant in the determination of the answer y_i .

ANOVA assumes that $x_{ij}^{(k)}$ for a given j and $k = 0$ or $k = 1$ are independently and normally distributed as follows:

$$x_{ij}^{(k)} \sim \mathcal{N}(\mu_k, \sigma^2) \quad 1 \leq i \leq n_k, \quad k = 0, 1 \quad (2.5)$$

or equivalently

$$x_{ij}^{(k)} = \mu_k + \epsilon_{ij}^{(k)} \quad \text{with } \epsilon_{ij}^{(k)} \sim \mathcal{N}(0, \sigma^2) \quad (2.6)$$

We additionally define:

- $SS_B = \sum_{k=0}^1 n_k (\bar{x}_j^{(k)} - \bar{x}_j)^2$ as the "between groups" sum of squares
- $SS_W = \sum_{k=0}^1 \sum_{i=1}^{n_k} (x_{ij}^{(k)} - \bar{x}_j^{(k)})^2$ as the "within group" sum of squares

Theorem 1. Under H_0 , $\frac{SS_B}{K-1} \frac{n-K}{SS_W}$ follows an F -distribution (Fisher-Snedecor) (see [78] for proof):

$$\frac{SS_B}{K-1} \frac{n-K}{SS_W} \sim F_{K-1, n-K}$$

with K , the number of groups considered.

We can then compute p -values and assess statistical significance. The interpretation of p -values is the following : it is the estimated probability of observing $x_{.j}$ under H_0 . Therefore, it can be criticised since we compute $\mathbb{P}[x_{.j}|H_0]$ when we want $\mathbb{P}[H_0|x_{.j}]$ which are not equal quantities. It is however widely used and provides often satisfying results. Then we compare the p -value to a significance threshold α chosen before the computations. If $p \leq \alpha$, we reject H_0 and that means that the variable is statistically significant. Otherwise, we reject H_a and that means that the variable is not statistically significant. There is no specific rule for fixing α . Fisher suggests in [20] to use 0.05. Others use 0.01. Nevertheless, it can vary if a task demands a big precision. The whole ANOVA procedure is given in Algorithm 1.

Normality test

To test for normality, we choose the Shapiro-Wilk test whose good power properties are established by Yap & Sim [95] when no specific hypothesis can be done about the data. This test has been proposed by Shapiro and Wilk in [79]. It tests the null hypothesis that the sample $\mathbf{u} = (u_1, \dots, u_n)$ is normally distributed. To test for normality, the following test statistic is computed:

$$W = \frac{(\sum_{i=1}^n a_i u_{(i)})^2}{\sum_{i=1}^n (u_i - \bar{u})^2} \quad (2.7)$$

where $u_{(i)}$ corresponds to the i^{th} order statistic, meaning the i^{th} number when the sample is ordered in ascending order. $\bar{u}$ is the sample mean. The a_i are determined the following way:

$$\mathbf{a} = (a_1, \dots, a_n) = \frac{m^T V^{-1}}{C} \quad (2.8)$$

Consider the random variables $z_1, \dots, z_n \sim \mathcal{N}(0, 1)$ and take their order statistics $z_{(1)}, \dots, z_{(n)}$. The vector m is defined as follows:

$$m = (m_1, \dots, m_n) = (\mathbb{E}[z_{(1)}], \dots, \mathbb{E}[z_{(n)}]) \quad (2.9)$$

V , on the other hand, is the covariance matrix of $z_{(1)}, \dots, z_{(n)}$ such that

$$V_{kl} = \mathbb{E}[(z_{(k)} - m_k)(z_{(l)} - m_l)] \quad (2.10)$$

C is a normalisation factor:

$$C = \|V^{-1}m\|_2 = (m^T V^{-1} V^{-1} m)^{\frac{1}{2}} \quad (2.11)$$

The W distribution has no name. The cut-off value for the statistics are computed using Monte-Carlo simulations. To apply it to our problem, it is enough to replace u_i by $x_{ij}^{(k)}$ given $1 \leq i \leq n_k$ and $k = 0, 1$. The detailed procedure is given in Appendix B in Algorithm 7.

Homoscedasticity test

We now have to test for homoscedasticity ($H_0: \mathbb{V}[x_j^{(0)}] = \mathbb{V}[x_j^{(1)}] = \sigma^2$ for each j). Most common statistical tests are Levene, Bartlett and Brown-Forsythe tests. However, Bartlett is sensitive for non-normality and Levene or Brown-Forsythe are preferred [13]. Between these, B-F test is the most robust one against many types of non-normal data while preserving good statistical power [15]. This is the one we use here for this reason.

The Brown-Forsythe test is based on an ANOVA on a transformation of the variable. Let us define

$$v_{ij}^{(k)} = |x_{ij}^{(k)} - \tilde{x}_j^{(k)}| \quad (2.12)$$

$\tilde{x}_j^{(k)}$ being the median of variable j for the group k . Performing a one-way ANOVA on $v_{ij}^{(k)}$, it gives:

$$F = \frac{n - K}{K - 1} \frac{\sum_{k=0}^{K-1} n_k (\tilde{v}_j^{(k)} - \tilde{v}_j)^2}{\sum_{k=0}^{K-1} \sum_{i=1}^{n_k} (v_{ij}^{(k)} - \tilde{v}_j^{(k)})^2} \quad (2.13)$$

With n being the total number of samples, K being the number of groups (two in this case), n_k corresponding to the number of samples in group k , $\tilde{v}_j^{(k)}$, the median of v_{ij} for group k , $\tilde{v}_j$ the overall median of v_{ij} . It is an F-statistic with degrees of freedom $d_1 = K - 1$ and $d_2 = n - K$. So under H_0 , the variances for group 0 and group 1 are equal. One could argue that ANOVA is used on v_{ij} without checking if independence, normality and homoscedasti-

city are met. The detailed procedure is given in Appendix B in Algorithm 8.

Algorithm 1: ANOVA

```

1 Determine significance threshold  $\alpha$ 
2 for  $j = 1 : p$  do
3    $x \leftarrow x_{.j}$ 
4    $x^{(0)} \leftarrow x_{ij}$  such that  $y_i = 0$ 
5    $x^{(1)} \leftarrow x_{ij}$  such that  $y_i = 1$ 
6    $n_0 \leftarrow \text{length}(x^{(0)})$ 
7    $n_1 \leftarrow \text{length}(x^{(1)})$ 
8    $K \leftarrow 2$ 
9   test normality  $x^{(0)}$ 
10  test normality  $x^{(1)}$ 
11  test homoscedasticity between  $x^{(0)}$  and  $x^{(1)}$ 
12  Evaluate sample independence likelihood
13  if All assumptions met then
14    | The following results are reliable
15  else if Independence and (normality or homoscedasticity) assumptions are met then
16    | The following results still reliable
17  else
18    | The following results must be interpreted carefully
19  end
20
21   $\bar{x}^{(0)} \leftarrow \frac{1}{n_0} \text{sum}(x^{(0)})$ 
22   $\bar{x}^{(1)} \leftarrow \frac{1}{n_1} \text{sum}(x^{(1)})$ 
23   $\bar{x} \leftarrow \frac{1}{n} \text{sum}(x)$ 
24   $\text{SS}_B \leftarrow n_0(\bar{x}^{(0)} - \bar{x})^2 + n_1(\bar{x}^{(1)} - \bar{x})^2$ 
25   $\text{SS}_W \leftarrow \sum_{i=1}^{n_0} (x_j^{(0)} - \bar{x}^{(0)})^2 + \sum_{i=1}^{n_1} (x_j^{(1)} - \bar{x}^{(1)})^2$ 
26   $F \leftarrow (n - K) \text{SS}_B / ((K - 1) \text{SS}_W)$ 
27   $p \leftarrow p\text{-value obtained from } F \text{ by numerical computation}$ 
28  if  $p \leq \alpha$  then
29    | Reject  $H_0$ , this implies that variable  $j$  is statistically significant
30  else
31    | Reject  $H_a$ , this implies that variable  $j$  is not statistically significant
32  end
33 end

```

ANOVA suitability for this situation

ANOVA may not seem suitable for our situation because of its assumptions. We do not know yet whether normality and homoscedasticity can be respected. It is known that AN-

OVA is robust for non-normality (see [78] for details). But we are sure that the assumption of independence is not met. Indeed, the measures are repeated each year in each country. Furthermore, it would be difficult to accept that measurements in neighbouring countries in the same year are independent. It is even barely acceptable to assume that two measures in any two countries in a same year are independent because of globalisation.

Example 1 (continued)

Looking at our conceptual example, the corruption value in Angola in 2001 and the one in 2002 are related. They are not independent. In addition, it is difficult to imagine that there is no link between Unemployment values in RDC and in Angola which are neighbouring countries. Furthermore, we cannot pretend, in a globalised world, that Yemen (2007) and Canada (2008) Unemployment value are totally independent variables.

The non-respect of independence assumption makes the interpretation of ANOVA test difficult. Most of the time, it leads to a higher ratio for the F-statistics and consequently a lower significance level, but the opposite can also be true [76]. It potentially results in rejecting null-hypothesis too often, a type I error. A variable can then be declared significant, although it is not.

More generally, Sahai & all. [76] provide us with the following conclusion. ANOVA results might still be valid if one of the two last assumptions is not respected (normality or homoscedaticity). However, one cannot trust results of an ANOVA with two or more violated assumptions.

However, ANOVA has the advantage of being easy to interpret. It is used most of the time as a first test to gain understanding into the significance level of variables.

We could take more complex models such as repeated measures ANOVA, but it is little adapted to our problem since not all countries have the same number of years of measurements. We would lose the advantage of keeping a basic model.

Another important thing to discuss is that we take each variable separately. Which means that joint-significance is not tested. We could imagine the situation of two variables being insignificant if taken separately but jointly significant. Therefore, multivariate ANOVA should be part of future work.

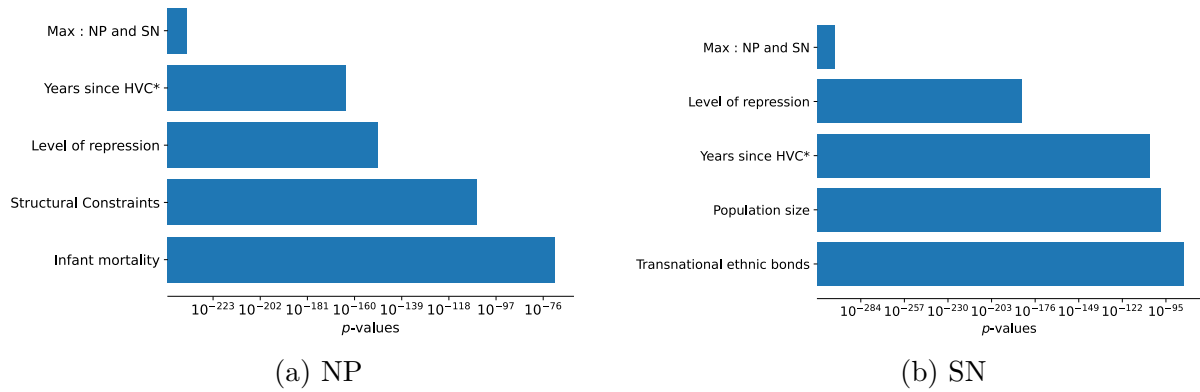


Figure 2.1: The five most significant variables according to ANOVA and their p -values on a log-scale. The exact interpretation of the p -value is the following : it is the probability to observe the sample $x_{.j}$ given that H_0 is verified, meaning that the mean of the values for positive instance is the same as the mean of the values for the negative instance for the variable given on the y -axis. Let us recall that, the smaller the p -value, the most significant the variable is.

2.1.2 Results of ANOVA and assumptions test

The results are reported in Table 2.2. By taking the five smallest p -values for each category, one can derive the most significant variables. They are reported in Figure 2.1.

We first observe that, even at a significance level of 0.01, the null hypothesis of normality is rejected for the two groups, both for SN and NP cases (except for economic openness in the NP case for group 1). Because of this, it is necessary to choose a test for homoscedasticity robust for non-normality. Fortunately, as explained above, the Brown-Forsythe test is barely affected by non-normality. Furthermore, as said in the previous section, it is known that ANOVA is robust for non-normality.

Additionally, the hypothesis of homoscedasticity is rejected for all variable except for Transnational Ethnic Bonds, Economic Openness, Income Inequality, Unemployment Rate and Water Stress for NP cases; Economic Openness, Oil Production, Population Size, Regime type and Structural Constraints for the SN case (at 0.01 significance level). For the pre-cited variables, we cannot reject the hypothesis of equal variances for both group. Note that at a 0.05 significance level, it is even worse. Since we also reject the null hypothesis for Unemployment Rate and Water stress (NP case) and Income inequality and population size (SN case). The non-respect of equal variance assumption has a limited impact on the results as long as the samples are balanced, which is not the case here (2348 vs 592 and 2046 vs 894 for NP and SN respectively, the biggest number corresponding to the negative instances). It then compromises the validity of the test. Generally, we observe that the hypothesis for ANOVA are

Variables	National Power				Subnational level			
	Normality group 0	Normality group 1	Homoscedaticity	ANOVA	Normality group 0	Normality group 1	Homoscedaticity	ANOVA
Years since HVC*	0.00E+00	4.51E-29	4.08E-235	8.62E-165	0.00E+00	1.02E-37	4.63E-106	4.63E-106
Max value: national and subnational intensity	0.00E+00	8.47E-25	1.97E-89	1.22E-235	0.00E+00	1.89E-29	1.28E-219	8.93E-301
Neighbours with highly violent conflict	0.00E+00	2.15E-29	6.21E-40	1.30E-64	0.00E+00	9.91E-33	1.32E-11	6.11E-40
Corruption	7.68E-35	2.57E-04	2.07E-18	1.86E-62	2.21E-33	8.39E-20	2.56E-11	9.50E-60
Transnational ethnic bonds	2.80E-45	2.64E-27	2.27E-01	2.31E-20	4.79E-43	4.24E-35	1.06E-03	1.06E-84
Economic openness	4.97E-05	5.88E-01	5.68E-01	7.69E-33	3.74E-04	6.43E-03	5.95E-01	6.88E-26
Empowerment rights*	6.17E-29	2.75E-09	4.49E-14	3.69E-23	1.64E-28	9.76E-12	2.33E-20	1.67E-16
Ethnic power change /compilation	0.00E+00	2.27E-43	9.72E-08	9.72E-08	0.00E+00	1.26E-35	4.91E-30	1.75E-46
Food security*	1.31E-18	4.18E-10	1.36E-21	6.09E-54	9.99E-15	4.49E-19	9.12E-04	3.07E-33
Oil production	2.94E-44	6.34E-25	6.44E-07	1.41E-05	6.67E-43	4.99E-28	5.15E-01	6.22E-03
GDP per capita*	1.88E-14	9.76E-12	2.52E-09	3.65E-49	8.37E-15	1.37E-10	7.43E-08	3.34E-40
Government effectiveness*	4.68E-28	3.64E-09	7.06E-09	9.38E-71	3.55E-28	1.19E-09	1.66E-03	3.95E-39
Homicide rate	3.05E-22	1.67E-08	5.56E-17	8.10E-08	7.84E-22	2.80E-08	6.37E-19	1.34E-10
Income inequality	5.60E-13	4.78E-10	3.81E-01	3.54E-02	2.01E-13	1.41E-06	5.44E-03	1.07E-01
Infant mortality	3.63E-20	2.87E-13	4.84E-33	1.86E-71	1.82E-18	5.68E-19	5.25E-09	1.79E-54
Population size	6.15E-15	4.78E-10	7.22E-05	3.17E-32	1.56E-15	2.02E-10	7.83E-02	5.92E-99
Lack of democracy	3.98E-39	3.82E-17	1.72E-08	8.05E-13	3.40E-37	6.06E-22	1.76E-13	1.26E-01
Regime type	0.00E+00	4.25E-28	1.82E-05	5.79E-16	0.00E+00	1.39E-34	4.14E-01	1.58E-14
Level of repression	1.27E-35	5.23E-24	5.49E-03	1.20E-150	1.46E-34	1.19E-26	1.60E-03	8.12E-185
Structural constraints	9.90E-27	8.18E-20	1.84E-12	3.71E-106	4.11E-25	5.98E-22	2.69E-01	5.66E-69
Unemployment rate	2.25E-14	9.33E-11	2.74E-02	7.20E-01	1.12E-14	4.15E-12	1.16E-14	4.34E-02
Water stress	5.42E-24	1.04E-16	1.56E-02	4.26E-05	1.72E-23	4.04E-15	6.51E-16	3.89E-02
Youth bulge	1.51E-29	4.17E-16	1.32E-72	5.30E-54	3.30E-28	1.51E-23	1.39E-25	1.24E-27

Table 2.2: p -values obtained for the Shapiro-Wilk, Brown-Forsythe and ANOVA tests for each variable (group 0 being the negative instances and group 1 the positive ones). The p -values highlighted in bold are the one bigger than 0.01. For these, we cannot reject H_0 at 0.01 significance level.

barely respected.

For the ANOVA test itself, on the other hand, we observe significance for all variables (event at a 0.01 significance level) except for Income Inequality, Unemployment Rate and Water Stress for the NP case and Income Inequality, Lack of Democracy, Unemployment Rate and Water Stress for the SN case. At a 0.05 significance level, only Income Inequality (NP and SN cases) and Lack Of Democracy (SN case) are insignificant. In other words, most of the time, we have to reject the null hypothesis of equal means for conflict and peace instances.

In Figure 2.1, we observe that among the five most significant variables, the first three are the same for NP and SN but not in a similar order. The first one, the maximal value between NP and SN conflict intensity score on the given year, seems logical. The severity of actual conflict is a good indicator to predict future ones [35]. The same reasoning applies to the variable Years Since HVC*. The way we can interpret the p -values is the following. With a score of 1.22E-235, the conflict intensity (Max: NP and SN) is 10^{70} more significant than Years Since HVC*. However, as we discuss later, the non-respect of the assumptions makes the result interpretation less reliable.

As mentioned earlier, the non-respect of independence assumption makes the interpretation of ANOVA test difficult. It potentially results in rejecting null-hypothesis too often, a type I error. With the above-shown results, type I errors are particularly problematic. It would not have been a problem to have a higher type II errors rate with such results.

Since most of the time two or more assumptions are violated, we only consider the p -values obtained here carefully. In Chapter 3, we will observe the model's performances by removing the "insignificant" variables and compare it with two other models.

2.2 Mutual information

ANOVA showing limits and assuming linearity, we use a method that can account for non-linearity. We need mutual information, a concept coming from information theory. It can be summarised as "what amount of information do we have about random variable a if we know random variable b ?". This is possible both between the outcome and each input variable or for pairs of input variables.

Example 1 (continued)

Back to our conceptual example, what can we learn from random variable $\mathbf{y}$, the outcome if we know the Corruption vector $x_{.1} = (0.16 \ 6.69 \ 8.69 \ 7.76 \ 8.06 \dots)^T$. Additionally, we can consider the mutual information between Corruption and Unemployment.

2.2.1 Mutual information: theory

Except if specified otherwise, the content of this section is based on [31]. Information theory is the scientific discipline that theorises information quantification and communication. It has been formally introduced by Claude Shannon in the 40s, although Hartley already published a paper about concepts used later in this field [68]. One of the key ideas is to measure the information contained in a probability distribution. Shannon proposed a first measure called self-information:

$$I(x) = \log \left(\frac{1}{p(x)} \right) \quad (2.14)$$

where $p(x) = \mathbb{P}[X = x]$ and X is a random variable [7]. The base of the log only modifies the expression with a scaling factor. If this base is two, the self-information unit is the bit.

Example 2

To get this intuition, one can think about x as a message. For example, let us consider weather forecast announce 1: "this night, it will be dark until morning" and announce 2: "this night, there will be a snowstorm". The probability that it will be dark in the night is high, and this message does not contain much information. However, the second one contains more information since snowstorm are less likely (at least in our country).

Shannon introduced a second measure which gives the uncertainty over a probability space, the entropy¹:

$$H(X) = \mathbb{E}[I(x)] = - \sum_{x \in X} p(x) \log p(x) \quad (2.15)$$

Low entropy indicates low uncertainty over the outcome like the possibility of a volcanic eruption tomorrow : we can say that, most probably, it will not happen, so the uncertainty is low. Conversely, a high entropy indicates a high uncertainty over the outcome like the place where the lightning strikes. This measure can be extended: joint entropy for X and Y applies to their joint distribution.

$$H(X, Y) = \mathbb{E}_{X, Y}[-\log p(x, y)] = - \sum_{x, y} p(x, y) \log p(x, y) \quad (2.16)$$

If X and Y are independent, this corresponds to the sum of their own entropies. The conditional entropy of X given $Y = y$ is defined as

$$H(X|y) = \mathbb{E}_{X|y}[-\log p(x|y)] = - \sum_{x \in X} p(x|y) \log p(x|y) \quad (2.17)$$

with

$$p(x|y) = \frac{p(x, y)}{p(y)} \quad (2.18)$$

¹The name comes from the similarity of its expression with thermodynamic entropy

The conditional entropy of X given Y is given by

$$H(X|Y) = \mathbb{E}_Y[H(X|y)] = - \sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) \log p(x|y) \quad (2.19)$$

Conditional entropy obeys:

$$H(X|Y) = H(X, Y) - H(Y) \quad (2.20)$$

The intuition is: the uncertainty about X knowing Y is the uncertainty over X and Y minus the uncertainty over Y . All the above metrics can also be extended to continuous distributions if the sums are replaced by integrals and probability mass functions by probability density functions $f(x)$. This gives:

$$H(X) = - \int_X f_X(x) \log f_X(x) dx \quad (2.21)$$

$$H(X, Y) = - \int_Y \int_X f_{X,Y}(x, y) \log f_{X,Y}(x, y) dx dy \quad (2.22)$$

$$H(X|y) = - \int_X f_{X|Y}(x|y) \log f_{X|Y}(x|y) dx \quad (2.23)$$

$$H(X|Y) = \int_Y \int_X f_{X,Y}(x, y) \log \frac{f_Y(y)}{f_{X,Y}(x, y)} dx dy \quad (2.24)$$

In order to define mutual information, we provide here the formula of the Kullback-Leibler divergence. It is a measure of dissimilarity for two probability distributions:

$$D_{\text{KL}}(p(X)||q(X)) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \quad (2.25)$$

In the case of continuous variable, it takes the form

$$D_{\text{KL}}(f(X)||g(X)) = \int_X f(x) \log \left(\frac{f(x)}{g(x)} \right) dx \quad (2.26)$$

It is not symmetric. Therefore, it is not a true distance function in the mathematical sense.

We can then define the mutual information:

$$I(X, Y) = D_{\text{KL}}(P_{XY}||P_X \otimes P_Y) \quad (2.27)$$

where P_{XY} is X and Y joint distribution while P_X and P_Y are the marginal distributions. If the variables are independent, their joint distribution is equal to the product of their marginal distributions and the mutual information to zero. This quantity, always greater or equal to zero, is described as "the amount of information one can obtain from one random variable given another" [42]. More formally, it is the amount of bits (if the log is in base 2) that can be learned for one distribution by observing the other one. It can also be defined as

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} p_{X,Y}(x, y) \log \frac{p_{X,Y}(x, y)}{p_X(x)p_Y(y)} dy dx \quad (2.28)$$

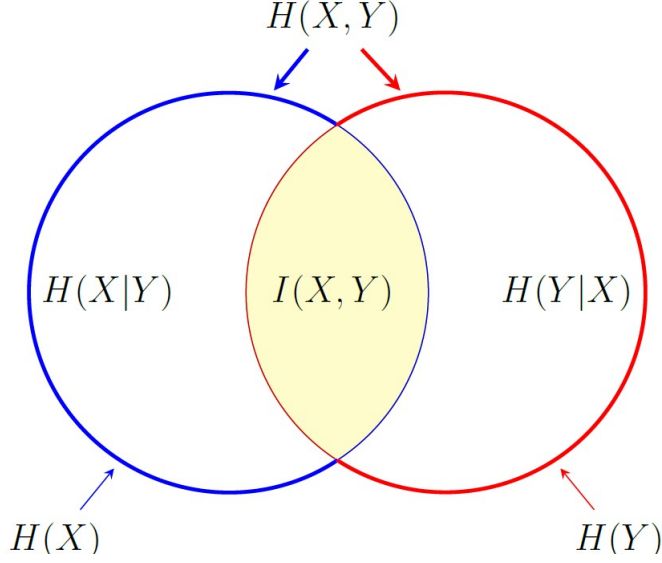


Figure 2.2: Schematic representation of mutual information and entropy in form of a Venn diagram extracted from [90]

for discrete variables and

$$I(X, Y) = \int_X \int_Y f_{X,Y}(x, y) \log \frac{f_{X,Y}(x, y)}{f_X(x)f_Y(y)} dx dy \quad (2.29)$$

for continuous variables. Alternatively, one can understand mutual information with the following equalities. A Venn diagram is given in Figure 2.2.

$$I(X, Y) = \begin{cases} H(Y) - H(Y|X) \\ H(X) - H(X|Y) \\ H(X) + H(Y) - H(X, Y) \end{cases} \quad (2.30)$$

The first one can be understood as: the mutual information is equal to the uncertainty in Y minus the uncertainty in Y that remains even if we know X . The second one is similar to the first one because of symmetry. The third one can be interpreted as the intersection between uncertainty of X and uncertainty of Y . If we know one of these, we remove the uncertainty it shares with the other one.

2.2.2 How to apply mutual information to our problem

We want to apply mutual information both between input variables x_k, x_l (with $k \neq l$) and between input variables x_k and outcome variable $\mathbf{y}$. This implies the calculation of $I(x_k, x_l)$ for $1 \leq k, l \leq p, k \neq l$ and $I(x_j, \mathbf{y})$ such that $1 \leq k \leq p$. Since mutual information is symmetric, we save half the computations for the between input-variables computations. This way, we seek to get a better insight into the link between the variables. And we also want to determine which variables are the most important ones for the prediction of the

outcome in light of this tool. However, because of its expression, mutual information is sometimes difficult to compute. When X and Y are discrete, it can be done using equation 2.28 or an equivalent one. However, when one or the two variables are continuous, most of the time, we do not dispose of the probability density function expression. Therefore, MI needs to be estimated and we have chosen two techniques that are adapted to the task. For the between input variables mutual information, we selected a technique presented in [25] since it can handle all situations, both discrete, both continuous or a mix of the two. We use the implementation provided by the authors. For the mutual information between input variables and outcome variable, we use the toolbox in scikit learn since it is specifically built for classification target variables and can handle both continuous and discrete other variables [64]. It is based on techniques developed in [50] and [73].

Mutual information is a powerful tool since it can handle any type of link between the variables, contrarily to Pearson correlation. However, as it needs to be estimated when dealing with continuous variables, it can suffer from imprecise results. It is still a wide research field today [25]. However, we do not aim to discussing mutual information estimation here. Additionally, it can be difficult to interpret the mutual information values if taken alone.

Example 1 (continued)

For example, are 2 bits of information enough to get the information about Corruption contained in Unemployment ? If we only dispose of partial information, is it a part of information that is relevant to my problem ?

The main interest is to compare these numbers between them in order to rank the link between variables. This is not as Pearson's correlation which indicates the sense of the relation and whose value is interpretable.

2.2.3 Pairwise mutual information

In Figure 2.3, the mutual information is shown for each pair of variables. We observe that Water Stress is linked with many other variables since the values are large on this line. All these are above 0.5 except Years Since HVC*, Max Value: NP and SN, (Economic) Openness and Ethnic Power Change. This implies that we can learn many things from this variable. Water Stress is a constant variable over time. This indicates that it is general information about the country. As we know, the other variables tend to show small evolution from years to years. We observe that the variables it is less related to are the ones changing more often : Years Since HVC*, Max Value : NP and SN and Ethnic Power Change tend to vary more than Structural Constraints. Therefore, we suppose that Water Stress acts as a proxy variable for the country, meaning that we could use it instead of the country name. It allows us to derive many national characteristics that do not change much over time. But this does not imply that it has a good predictive performance. Since it does not evolve

from year to year, it is not possible for the model to take account of any change in the country.

The same trends apply, to a lesser extent, for Corruption, Oil Production, GDP Per Capita*, Government Effectiveness*, Food Security* and Homicide Rate. Still to a lesser extent, Structural Constraints is also linked with many other variables. One could imagine that we could keep most of the information by taking only these variables. But it is not straightforward. Even if these specific variables show higher link with all the other ones, it is difficult to assess whether this bound is strong enough to transmit the needed information. Water Stress is a typical example of such problem: it would maybe allow to determine the mean value of Government Effectiveness* since their mutual information is 2.2 (even though we are not sure that 2.2 bit of information is enough to determine this mean), but it does not allow us to determine Government Effectiveness* evolution. Additionally, it is possible that the information carried by this subset is redundant.

One can also spot the variables that have lower mutual information with all other variables. This suggests that we need to keep these variable if they are pertinent for the prediction because we cannot find the information it contains in other variables (or only a small part). These are Ethnic Power Change, Years Since HVC*, Neighbours with HVC, Max Value: NP and SN, and (Economic) Openness.

2.2.4 Mutual information between input variables and outcome

The mutual information values are available in Table 2.3. The five variables with highest scores are given in Figure 2.4. We observe that they are the same, which allows us to plot them on one graph. As Water Stress has the highest MI with $\mathbf{y}$, we imagine that the best predictor for the outcome is the country itself. Relation between water stress and conflicts has been strongly established in [43, 49] and this gives credit to our results. Except the fifth one, all the others had high MI with other variables. It can be that they contain much information that are related to the outcome. What is curious is that SN has higher scores than NP. It can be explained by the fact that NP fights are more seldom events, with less regular patterns. The second surprising element is that almost the same order is observed for both, up to one element. Looking at the Table 2.3, we observe that Ethnic Power Change has the lowest MI with the outcome for NP predictions. This could indicate that, although we cannot retrieve information about this variable in other ones due to weak MI, it is not needed for the prediction since not as pertinent as other elements for the outcome. As we shall see in the discussion, this reasoning has its limitations.

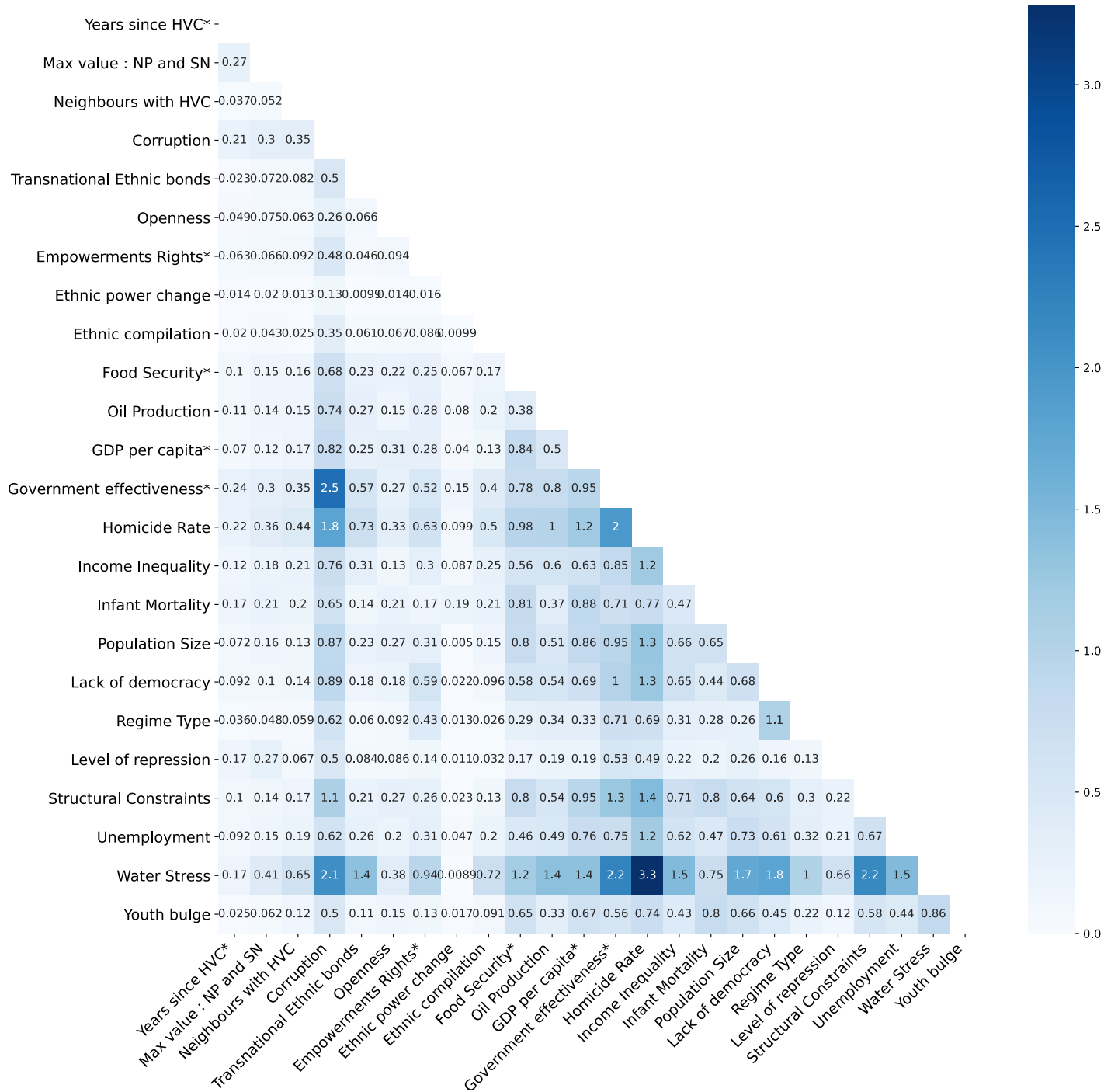


Figure 2.3: Mutual information matrix between the input variables. Entry k, l is equal to $I(x_k, x_l)$ or intuitively to the amount of information about variable k that we can retrieve by knowing variable l (or conversely since mutual information is symmetric).

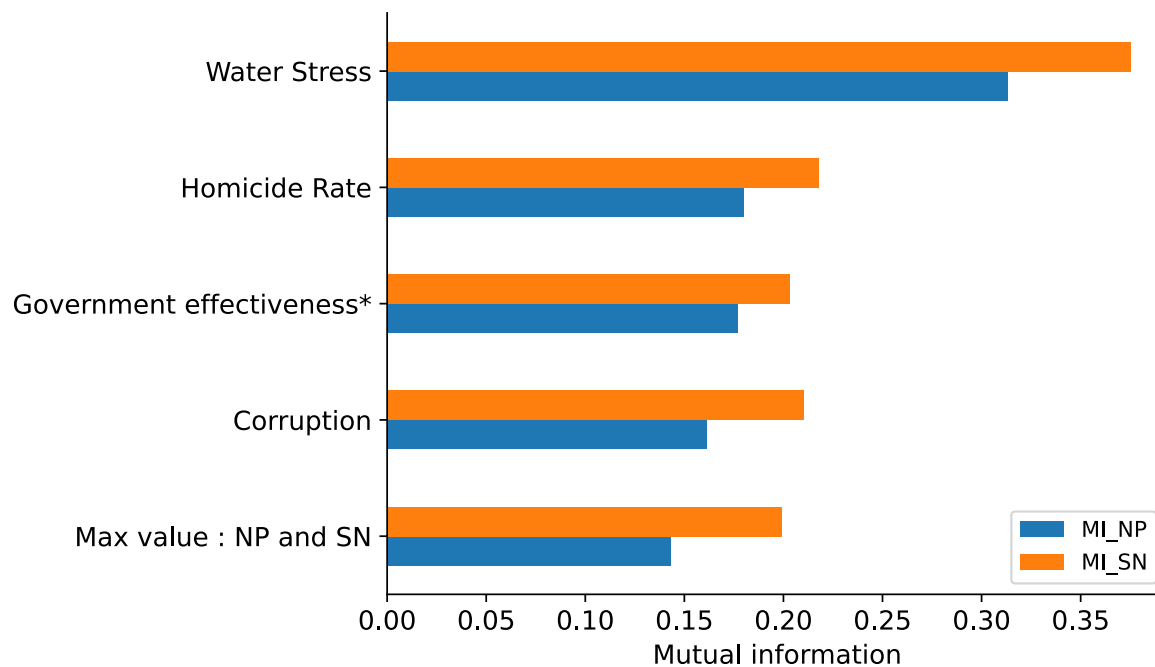


Figure 2.4: The five variables with highest mutual information with $\mathbf{y}$. On the ordinate, the name of the considered variable and on the abscissa, the mutual information $I(x_k, \mathbf{y})$. The values are ordered following NP scores. It is not more important than SN, but one order had to be chosen.

	NP	SN
Water Stress	0,313	0,375
Homicide Rate	0,18	0,218
Government effectiveness*	0,177	0,203
Corruption	0,161	0,21
Max value: NP and SN	0,143	0,199
Income Inequality	0,122	0,096
Level of repression	0,119	0,142
Structural Constraints	0,115	0,088
Years since HVC*	0,092	0,074
GDP per capita*	0,09	0,057
Oil Production	0,076	0,093
Youth bulge	0,07	0,043
Food Security*	0,068	0,078
Infant Mortality	0,058	0,055
Lack of democracy	0,055	0,059
Neighbours with HVC	0,052	0,031
Population Size	0,051	0,116
Unemployment	0,048	0,047
Openness	0,041	0,044
Regime Type	0,036	0,029
Empowerments Rights*	0,029	0,03
Transnational Ethnic bonds	0,017	0,066
Ethnic power change/compilation	0,005	0,048

Table 2.3: Mutual information between $\mathbf{y}$ and x_j for each j . It is ordered according to the NP scores. It is not more important than the SN scores, but we had to choose one of the two to increase readability.

2.3 ANOVA and mutual information: comparison and discussion

The main problem with ANOVA is the non-respect of the hypothesis making the results less reliable. Contrarily to MI, it assumes linearity, normality and homoscedasticity, none of which is respected for all variables. MI can integrate all types of links which makes it particularly interesting. But MI suffers from lack of interpretability as explained in section [2.2.2](#).

These methods are principally useful to discover links between the variables and the outcome. However, if one wants to do feature selection, it is needed to be aware of the following problem. By keeping only the five variables with the highest score, one risks to keep features carrying the same information and discarding variables containing determinant data for the prediction. That problem can be solved by computing the MI between the variables like we did. In a future work, one could also use dimensionality reduction techniques.

When comparing the results, we observe that only one variable is common to the top five, the maximal current intensity between NP and SN. Water Stress which has the highest mutual information score with the outcome is insignificant for SN at 0.01 significance level according to ANOVA. This might be due to non-linear effects. The other variables in MI top five are statistically significant. Level Of Repression which is in ANOVA top five variables is in the top ten for MI, both for SN and NP. The same applies for Population Size in SN cases.

The two techniques give different and sometimes contrary results. On the one hand, we could think that one of them (or even both) is not reliable. On the other hand, we could imagine that it brings different information to light because of the different nature of both techniques, and that both should be considered with interest while keeping limitations in mind.

2.4 Conclusion

In this chapter, we gained insight into the dataset features and how they are related between them and to the outcome. First, we used analysis of variance which assumes linearity. For almost all samples, the test assumptions are not met which is a limitation. However, the majority of the variables are statistically significant, even at a 0.01 significance level. Future work could consist in considering multivariate ANOVA for checking pairwise statistical significance. In the second part of the chapter, we used mutual information to assess dependence between the outcome and the variables and pairwise dependence of input variables. It produced different results from ANOVA. The main advantage of this method is that it can handle non-linear relations. It is however difficult to estimate when dealing with continuous variables, which is a limitation. This chapter allowed us to get a better understanding of the interactions in the dataset and the importance of the variables. In Chapter [3](#), we use this insight to improve the classification models performances.

Chapter 3

Classification approach

In this chapter, we will present and apply machine learning technique to predict conflict outbreaks within the next four years. For the sake of clarify, we will recall some notions already introduced in Chapter 1. Machine learning is a sub-field of artificial intelligence. As explained in the introduction, we do supervised learning (which is a part of machine learning) : an algorithm takes data as input and processes it to output either a real number (prediction) or a class label (classification). In order to make prediction accurately and to determine the values of the parameters it is using, the algorithm is trained with data for which we know the desired output. It adapts its parameters in order to decrease the final error.

Example 1 (continued)

We go back to our small example in which we only dispose of two variables, Corruption and Unemployment. For the sake of simplicity, we only consider NP fights. Part of the fictitious dataset is given in Table 3.1.

	Country	Year	Corruption	Unemployment	Outcome (NP) y
1	Canada	2008	0.16	4.93	0
2	Yemen	2007	6.69	7.43	1
3	Angola	2001	8.69	5.3	1
4	Angola	2002	7.76	5.2	0
5	RDC	2002	8.06	5.73	1
...	...	...	...	...	...

Table 3.1: Toy dataset for conceptual example

If we wanted to do regression, we would seek to predict, for example, the number of casualties based on Corruption and Unemployment values. In this thesis, we want to perform classification, meaning assigning a class to each sample x_i . Here, the class is given by the outcome y_i . We want to find a function $\hat{f}$ such that $\hat{f}(0.16, 4.93) = 0$ and $\hat{f}(6.69, 7.43) = 1$.

We recall here some notions introduced in Chapter 1. Formally, we seek to find $\hat{f} : \mathbb{R}^p \rightarrow \{0, 1\} : x_i \rightarrow \hat{f}(x_i) = \hat{y}_i$ which provides the best estimate for y_i given x_i . y_i states whether instance i belongs to class $\mathcal{C}_0$ or $\mathcal{C}_1$ which are negative and positive classes, corresponding to peace and conflict, respectively. Back to the machine learning task, we find the parameters for $\hat{f}$ that minimises a measure of difference between y_i and $\hat{y}_i$. We will define several measures of difference in following sections and discuss them.

Most of the time, classification functions do not provide a binary output but a probability of belonging to one class. Since it is binary, we only consider the probability of belonging to the positive class $\mathcal{C}_1$, the conflict case. It is defined as

$$\hat{p} : \mathbb{R}^p \rightarrow [0, 1] : x_i \rightarrow \hat{p}(x_i) \quad (3.1)$$

From there, we derive $\hat{f}$ the following way, by fixing a discriminating threshold t :

$$\hat{f} = \begin{cases} 0 & \text{if } \hat{p}(x_i) < t \\ 1 & \text{if } \hat{p}(x_i) \geq t \end{cases} \quad (3.2)$$

In this chapter, we first explain how to assess the quality of $\hat{f}$ and $\hat{p}$. Then, we explain the methodology for deriving them and measuring systematically their performance. After that, we introduce two classification techniques, perform experiences and compare them.

3.1 Performance metrics: how to assess $\hat{f}$ and $\hat{p}$

Assessing $\hat{f}$ and $\hat{p}$ is part of the methodology. That is why we introduce performance metrics before deriving them.

3.1.1 The confusion matrix and basic metrics

Binary classification results are often represented in a confusion matrix which is explained in Table 3.2.

		What has been predicted	
		Conflict ($\hat{y}_i = 1$)	No conflict ($\hat{y}_i = 0$)
What has really happened	Conflict ($y_i = 1$)	True positive	False negative
	No conflict ($y_i = 0$)	False positive	True Negative

Table 3.2: Confusion matrix

Example 1 (continued)

For example, if $\hat{f}(x_1) = 1$, which is the prediction for the Canada in 2008, it is a false positive since the predictor forecasts a conflict when there is none.

We use the following abbreviations for conciseness.

Abbreviation	Category
TP	True positive
FP	False positive
TN	True negative
FN	False negative

Table 3.3: Abbreviations used for the classifier results

The number of instances in each category is characterised by the abbreviation preceded by a $\#$. When the context is clear, this symbol will be omitted.

When considering $\hat{p}$, many performance metrics are defined as a function of the discriminating threshold t .

- The true positive rate (also known as sensitivity or recall)

$$\text{TPR}(t) = \frac{\#TP(t)}{\#TP(t) + \#FN(t)} \quad (3.3)$$

where $\#TP(t)$ and $\#FN(t)$ correspond to the number of respectively true positives and false negatives for threshold t . This corresponds to the proportion of positive instances that were correctly classified. It is also understood as the proportion of relevant instances that are correctly identified.

- False positives rate (also known as fall out rate):

$$\text{FPR}(t) = \frac{\#FP(t)}{\#FP(t) + \#TN(t)} \quad (3.4)$$

This corresponds to the negative instances that were incorrectly classified. In other words, it corresponds to the proportion of irrelevant instances that were identified as relevant ones.

- Positive predictive value (also known as precision):

$$\text{PPV}(t) = \frac{\#TP(t)}{\#TP(t) + \#FP(t)} \quad (3.5)$$

This corresponds to the proportion of instances that were predicted as positive and

ROC AUC	
0.5	As good as tossing a coin
between 0.5 and 0.7	Poor score
between 0.7 and 0.8	Acceptable discrimination
between 0.8 and 0.9	Excellent discrimination
above 0.9	Outstanding discrimination

Table 3.4: Assessment of the model with ROC AUC score by Hosmer & Lemeshow [41]

that are actually positive or said otherwise: the proportion of identified cases that are relevant.

The above presented metrics do depend on the threshold t and do not give a global overview of the classifier since the threshold can be changed. This can be resolved by plotting curves of the metrics as a function of t or by taking more complex metrics.

It has been deliberately decided to avoid measuring the accuracy, $(\#TN(t) + \#TP(t))/n$, which corresponds to the percentage of correctly classified instances, either positive or negative. Accuracy is highly sensitive to class imbalance. If the prevalence (i.e. the fraction of positive instances) is low, classifying all the instances as negative gives a good accuracy score while the model is useless. As mentioned in section the dataset used in this thesis for conflict prediction is imbalanced as peace remains in most cases.

3.1.2 The ROC AUC

The ROC or Receiver Operating Characteristic curve shows the true positives rate versus the false positives rate by letting the discriminating threshold t vary. That gives the following parametric curve:

$$(x(t), y(t)) = (\text{FPR}(t), \text{TPR}(t)) \text{ with } 0 \leq t \leq 1 \quad (3.6)$$

It gives an idea of the classifier performance by showing the trade-off between classifying all cases as positives or all as negatives. As an illustration, two different ROC are drawn in Figures 3.1a and 3.1b. We directly see that the second curve is worse than the first one. To achieve the same true positive rate, the second model produces a higher false positive rate than the first one. The ideal classifier ROC curve includes point (0,1) such that there are no false positives and no false negatives. In order to quantify the goodness of a particular curve, a commonly used metric is the measure of the area under the curve (the ROC AUC), which can be obtained by integrating the TPR from 0 to 1 with respect to the FPR. Although there is no absolute rule, Hosmer & Lemeshow [41] provide indications about values of AUC that are reproduced in Table 3.4.

We must consider these values carefully since the requirement highly depends on the field of application. It might be more relevant to use it as a mean of comparison between different

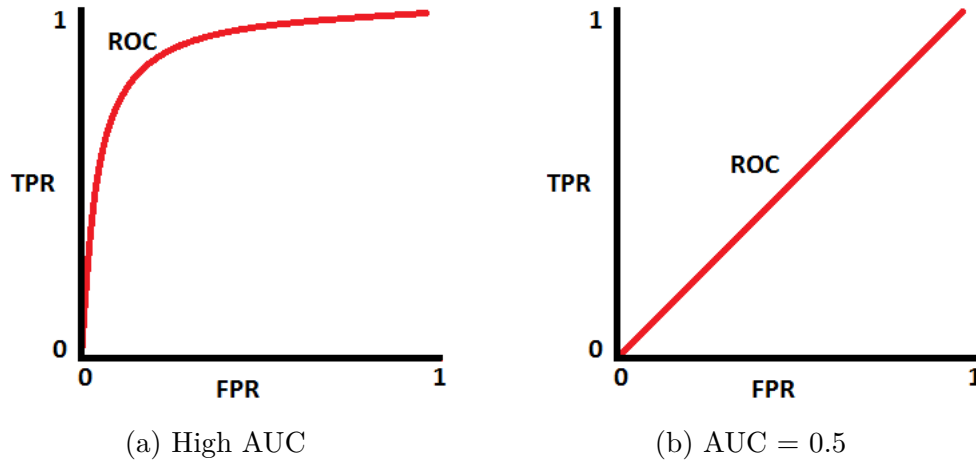


Figure 3.1: Illustration of ROC which is the trade-off between the TPR, the proportion of identified relevant instances and the FPR, the proportion of irrelevant instances that are classified as positives. In the Figure on the left, we observe that it is possible to achieve a higher TPR for a same FPR than in the Figure on the right. In the Figure on the right, it is the score obtained by tossing a coin for determining if a phenomenon will happen (if the phenomenon has a probability of 0.5 to happen).

models. Additionally, assessing a model is generally done with the help of several metrics. As an example, a good ROC AUC score could hide poor performances due to important class imbalance [47] and we will pay attention to this because of the low prevalence in the dataset that we use. However, we have to present this metric since it is used in one of the experiments that we reproduce here. We will frequently use the abbreviation AUC for this metric.

Example 1 (continued)

Back to our conceptual example, we imagine that there are two probability-assessing functions $\hat{p}_1$ and $\hat{p}_2$ which ROC curves are given in Figure 3.1. Looking at $\hat{p}_1$, if we set a sufficiently high value for t , it is possible to achieve a TPR, the proportion of identified relevant instances, of 0.9 with an FPR, which is the proportion of ill-classified negative instances, of 0.1 approximately. By setting a lower t , we catch more positive instances but the proportion of false positives increases. In terms of conflict predictions, it results in omitting fewer quarrels but having more peace situation that are declared as conflicts. The second function $\hat{p}_2$ performs worse since it gives an FPR of 0.9 for a TPR of 0.9. This means that there is a bigger proportion of false positives for the same proportions of identified relevant instances.

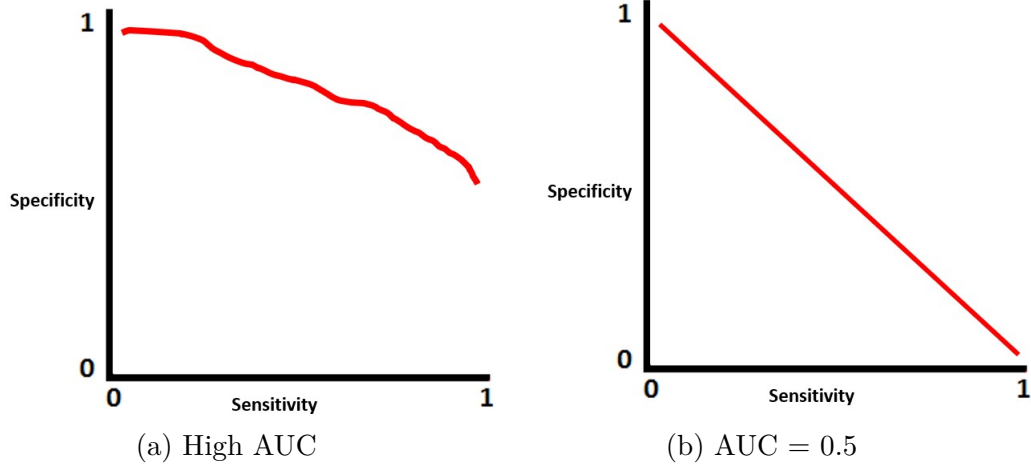


Figure 3.2: Illustrations of precision recall curves. These represent the trade-off between sensitivity, the proportion of relevant instances, and the specificity, the identified instances that are relevant. In the Figure on the left, we observe that it is possible, for the same sensitivity, to achieve a higher specificity than for the classifier depicted in the Figure on the right. Based on this criterion, the model for which the performance is given in the Figure on the left is a better classifier.

3.1.3 The precision-recall curve AUC

The curve presented here is defined by the following parametric curve:

$$(x(t), y(t)) = (\text{TPR}(t), \text{PPV}(t)) \text{ with } 0 \leq t < 1 \quad (3.7)$$

It shows the trade-off between the proportion of relevant instances that are identified and the proportion of identified cases that are relevant. Both the curve and the area under the curve can be used as metrics to assess the performance of the model. Note that $t = 1$ is not included, contrarily to the ROC curves since PPV (specificity) is not defined for $t = 1$. Illustrations of two precision recall curves are given in Figure 3.2. The left-one depicts a better classifier than the right one. We also observe that points $(0, 1)$ and $(1, 0)$ are not part of the curves. Indeed, it is impossible to have the specificity = 1 and the sensitivity = 0 or the opposite at the same time. As a matter of fact, they have the same numerator. It can seem that these points are part of the curve on some plots since it is possible to have a specificity of 1 with a low value for sensitivity, due to the high number of false negatives, especially for phenomena with low prevalence.

Example 1 (continued)

In Figure 3.2, we have a better probability-assessing function $\hat{p}_1$ represented on the left than the one on the right. If we want to increase the sensitivity (i.e. the proportion of identified relevant instances), the specificity decreases less for $\hat{p}_1$ than for $\hat{p}_2$.

Saito & Rehmsmeier explain in [77] that this metric is more adapted for imbalanced classification problems. Indeed, it does not use FPR where the true negatives in the denominator hide variation since it is a larger quantity than false positives (for a good classifier) in case of low prevalence.

3.1.4 The Brier score

Also called Brier inaccuracy, it is defined as:

$$\text{BS} = \frac{1}{n} \sum_{i=1}^n \left(\hat{p}_i - y_i \right)^2 \quad (3.8)$$

It is designed for probability assessing models. If applied on classification models working without probability assessing, it is equivalent to the mean squared error:

$$\text{BS} = \frac{1}{n} \sum_{i=1}^n \left(\hat{y}_i - y_i \right)^2 \quad (3.9)$$

Example 1 (continued)

To get the intuition, we compute it on the dataset introduced for the conceptual example given in table 3.5. For the below instances, it is equal to

$$\text{BS} = \frac{1}{5} \left((0.234-0)^2 + (0.95-1)^2 + (0.55-1)^2 + (0.54-0)^2 + (0.45-1)^2 \right) = 0.1708 \quad (3.10)$$

	Country	Year	Outcome (NP) y	$\hat{p}$
1	Canada	2008	0	0.234
2	Yemen	2007	1	0.95
3	Angola	2001	1	0.55
4	Angola	2002	0	0.54
5	RDC	2002	1	0.45
...	...	...	...	...

Table 3.5: Toy dataset and predicted probability $\hat{p}$ for conceptual example

It ranges from 0 (no single error) to 1 (all the instances are incorrectly classified). In practice, it never reaches 1. In this case, it would be enough to inverse the prediction in order to get a perfect classifier. We can compute a theoretical bound for Brier Score if the predicted probability is the prevalence in the dataset:

$$\rho = \frac{1}{n} \sum_{i=1}^n y_i \quad 1 \leq k \leq n \quad (3.11)$$

We note this quantity ρ for the sake of conciseness. We can note that this type of prediction

would be useless when considering a classification task. All instances would be classified as positive or negative depending on the threshold t . Here, we just derive a theoretical upper bound. This gives:

$$\text{BS} = \frac{1}{n} \sum_{i=1}^n (\rho - y_i)^2 = \rho(1 - \rho)^2 + (1 - \rho)\rho^2 = \rho - \rho^2 \quad (3.12)$$

For a balanced dataset, it would give 0.25. In our case, it yields 0.1608 and 0.2116 for NP and SN respectively. It means that, if the Brier score reaches this bound, it does not bring us more information than the dataset prevalence meaning that it is a non-informative model.

This metric does not deal well with class imbalance. In case of low prevalence, categorising all instances as negatives leads to low Brier Score while the classifier is useless. Brier score is not presented here because of its relevance for the general model but because it is used in the experiment that we reproduce, first realised by Halkia et al. [35].

3.1.5 Cohen's Kappa

As explained above, ROC AUC and Brier Score do not represent the real performances of the classifier for imbalanced dataset. Therefore, the only relevant metric is the precision-recall curve. One of its problem is that it does not give much importance to negative instances. And that is why we also consider the Cohen's kappa. It is defined as

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \quad (3.13)$$

where p_0 corresponds to the accuracy

$$p_0 = \frac{\text{TP} + \text{TN}}{n} \quad (3.14)$$

and

$$p_e = \left(\frac{\text{TP}(t) + \text{FP}(t)}{n} \right) \left(\frac{\text{TP}(t) + \text{FN}(t)}{n} \right) + \left(\frac{\text{FN}(t) + \text{TN}(t)}{n} \right) \left(\frac{\text{FP}(t) + \text{TN}(t)}{n} \right) \quad (3.15)$$

The intuition behind p_e is that it represents the probability of random agreement between reality and prediction. κ ranges from -1 to 1. It has been introduced by Cohen in [12] in order to assess the agreement between two classifications (e.g. two types of diagnosis in medicine). The idea here is to remove the agreement between the two classifications due to chance. It is particularly useful when dealing with imbalanced classes [47]. There is no agreement in the scientific community about the interpretation. However, Landis & Koch [52] propose one transcribed in Table 3.6 that we consider carefully. We mainly use this metric to compare models. The weakness of this metric is its dependence to one particular threshold. For this reason, we take the maximum value over all t . Another drawback is its difficulty of interpretation.

Range of κ	Strength of agreement
$\kappa < 0$	Poor
$0 \leq \kappa \leq 0.2$	Slight
$0.2 < \kappa \leq 0.4$	Fair
$0.4 < \kappa \leq 0.6$	Moderate
$0.6 < \kappa \leq 0.8$	Substantial
$0.8 < \kappa \leq 1$	Almost perfect

Table 3.6: One interpretation of Cohen’s κ by Koch & Landis [52]

3.1.6 Final choice of metrics

One of our goals is to compare our models to the experiment conducted by Halkia et al. in [35]. For this purpose we use ROC AUC and the Brier score. In addition, other models are tested and compared. As mentioned before, ROC AUC and Brier score do not represent the real performances well when used on imbalanced datasets. For this reason, we mainly use precision-recall curve and its AUC in addition to the Cohen’s kappa for the different algorithms in order to perform our final analysis. The interest of combining both is the following: Cohen’s kappa is difficult to interpret while PR curve is easier. On the other hand, PR curve does not give importance to the negative class while Cohen’s kappa is a more general measure. We also provide the confusion matrix for $t = 0, 5$ since it is the most readable performance assessment.

3.2 Methodology

3.2.1 Variables consideration

All variables (even though already rescaled from 0 to 10) have been centred (removing of the mean) and divided by their standard deviation since this can improve model performances.

Reducing the input dimension p can also be a way to improve prediction performances. Typically, considering too many variables can lead to overfitting. This is a phenomenon where the model fits the data it has been trained with so well that it performs poorly on unseen samples. For some tests we discard variables that are insignificant according to ANOVA tests.

Example 1 (continued)

For example, it is possible that, at a 0.01 significance level, Unemployment is not statistically significant for the toy dataset. We would discard this variable and make predictions only based on Corruption.

For other experiments, we tune hyperparameters to avoid this phenomenon. Hyperparameters are specifications that control the learning process while parameters are optimised during the learning.

3.2.2 Procedure

Before anything else, we need to adapt the learning loss function in order to take class imbalance into consideration. The term that accounts for error of sample point i is multiplied by $\frac{n}{2n_1}$ for positive instances and $\frac{n}{2n_0}$ for negative instances with n being the number of samples considered, and n_1 and n_0 , the amount of respectively positive and negative instances in the considered set.

In order to use the whole dataset alternatively as train and test set to avoid performance measures depending too much on the data points included in the test set, we use k -fold cross validation. The use of k -fold validation consists in dividing the dataset into k equal parts and repeating the process k times. For each repetition, the part containing $(k-1)/k$ proportion of the instances is used for training while the other part is used for assessing the performances. There is an illustration of this process in Figure 3.3. It also allows us to check for model overfitting by comparing the performances on the training and the test set. It is advised to choose $k = 10$ cross validation for real-world data with many predictors [3]. The distinction between test and train set is crucial. It allows us to observe the model performances on unseen data.

Example 1 (continued)

If we apply k -fold validation on the conceptual example, we train five models, each on four samples, and we keep the last one for assessing the model. For the first fold, we would train the classifier on instance 1,2,3 and 5 and test it on RDC 2002 situation.

In order to test our model with several configurations and derive robust model characteristics, we repeat the experience 10 times and shuffle the data for each round. This changes the compositions of training and test. To avoid confusion, we call this k -fold validation the "external" cross validation. We will explain the "internal" cross validation in the next paragraph.

The hyper-parameters (as a reminder, these are the parameters that control the learning process) are chosen among several candidates using grid-optimisation and internal 10-fold cross validation. That happens inside one external validation fold. The objective function is a determined criterion (e.g. ROC AUC or Brier score). To get an illustration of this process, a small conceptual explanation of the grid search is provided here. We consider q hyper-parameters: $\alpha_1, \alpha_2, \dots, \alpha_q$. The set A_i contains the values that have to be tested for α_i (it can be numeric, categorical, abstract value like 'Euclidean' or 'Cosine' for distance measure). The grid search algorithm fits the model for each combination of hyper-parameters

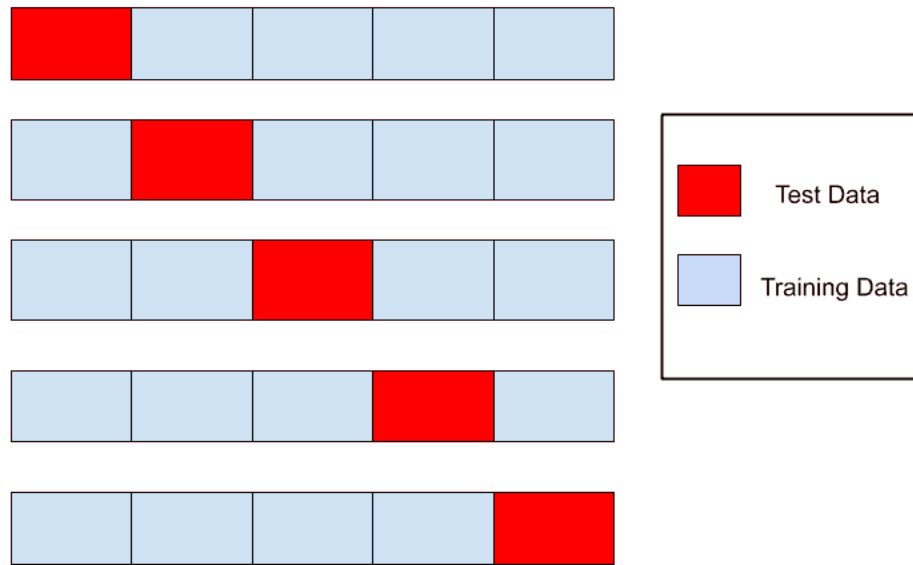


Figure 3.3: 5-fold test scheme with each line representing an independent run (coming from [59]). Each row represents a train and test phase for the model such that there are 5 different models. The blue boxes correspond to the part of samples used for the training phase and the red boxes correspond to the part of the samples used for the test phase.

$(\alpha_1, \alpha_2, \dots, \alpha_q) \in A_1 \times A_2 \times \dots \times A_q$ and save the chosen value of the objective function. At the end of the process, it outputs the best set of hyper-parameters to be kept for the model fit. Note that here, it performs 10-fold internal cross validation for each point in the grid. The performances are obtained by taking the mean of the performances for each test.

After, it fits a model on the train set by keeping the best hyperparameters. Since we evaluate 100 models (the number of external cross folds times the number of repetitions), we report the median scores, curves and their interquartile ranges. We could not take the mean curves since their distribution are skewed for some of them.

The whole methodology is summarised in Algorithm 2.

Algorithm 2: General methodology

```

1 Determine candidate values for hyperparameters
2 Chose a performance metric for the grid search
3 for  $iSeed = 1 : 10$  do
4     shuffle  $\mathbf{X}$  rows and  $\mathbf{y}$  the same way with seed = Iseed (for reproducibility)
5     divide  $\mathbf{X}$  and  $\mathbf{y}$  in 10 folds
6     for  $iExtFold = 1 : 10$  do
7          $X^{test}$  and  $y^{test}$  are composed by the  $iExtFold^{th}$  fold
8          $X^{train}$  and  $y^{train}$  are composed by the other folds
9         for  $j = 1 : p$  do
10              $\mu_j \leftarrow \text{mean}(X_j^{train})$ 
11              $\sigma_j \leftarrow ((\text{Var}[X_j^{train}]))^{1/2}$ 
12             remove  $\mu_j$  from each element of  $X_j^{train}$  and  $X_j^{test}$  and divide the result by  $\sigma_j$ 
13         end
14         divide  $X^{train}$  and  $y^{train}$  in 10 folds
15         for each combination of hyperparameters do
16             for  $iIntFold = 1 : 10$  do
17                 fit the model on all the internal folds except the  $iIntFold^{th}$  one
18                 predict  $\hat{p}$  and  $\hat{y}$  for the  $iIntFold^{th}$  fold
19                 compute and store performance score for the 10th fold prediction
20             end
21             compute and store mean performance score over the all the internal folds
22         end
23         Keep the subset of parameters with highest mean performance score
24         fit the model on  $X^{train}$ 
25         predict  $\hat{p}$  and  $\hat{y}$  for the test set
26         Compute the ROC and PR curve, ROC and PRC AUC, Brier Score, Cohen's
            Kappa and confusion matrix (for  $t = 0.5$ ) for the test and train set predictions
            separately, store.
27     end
28 end
29 Compute the median scores and curves and their interquartile ranges.

```

3.3 Logistic regression

In this section, we introduce a first machine learning algorithm used for the classification task.

3.3.1 Logistic regression: theory

Recall that we want to find a function $\hat{f}$ to estimate whether instance i belongs to class $\mathcal{C}_0$ or class $\mathcal{C}_1$. The proposed method is the following: it guesses a probability of being a positive instance by fitting a sigmoid function (see next equation) to data points.

$$p_i = \frac{e^{\gamma_i}}{1 + e^{\gamma_i}} \quad 1 \leq i \leq n \quad (3.16)$$

p_i being the estimated probability that data point i belongs to $\mathcal{C}_1$ and where

$$\gamma_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} = \beta_0 + \boldsymbol{\beta}^T x_i \quad 1 \leq i \leq n \quad (3.17)$$

Example 1 (continued)

In Figure 3.4, we have the sigmoid function if we only consider Corruption in the conceptual example dataset. If we consider both variables, it would give a function of this form :

$$\hat{p}(x_i) = \frac{1}{1 + e^{\beta_0 + \beta_1 \text{Corruption} + \beta_2 \text{Unemployment}}} \quad (3.18)$$

The coefficients are estimated using the maximum likelihood method [63]. We can rewrite equation 3.16 as

$$p_i = \frac{1}{1 + e^{-\gamma_i}} = \frac{1}{1 + e^{-\beta_0 - \boldsymbol{\beta}^T x_i}} \quad (3.19)$$

It is renamed as $h_{\beta_0, \boldsymbol{\beta}}(x_i)$ We assume that

$$P(y_i = 1 | x_i, \beta_0, \boldsymbol{\beta}) = h_{\beta_0, \boldsymbol{\beta}}(x_i) \quad (3.20)$$

$$P(y_i = 0 | x_i, \beta_0, \boldsymbol{\beta}) = 1 - h_{\beta_0, \boldsymbol{\beta}}(x_i) \quad (3.21)$$

And this can be rewritten as

$$p(y_i | x, \beta_0, \boldsymbol{\beta}) = (h_{\beta_0, \boldsymbol{\beta}}(x_i))^{y_i} (1 - h_{\beta_0, \boldsymbol{\beta}}(x_i))^{1-y_i} \quad (3.22)$$

With n data points, generated independently¹, we write the likelihood of the parameters as

$$L(\beta_0, \boldsymbol{\beta}) = p(\mathbf{y} | \mathbf{X}, \beta_0, \boldsymbol{\beta}) \quad (3.23)$$

$$= \prod_{i=1}^n p(y_i | x_i, \beta_0, \boldsymbol{\beta}) \quad (3.24)$$

$$= \prod_{i=1}^n (h_{\beta_0, \boldsymbol{\beta}}(x_i))^{y_i} (1 - h_{\beta_0, \boldsymbol{\beta}}(x_i))^{1-y_i} \quad (3.25)$$

¹The samples' independence has already been discussed in section 2.1.1, it is impossible to guarantee it.

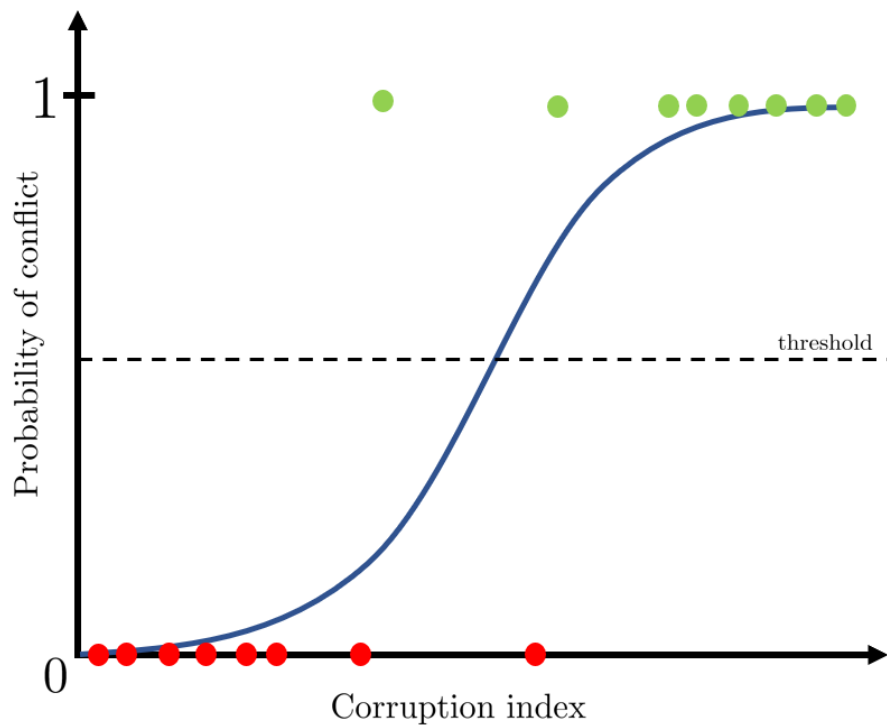


Figure 3.4: Illustration of a one-dimensional logistic regression (the green and red dots are positive and negative instances respectively). On the ordinate, we have the assessed probability. On the abscissa, we have the feature value. An instance is classified as positive if the assessed probability is greater or equal to the threshold t which is represented by the dashed line. Otherwise, it is classified as negative.

By taking the log likelihood, we have:

$$\ell(\beta_0, \boldsymbol{\beta}) = \log(L(\beta_0, \boldsymbol{\beta})) \quad (3.26)$$

$$= \sum_{i=1}^n \left(y_i \log(h_{\beta_0, \boldsymbol{\beta}}(x_i)) + (1 - y_i) \log(1 - h_{\beta_0, \boldsymbol{\beta}}(x_i)) \right) \quad (3.27)$$

$$= \sum_{i=1}^n \left(y_i \log\left(\frac{1}{1 + e^{-\beta_0 - \boldsymbol{\beta}^T x_i}}\right) + (1 - y_i) \log\left(1 - \frac{1}{1 + e^{-\beta_0 - \boldsymbol{\beta}^T x_i}}\right) \right) \quad (3.28)$$

The following optimisation is then derived:

$$\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^n \left(y_i \log(1 + e^{-\beta_0 - \boldsymbol{\beta}^T x_i}) + (1 - y_i) \log(1 + e^{\beta_0 + \boldsymbol{\beta}^T x_i}) \right) \quad (3.29)$$

This is the basic form of the problem. However, this formulation can lead to several problems. One of them is overfitting. In addition, if the dataset is linearly separable, then the optimisation process will not converge [1]. The objective function can be as close as wanted to zero by choosing β sufficiently large since the maximum likelihood estimator does not exist [1].

Linear separability implies that the dataset can be separated with a hyperplane, having the positive instances on one side of the hyperplane and the negative instances on the other side. To formulate this mathematically, we need a variable z_i such that

$$z_i = \begin{cases} -1 & \text{if } y_i = 0 \\ 1 & \text{if } y_i = 1 \end{cases} \quad (3.30)$$

The dataset is linearly separable if there exists a vector $\mathbf{a} \in \mathbb{R}^p$ and a scalar b such that

$$z_i(\mathbf{a}^T x_i + b) \geq 1 \quad 1 \leq i \leq n \quad (3.31)$$

We explain here why the optimisation problem has no solution for the one-dimensional case. Suppose without loss of generality that $y_i = 0$ for $x_i < 0$ and $y_i = 1$ for $x_i > 0$ and that there are no data points at $x = 0$. The data is centred round $x = 0$ such that we can neglect the intercept for simplicity of notation. Let us consider the following limits:

$$\lim_{\beta \rightarrow \infty} h_{\beta}(x_i) = \lim_{\beta \rightarrow \infty} \frac{1}{1 + e^{-\beta x_i}} = 1 \text{ for } x_i > 0 \quad (3.32)$$

$$\lim_{\beta \rightarrow \infty} h_{\beta}(x_i) = \lim_{\beta \rightarrow \infty} \frac{1}{1 + e^{-\beta x_i}} = 0 \text{ for } x_i < 0 \quad (3.33)$$

Taking the objective function again, that yields:

$$\min_{\beta} \sum_{i=1}^n \left(y_i \log(h_{\beta}(x_i)) + (1 - y_i) \log(1 - h_{\beta}(x_i)) \right) \quad (3.34)$$

For both parts of the sum, the term in the log tends to one and the overall result is zero. No particular β is chosen as a solution. That is why regularisation is needed in case of linear

separability.

The regularised problem is defined as:

$$\min_{\beta_0, \beta} F(\beta) + C \sum_{i=1}^n \left(y_i \log(1 + e^{-\beta_0 - \beta^T x_i}) + (1 - y_i) \log(1 + e^{\beta_0 + \beta^T x_i}) \right) \quad (3.35)$$

C is called the regularisation parameter and allows us to weigh the main term. Low values for C gives a strongly-regularised model and conversely.

Here, we present the most classical regularisation functions for logistic regression: ℓ_1 and ℓ_2 . Both are based on the ℓ_1 - and ℓ_2 -norm of the weights, respectively. This gives:

$$F_{\ell_1}(\beta) = \|\beta\|_1 = \sum_{j=1}^p |\beta_j| \quad (3.36)$$

ℓ_1 -regularisation minimises the overall sum of the coefficients (in absolute value). It is also an approximation for ℓ_0 [74] which is defined as:

$$\|v\|_0 = |\{k \text{ s.t. } |v_k| > 0\}| \quad (3.37)$$

Regularisation with ℓ_0 -norm² is equivalent to the reformulation of the problem by integrating features selection. However, this adds a discontinuous term in the optimisation problem since for $v_k = 0$, the contribution of the k^{th} variable to the norm is 0 but for all $|v_k| > 0$, the contribution of the k^{th} variable to the norm is 1. ℓ_1 serves as an approximation for ℓ_0 because it is continuous and even convex, which is a good advantage for the optimisation. This type of regularisation allows us to avoid overfitting since only the most important features are considered (the other ones have low weights). It is its main advantage. The ℓ_2 -regularisation is defined as follows:

$$F_{\ell_2}(\beta) = \frac{1}{2} \|\beta\|_2^2 = \frac{1}{2} \beta^T \beta \quad (3.38)$$

It penalises coefficients with significantly greater values than others while minimising the sum of the square of all coefficients. The main goal of ℓ_2 -regularisation is to guarantee the existence of a solution. Indeed, as explained previously, if the dataset is linearly separable, then the optimisation problem has no solution.

During the training phase, in the internal CV, the best regularisation-parameter C in equation 3.35 is found by grid-search optimisation.

Until here, only probabilities have been assessed, no classification has been done yet. That

²It is called a norm by abuse but it is not since it does not have homogeneity-property stating that $\|\alpha x\| = \alpha \|x\|$ for any $x \in X$ and α being a scalar.

is why we have to choose a probability threshold and derive $\hat{f}$ from $\hat{p}$ by using the previously presented equation :

$$\hat{y}_i = \begin{cases} 0 & \text{if } \hat{p}(x_i) < t \\ 1 & \text{if } \hat{p}(x_i) \geq t \end{cases} \quad (3.39)$$

To get the intuition, one can take a look at Figure 3.4. The first red dot is predicted as negative since the sigmoid function value is less than the threshold for this abscissa. But the red point with the highest x value (the rightmost one) will be predicted as positive since the value of the sigmoid function for this abscissa is above the threshold.

Still, the delicate question of choosing the threshold remains. The choice has to be made by the final user of the algorithm by determining a trade-off between false positives and false negatives.

The complete procedure for applying logistic regression is detailed in algorithm 3.

Algorithm 3: Logistic regression

```

1 Choose  $t$ 
2 choose regularisation
3 determine  $C$ 
4 if  $\ell_1$ -regularisation then
5   | solve  $\min_{\beta_0, \beta} \sum_{j=1}^p |\beta_j| + C \sum_{i=1}^n \left( y_i \log(1 + e^{-\beta_0 - \beta^T x_i}) + (1 - y_i) \log(1 + e^{\beta_0 + \beta^T x_i}) \right)$ 
6 else if  $\ell_2$ -regularisation then
7   | solve  $\min_{\beta_0, \beta} \frac{1}{2} \beta^T \beta + C \sum_{i=1}^n \left( y_i \log(1 + e^{-\beta_0 - \beta^T x_i}) + (1 - y_i) \log(1 + e^{\beta_0 + \beta^T x_i}) \right)$ 
8 else
9   | solve  $\min_{\beta_0, \beta} \sum_{i=1}^n \left( y_i \log(1 + e^{-\beta_0 - \beta^T x_i}) + (1 - y_i) \log(1 + e^{\beta_0 + \beta^T x_i}) \right)$ 
10 store  $\beta_0$  and  $\beta$ 
11 for  $i = 1 : n$  do
12   |  $\hat{p}_i \leftarrow 1 / (1 + e^{-\beta_0 + \beta \cdot x_i})$ 
13   | if  $\hat{p}_i < t$  then
14     |  $\hat{y}_i \leftarrow 0$ 
15   | else
16     |  $\hat{y}_i \leftarrow 1$ 
17   | end
18   | store  $\hat{y}_i$  and  $\hat{p}_i$ 
19 end

```

3.3.2 Strength and weaknesses of logistic regression

Strengths

- The model's weights are easy to interpret
- It is easy to implement. Actually, it is available in many machine learning toolboxes. Furthermore, it is easy to understand and apply. Except for C , no hyperparameter has to be tuned.
- The data distribution does not have to be normal [40] This is the case for many statistical processes, but it does not have an impact on the logistic regression performances.

Weaknesses

- For the model to perform well, the dataset needs to be linearly separable (and then, regularisation is needed) in the sense of equation 3.31. This criterion is often not met. Having a majority of points satisfying equation 3.31 already leads to good performances.
- It requires independence of the sample points because of the maximum likelihood method.
- Correlated variables in the model lead to lower performances [69].
- It is unable to capture complex relations since it requires a linear relation between the independent variables and the log of the odds ($\log(\frac{p_i}{1-p_i}) = \beta_0 + \beta^T x_i$). As a consequence, it might be unable to produce reliable predictions for complex phenomena as conflicts outbreaks.
- It is unable to capture feature interactions except if the user encodes a combination of the variables, for example the product of them. However, this has two limitations : the first one is that a same value for the product can originate from different pairs of features' values depicting distinct situations. The second is that taking all pairwise interactions in account generates $\binom{23}{2} = 273$ additional variables. And this still does not take three or more variables interactions into account.

3.3.3 Analysis of the appropriateness of logistic regression for our analysis

The interpretability of the model and its straightforwardness make this model a good starting point. However, independence of the sample points is impossible to guarantee since different years for the same country are taken. And we can reasonably suppose that neighbouring countries might have correlated evolution of several indicators. Neither can we assume that variables for a same data point are uncorrelated. We cannot reasonably state that oil exportation is uncorrelated with regime type. And last, the inability to capture complex relations adds doubt about the usefulness of this model.

When weighing the advantages and the disadvantages, the model does not seem completely adapted to the problem. However, it has been used in [35] which we reproduce, and it is also a first straightforward technique mainly used in a preliminary analysis to gain insight into the problem.

3.3.4 Results obtained with logistic regression

Several tests are conducted, always following the above described procedure but with small changes. At the end, we select the most interesting model and discuss the relevance of the coefficients for each variable. For the hyper-parameter grid-search, the optimisation criterion is balanced accuracy which is defined as

$$\frac{TP/n_1 + TN/n_0}{n} \quad (3.40)$$

with TP, the number of true positives, TN, the number of true negatives, n_1 , the number of positive instances, n_0 , the number of negative instances and n , the total number of instances. It allows us to take class imbalance into account. This metric has been chosen as objective function for determining C , the regularisation parameter, since it gives the best compromise between all the metrics considered in this problem, as far as we could observe during the experiments. The median results for each classifier are given in Table 3.7.

Experiment 1: insignificant variables dropped and ℓ_2 -regularisation

The first test consists in removing the insignificant variables (determined by the ANOVA test in subsection 2.1.2) and applying a ℓ_2 -regularisation. We take a level of significance of 0.01 since we want to discard less important variables, making the new variables set different from the initial one. The removed variables are Income Inequality (NP and SN), Unemployment Rate (NP and SN), Lack Of Democracy (SN) and Water Stress (SN). Since we already removed some variables, we do not use ℓ_1 -regularisation which works as feature selection. The ℓ_2 -regularisation is applied in order to ensure solution existence of the algorithm since the possible linear separability of the dataset would compromise this. We have the following minimisation problem

$$\min_{\beta_0, \beta} \frac{1}{2} \beta^T \beta + C \sum_{i=1}^n \left(y_i \log(1 + e^{-\beta_0 - \beta^T x_i}) + (1 - y_i) \log(1 + e^{\beta_0 + \beta^T x_i}) \right) \quad (3.41)$$

The value for C is chosen in the following set:

$$\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100, 10^3, 10^4, 10^5\} \quad (3.42)$$

The log scale allows us to cover many orders of magnitude without testing too many cases. The optimal value for C is chosen using the internal cross validation and can vary depending on the external fold. In order to solve the minimisation problem described in equation 3.41, the machine learning toolbox uses LBFGS-solver (see [8] for more details about this method).

The median precision-recall curves are given for both classifiers in Figures 3.5a and 3.6a.

Since we had about 100 curves for each classifier because of the 10 repetitions of the procedure, each comporting 10 external folds, we took the median curve. For the test sets only, the interquartile range is also plotted. The numerical scores are available in Table 3.7. The curve and score distributions are skewed and that is why we do not report the mean and standard deviation.

The model is not subject to overfitting since the train curve is near the test curve. This is also reflected by the scores that are almost the same even if they are slightly better for train sets. For example, the median PRC AUC are 0.742 for the test set and 0.748 for the train set for NP case. For SN cases, the score is 0.82 for both curves. The median max kappa score for NP cases is 0.603, both for train and test sets. For the SN cases, it is 0.632 and 0.649 respectively. These four kappa scores are qualified as substantial by Koch & Landis [52]. In Figure 3.5a and 3.6a, we observe that for a higher sensitivity, there is a smaller variation in the specificity of the models. The SN cases show less variation than the NP ones for the PRC. As expected, the area under the train curves has a lower interquartile range than for the area under the test curves with values of respectively 0.028 and 0.064 for the PRC, and 0.015 and 0.029 for the ROC curve for NP cases. For SN cases, it is respectively 0.006 and 0.044 for the PRC and 0.003 and 0.029 for the ROC curve. The interquartile ranges are equal or lower for SN cases. We also observe that the PRC AUC vary more than the ROC AUC. We can suppose that it is due to the consideration of the true negatives in the FPR. Since it is a bigger quantity, it hides other terms variations.

The ROC AUC (test sets) are 0.909 and 0.9 for the NP and SN cases respectively. It is an outstanding discrimination according to Hosmer & Lemeshow [41] (see Table 3.4). The median Brier scores for test sets are 0.118 and 0.125. These do not reach the bounds of a non-informative model derived in section 3.1.4 of 0.1608 and 0.2116 for NP and SN respectively. This means that we learn more with the model results than with the dataset prevalence.

The PRC is better for the SN case. It enables a higher specificity for the same sensitivity and this is reflected by the difference in the area under the curve which show important difference with 0.742 and 0.82, respectively for the test sets.

At last, we can take a look at the mean confusion matrix of the test sets. Around 83% of the instances are correctly classified. But comparing the false positives and the false negatives would not make any sense since they depend on the discriminating threshold t . A high value of t leads to many false negatives and few false positives while a small value leads to the opposite. The percentages have been rounded for readability, and they might not sum up to 100. Out of 295 cases, 245 are correctly identified. Knowing that the GCRI monitors 138 countries, it would correspond to 23 ill- and 115 correctly-classified instances. It is possible that the error rate is smaller: we could have a situation where, in the 294 countries-years of

the test sets, the same country appears several times for different years in the errors due to inherent particularity of the situation.

Experiment 2: insignificant variables dropped, no regularisation

The second test is the same as the first one but no regularisation is applied. Therefore, no value for C has to be chosen since we want to solve the following optimisation problem:

$$\min_{\beta_0, \beta} \sum_{i=1}^n \left(y_i \log(1 + e^{-\beta_0 - \beta^T x_i}) + (1 - y_i) \log(1 + e^{\beta_0 + \beta^T x_i}) \right) \quad (3.43)$$

Note that no internal fold is needed here since no regularisation parameter has to be chosen. The results obtained are given in Figures 3.5b and 3.6d. We see that the data is not linearly separable in the sense of equation 3.31 since the optimisation problem has a solution. Globally, we make the same observations as in the previous experiment. The difference is that scores are slightly better as can be seen in Table 3.7. We can reasonably think that this is due to no-regularisation which makes the optimisation problem "less constrained". Additionally, we note a bending in the precision-recall curve for precision values between 0.2 and 0.6 for the NP cases, which depicts lower performance.

Experiment 3: all variables, ℓ_1 -regularisation

For the third experience, all the variables are kept for the model but the ℓ_1 -regularisation is applied. It gives the following minimisation problem:

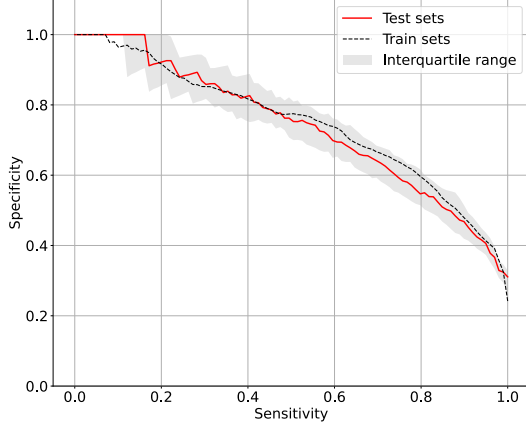
$$\min_{\beta_0, \beta} \sum_{k=1}^p |\beta_k| + C \sum_{i=1}^n \left(y_i \log(1 + e^{-\beta_0 - \beta^T x_i}) + (1 - y_i) \log(1 + e^{\beta_0 + \beta^T x_i}) \right) \quad (3.44)$$

In the toolbox used, another solver is applied when dealing with ℓ_1 -regularisation: it is based on coordinate descent and comes from the LIBLINEAR library which is built specifically for linear classification (see [94] for more details on coordinate descent). Dealing with linear classification does not imply that the objective function to minimise. Typically, in our case, the objective function is non-linear as can be observed in equation 3.44. The proposed values for C are the same as for experiment one. Results are depicted in Figures 3.5c and 3.6c and also in Table 3.8. The other results are similar to the ones we made for the previous experiments.

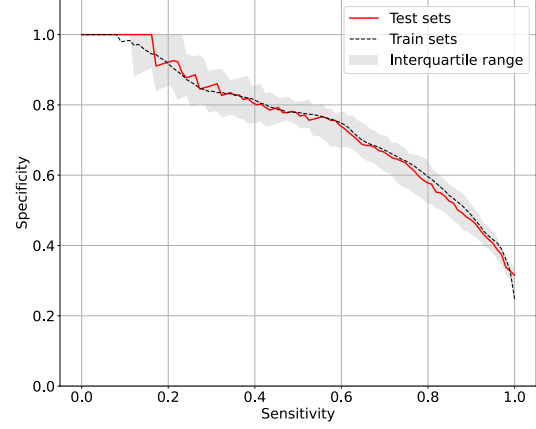
Experiment 4: all variables, no regularisation

No regularisation is applied and all the variables are kept. The solver used here is again LBFGS. The objective function is:

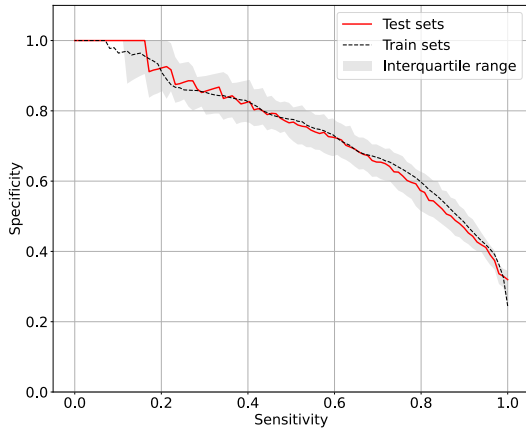
$$\min_{\beta_0, \beta} \sum_{i=1}^n \left(y_i \log(1 + e^{-\beta_0 - \beta^T x_i}) + (1 - y_i) \log(1 + e^{\beta_0 + \beta^T x_i}) \right) \quad (3.45)$$



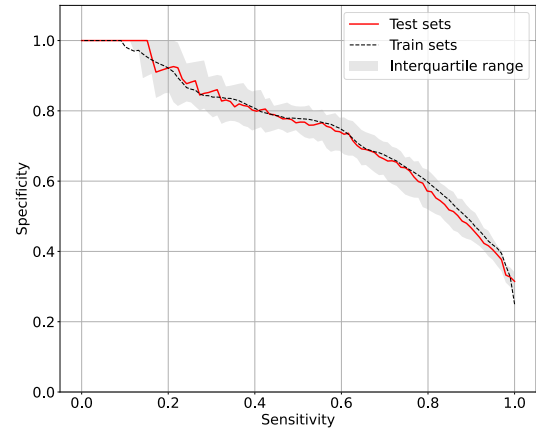
(a) Experiment 1: insignificant variables dropped, ℓ_2 -regularisation applied



(b) Experiment 2: insignificant variables dropped, no regularisation applied

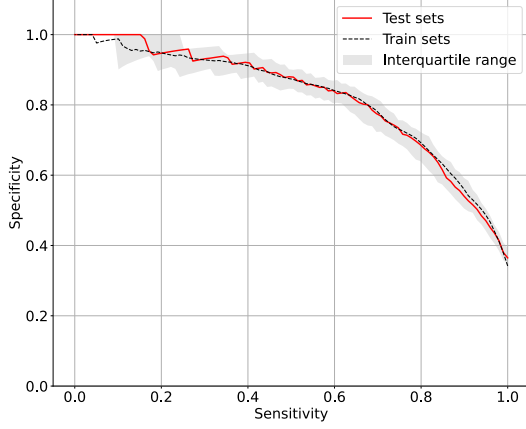


(c) Experiment 3: all variables kept, ℓ_1 -regularisation applied

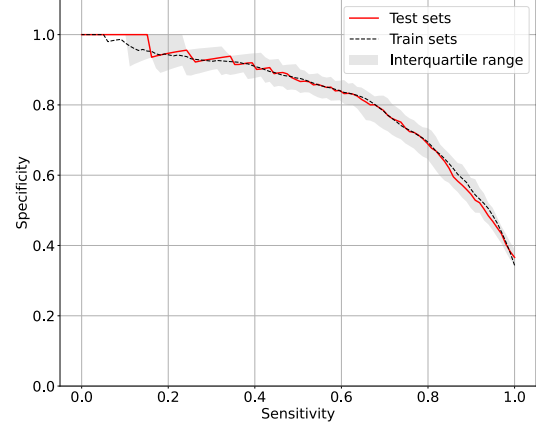


(d) Experiment 4: all variables kept, no regularisation applied

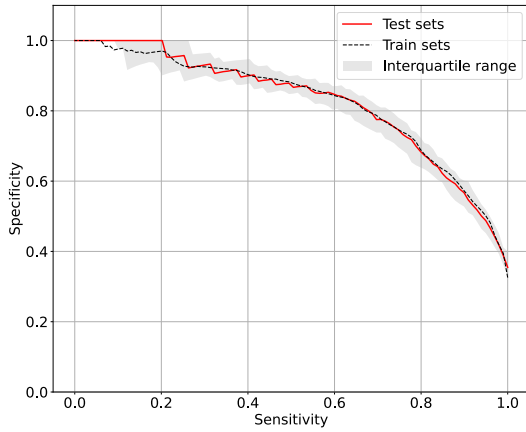
Figure 3.5: Median precision-recall curves of the logistic regression classifier, NP case. The specificity is given on the ordinate-axis and the sensitivity on the abscissa. The red curve shows the median performance on the test sets and the dashed black curve shows the median performance for the train sets. The shaded area corresponds to the interquartile range of the precision-recall curves on the test sets.



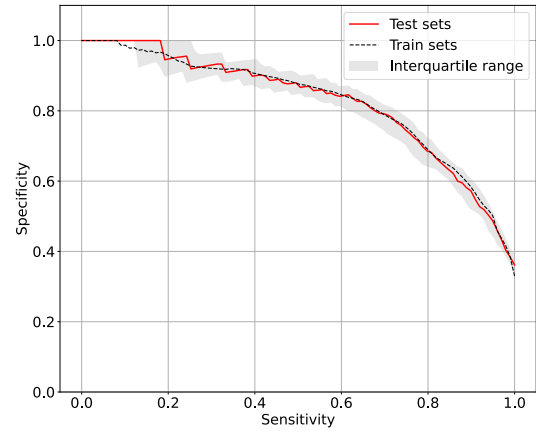
(a) Experiment 1: insignificant variables dropped, ℓ_2 -regularisation applied



(b) Experiment 2: insignificant variables dropped, no regularisation applied



(c) Experiment 3: all variables kept, ℓ_1 -regularisation applied



(d) Experiment 4: all variables kept, no regularisation applied

Figure 3.6: Precision-recall curves of the logistic regression classifier, SN case. The specificity is given on the ordinate-axis and the sensitivity on the abscissa. The red curve shows the median performance on the test sets and the dashed black curve shows the median performance for the train sets. The shaded area corresponds to the interquartile range of the precision-recall curves on the test sets.

	Experiment 1				Experiment 2				Experiment 3				Experiment 4			
	Insignificant variables dropped				Insignificant variables dropped				All variables				All variables			
	and ℓ_2 -regularisation				and no regularisation				ℓ_1 -regularisation				No regularisation			
	NP		SN		NP		SN		NP		SN		NP		SN	
Median PRC AUC – test sets	0.742 (0.064)		0.82 (0.044)		0.748 (0.06)		0.819 (0.046)		0.745 (0.061)		0.823 (0.041)		0.748 (0.06)		0.822 (0.043)	
Median PRC AUC – train sets	0.748 (0.028)		0.82 (0.006)		0.75 (0.006)		0.82 (0.005)		0.748 (0.007)		0.825 (0.005)		0.75 (0.006)		0.826 (0.005)	
Median max kappa – test sets	0.603 (0.064)		0.649 (0.055)		0.621 (0.07)		0.65 (0.061)		0.616 (0.06)		0.652 (0.061)		0.622 (0.059)		0.66 (0.06)	
Median max kappa – train sets	0.603 (0.043)		0.632 (0.008)		0.606 (0.01)		0.633 (0.007)		0.605 (0.009)		0.638 (0.008)		0.608 (0.01)		0.644 (0.008)	
Median ROC AUC – test sets	0.909 (0.029)		0.9 (0.029)		0.914 (0.026)		0.9 (0.028)		0.913 (0.026)		0.903 (0.025)		0.912 (0.026)		0.902 (0.027)	
Median ROC AUC – train sets	0.916 (0.015)		0.902 (0.003)		0.917 (0.003)		0.902 (0.003)		0.916 (0.003)		0.905 (0.003)		0.917 (0.003)		0.905 (0.003)	
Median Brier score – test sets	0.118 (0.024)		0.125 (0.017)		0.12 (0.016)		0.123 (0.015)		0.121 (0.016)		0.123 (0.013)		0.121 (0.016)		0.122 (0.014)	
Median Brier-score – train sets	0.127 (0.023)		0.123 (0.002)		0.117 (0.002)		0.123 (0.002)		0.12 (0.002)		0.122 (0.002)		0.118 (0.002)		0.121 (0.002)	
Median confusion matrix for discriminating threshold $t = 0.5$ - test sets	48 (16.3%)	11.5 (3.9%)	70 (23.8%)	20 (6.8%)	48 (16.4%)	11 (3.7%)	70 (23.9)	19 (6.5%)	48 (16.4%)	11 (3.7%)	71 (24.1%)	19 (6.4%)	48 (16.4%)	11 (3.8%)	71 (24.1%)	18.5 (6.3%)
	38 (12.9%)	197 (66.9%)	29 (9.9%)	175 (59.5%)	38 (12.9%)	196.5 (67%)	29 (9.9%)	175 (59.7%)	38 (12.9%)	197 (67%)	31 (10.6%)	173 (58.8%)	38 (13%)	196 (66.9%)	32 (10.9%)	173 (58.7%)

Table 3.7: Median performance metrics for the logistic regression classifier for each experiment (the numbers in brackets are the interquartile ranges except for the confusion matrix where it corresponds to the percentages). The best scores are highlighted in bold writing.

The results are depicted in Figures 3.5d and 3.6d but also in Table 3.7. Our observations are similar to the ones we made for the previous experiments. Additionally, one can note a bending in the PRC for sensitivity values between 0.2 and 0.6 for NP cases. Also in this case, the dataset is not linearly separable in the sense of equation 3.31 since the optimisation problem has a solution.

Overall comparison of the models and conclusions

The performances metrics for each experiment are available in table 3.7. Generally, the SN models show less variation in both the ROC and the PRC area under the curve. We observe that the model performances do not vary a lot both in the figures and in the numerical results. The best scores are highlighted in bold text. They have been determined by comparing the median score and, in case of equality, the interquartile range. It appears that the model of experiment 4 is the best for NP cases. It has the best scores for the PRC AUC, the Cohen’s kappa and the Brier score for the test set. For SN cases, it is more complex. The model of experiment 3 has best scores for PRC AUC and ROC AUC while the model of experiment 4

has the best kappa score. We conclude that our attempts to reduce possible overfitting fail to do so. We believe that there is no overfitting phenomenon as can be observed by looking at the PR curves and that the regularisation term "restricts" the optimisation and prevents it from finding the optimal solution.

We have not shown the results of an experiment keeping all variables and applying ℓ_2 -regularisation. It has been done and has produced similar results.

By looking at all these scores being almost the same with small interquartile ranges whichever the experiment, we can suppose that no better parametrisation of the model could significantly improve the scores. We believe that we have reached the limits of logistic regression and that a more complex model would be more adapted. The limit can come from the weaknesses of logistic regression: dependence of the samples or correlation between the variables. We have shown that the dataset is not linearly separable. Moreover, we can imagine that removing a few points does not lead to linear separability. More generally, it is possible that the log of the odds linear relation with the variables does not capture well the real relations. The limit could also come from the data. It is possible that more training data would improve the performances on the test sets.

The confusion matrices do not vary a lot depending on the experiment (maximal variation of 2 samples). We believe that 17 % of ill-classified instances in test sets is a lot. Twenty percent of the conflicts are not detected for a threshold t of 0.5 for NP cases, 23 % for SN cases. It is possible to lower t but this leads to an increase of false positives. The PRC shows that aiming at a sensitivity of 0.9 gives a specificity of 0.45 for NP cases and 0.5 for the SN cases. If we expect a specificity of 0.9, the sensitivity falls to 0.2 and 0.3 for NP and SN cases respectively. We believe that a more accurate model is needed for an efficient foreign policy.

3.3.5 Coefficients of the logistic regression

The coefficient of logistic regression allows us to determine the importance of each feature for the probability prediction. A positive coefficient indicates that a higher value for this feature increases the predicted risk of conflict while a negative coefficient indicates that a higher value decreases the predicted risk. They correspond to the β_i values in the sense of equation 3.17. The values reported in Table 3.8 are the means and standard deviations of these quantities over all the external folds and the repetition of the procedure for experiments 3 and 4 which achieved the best results. We can take the means and standard deviations since the coefficients distribution is symmetric. As previously mentioned, all the variables have been centred and divided by their standard deviation. Our implementation using Scikit-learn applies this standardisation separately with the means and standard deviations of the train sets for the train sets and the test set. Otherwise, the overall mean and standard deviations would bring information of the test sample into the train sample. This would imply that, during the training phase, the algorithm has some information about the test set which is

not recommended for reliable assessment.

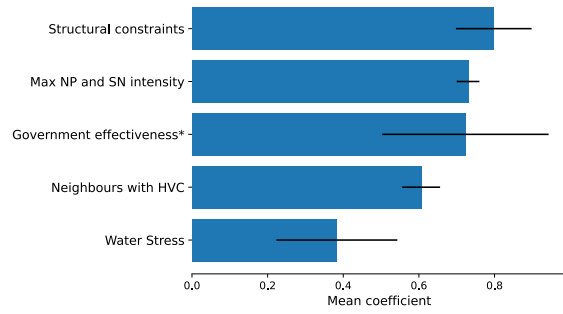
The high standard deviation of some coefficients compared to their own value makes it difficult to draw any conclusion. The concerned variables are:

- Neighbours with highly violent conflict (SN)
- Corruption (ℓ_1 -regularisation)
- Economic openness (all cases)
- Empowerment rights* (SN, ℓ_1 -regularisation)
- Ethnic power change (NP, ℓ_1 -regularisation)
- Food security* (all cases except NP with no-regularisation)
- GDP per capita* (ℓ_1 -regularisation)
- Government effectiveness* (SN, ℓ_1 -regularisation)
- Homicide rate (all cases)
- Income inequality (NP, ℓ_1 -regularisation)
- Lack of democracy (NP, ℓ_1 -regularisation)
- Regime type (SN, ℓ_1 -regularisation)
- Structural constraints (SN, ℓ_1 -regularisation)
- Unemployment rate (NP)
- Water stress (SN)

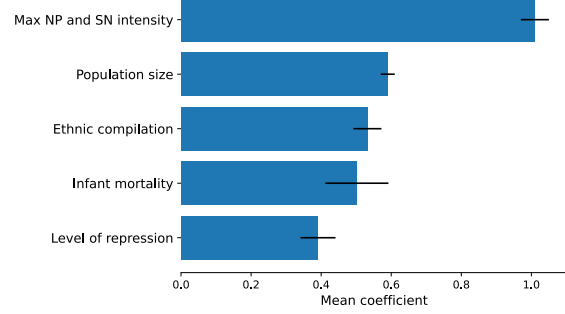
The five variables with the highest coefficients in absolute values are reported in Figure 3.7. We first observe that, whichever the regularisation, four out of five variables are the same. The only difference is Water Stress for ℓ_1 -regularisation and GDP Per Capita* for experiment 4. For SN cases, it is the Level Of Repression and the Corruption respectively.

We suppose that, when the weights are close to 0 for experiment 3 and not for experiment 4, it is a consequence of the feature selection. Examples of this are Corruption (NP), Empowerment Rights* (SN) or GDP Per Capita* (SN). This last variable has an unexpected influence on the predicted risk concerning the sign. One would expect that wealth is negatively related with conflicts since citizens generally have better life conditions and more to lose in beginning a quarrel.

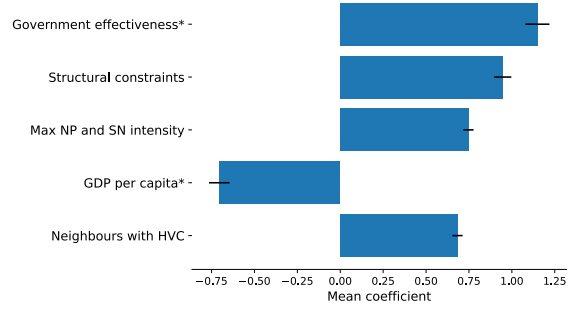
Looking at Table 3.8, we observe that both NP and SN weights have the same sign whichever the regularisation. This consistency gives credits to the computed coefficients. When discussing each variable, we mainly focus on the most interesting or counter-intuitive numbers.



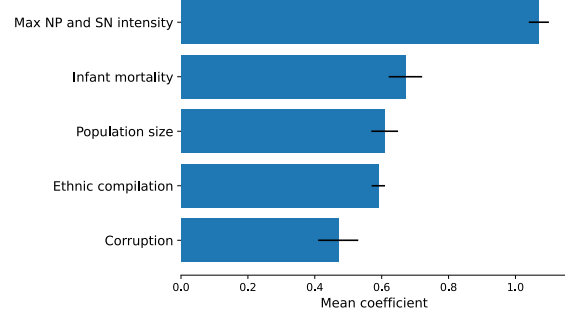
(a) Experiment 3: ℓ_1 -regularisation, NP



(b) Experiment 3: ℓ_1 -regularisation, SN



(c) Experiment 4: no regularisation, NP



(d) Experiment 4: no regularisation, SN

Figure 3.7: The five biggest mean coefficients of logistic regression for experiments 3 and 4 in absolute value. The black lines correspond to their standard deviations. The variables' names are given on the y -axis and the value of their coefficient on the x -axis.

Since the model is limited in terms of complexity, it is expected that some coefficients have the opposite sign as presumed.

It seems counter-intuitive that the weight attributed to Years Since HVC* is negative for SN. It is interesting to note that Transnational Ethnic Bonds have a positive relation with SN conflicts but a negative one with NP disputes. Empowerment Rights* have negative coefficients for all cases, whichever the regularisation. It might be that very repressive states give less possibilities to the citizens to let conflicts break out. The ethnic variable (Ethnic Compilation for SN and Ethnic Power Change for NP) does not show a clear relation with NP cases but a strong positive relation with SN cases. Food Security* also does not show a clear relation except for the no-regularised NP case where it seems illogical. It is negatively related with conflict. Surprisingly, the coefficients for Government Effectiveness* are negative for the SN case. In the SN case with no regularisation, the coefficient for Income Inequality is negative. Population Size's weights are positive in all cases which can seem logical. Lack Of Democracy has negative coefficients. As mentioned previously, anocracies tend to show more conflicts than pure autocracies or democracies. We could imagine that in this case, the dictatorship cases are more numerous than the democracies in the sense that the first one has more importance to determine the sign of the coefficient. The regime type's factors that we should interpret as the weight given to regime instability has a positive weight for NP cases but surprisingly, a negative weight for SN cases. Youth Bulge has a positive weight, that is particularly important in the NP-cases.

3.4 Random forest

3.4.1 Random forest: theory

In this section, another algorithm will be proposed for the classification task: random forest. In order to understand the concept of random forest, decision trees will first be introduced.

Decision tree

The concept of decision tree is explained in the context of machine learning and more precisely for classification tasks. To get the intuition, the reader can look at Figure 3.9. A decision tree is a procedure to follow with a boolean condition at each node. The number of condition nodes may or may be limited to a specific bound. The conditions have the following form:

$$x_{ij} < \theta_k \quad (3.46)$$

where θ is a determined threshold that applies specifically to variable j at node k . If the condition is met then the left path is taken, if not, then it is the right path. A theoretical scheme of a decision tree is given in Figure 3.8.

Variable	Experiment 3 NP, ℓ_1 -regularisation	Experiment 3 SN, ℓ_1 -regularisation	Experiment 4 NP, no regularisation	Experiment 4 SN, no regularisation
Years since HVC*	0.257 (0.03)	-0.121 (0.05)	0.299 (0.02)	-0.213 (0.02)
Max value: national and subnation intensity	0.73 (0.03)	1.01 (0.04)	0.747 (0.03)	1.07 (0.03)
Neighbours with highly violent conflict	0.606 (0.05)	-0.024 (0.03)	0.683 (0.03)	-0.074 (0.02)
Corruption	-0.088 (0.15)	0.277 (0.11)	-0.369 (0.07)	0.47 (0.06)
Transnational ethnic bonds	-0.168 (0.04)	0.326 (0.03)	-0.226 (0.03)	0.365 (0.02)
Economic openness	-0.011 (0.02)	-0.025 (0.02)	-0.05 (0.02)	-0.047 (0.02)
Empowerment rights*	-0.228 (0.06)	-0.09 (0.05)	-0.314 (0.05)	-0.158 (0.04)
Ethnic power change (NP)/ Ethnic compilation (SN)	0.01 (0.01)	0.532 (0.04)	0.013 (0.03)	0.59 (0.02)
Food security*	-0.048 (0.06)	0.004 (0.01)	-0.148 (0.04)	0.005 (0.04)
Oil production	0.169 (0.03)	0.103 (0.02)	0.148 (0.03)	0.084 (0.03)
GDP per capita*	-0.39 (0.16)	-0.07 (0.11)	-0.705 (0.06)	-0.303 (0.04)
Government effectiveness*	0.723 (0.22)	-0.043 (0.09)	1.149 (0.07)	-0.191 (0.06)
Homicide rate	-0.037 (0.04)	0.017 (0.02)	-0.08 (0.04)	0.027 (0.03)
Income inequality	-0.047 (0.03)	-0.141 (0.05)	-0.094 (0.03)	-0.236 (0.03)
Infant mortality	0.208 (0.12)	0.502 (0.09)	0.441 (0.06)	0.671 (0.05)
Population size	0.295 (0.05)	0.59 (0.02)	0.384 (0.03)	0.609 (0.03)
Lack of democracy	-0.057 (0.06)	-0.351 (0.05)	-0.141 (0.04)	-0.427 (0.04)
Regime type	0.186 (0.03)	-0.076 (0.05)	0.212 (0.03)	-0.143 (0.03)
Level of repression	0.286 (0.04)	0.391 (0.05)	0.348 (0.03)	0.47 (0.03)
Structural constraints	0.798 (0.1)	0.044 (0.07)	0.947 (0.05)	0.158 (0.05)
Unemployment rate	-0.005 (0.02)	-0.13 (0.03)	-0.035 (0.03)	-0.174 (0.02)
Water stress	0.383 (0.04)	0.007 (0.02)	0.441 (0.03)	0.03 (0.02)
Youth bulge	0.358 (0.05)	0.059 (0.03)	0.421 (0.04)	0.092 (0.04)

Table 3.8: Mean coefficients of the logistic regression obtained with ℓ_1 -regularisation or no regularisation, with all variables kept both for NP and SN. (the numbers in the brackets correspond to the standard deviations)

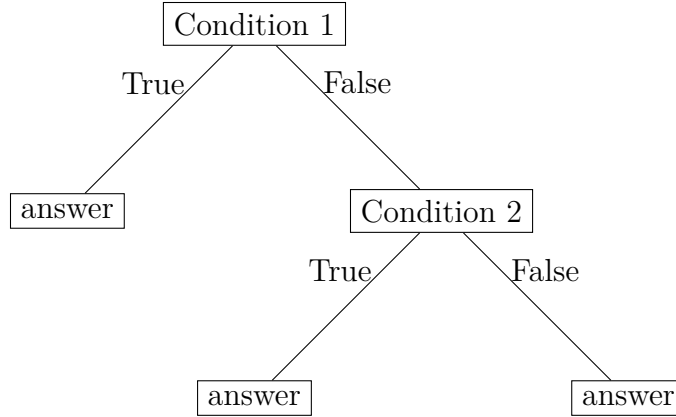


Figure 3.8: Theoretical scheme of a decision tree. At each node, there is a condition of the form $x_{ij} < \theta$. If this condition is met, we take the left branch, otherwise, we take the right branch. The leaves correspond to the final answer, either a class index (0 or 1) or a probability of belonging to $\mathcal{C}_1$.

The final answers are probabilities that $y_i = 1$. It is possible to transform them in $\hat{y}_i = 0$ or $\hat{y}_i = 1$ by fixing a discriminating threshold t .

Example 1 (continued)

For this example, we extend the dataset in Table 3.9.

	Country	Year	Corruption	Unemployment	Outcome (NP) y
1	Canada	2008	0.16	4.93	0
2	Yemen	2007	6.69	7.43	1
3	Angola	2001	8.69	5.3	1
4	Angola	2002	7.76	5.2	0
5	RDC	2002	8.06	5.73	1
6	Pakistan	2010	7.6	4.38	1
7	Afghanistan	2003	8.76	4.33	1
...	...	...	...	...	...

Table 3.9: Toy dataset for conceptual example

It gives the tree in Figure 3.9.

Trees are built by dividing the space into two half planes:

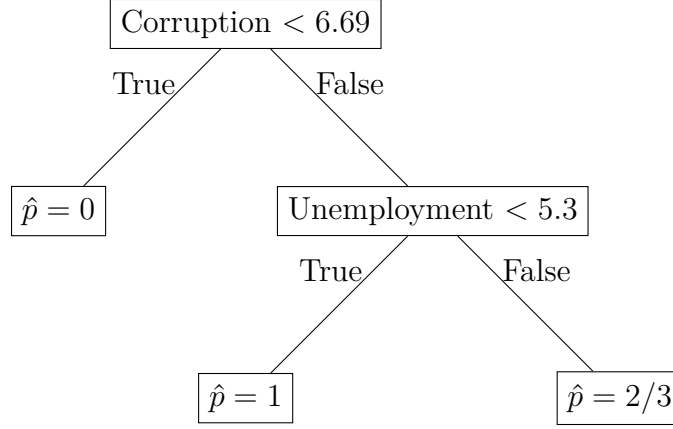


Figure 3.9: Conceptual example : decision tree

$$R_1 = \{x_i \text{ such that } x_{ij} < \theta\} \quad (3.47)$$

and

$$R_2 = \{x_i \text{ such that } x_{ij} \geq \theta\} \quad (3.48)$$

It corresponds to the division resulting from the boolean condition at each node. These regions can be split again in the same way. For each node k , the goal is to find the feature j and the threshold θ_k such that the new distinction maximises what is called node purity. In order to measure node purity, two metrics are used. One of them is the Gini impurity:

$$G_r = p_{r0}(1 - p_{r0}) + p_{r1}(1 - p_{r1}) \quad (3.49)$$

p_{r0} and p_{r1} being the proportion of respectively negative and positive instances in the r th region. A small Gini impurity reflects a high node purity. This means that there are few instances of the less commonly occurring class in each region.

The other metric is entropy, defined as:

$$D_r = -p_{r0} \log(p_{r0}) - p_{r1} \log(p_{r1}) \quad (3.50)$$

Again, a small value of the entropy implies a high node purity. For the purpose of clarity, we will refer to these criteria as node impurity. At each node, we have the following minimisation problem:

$$\min_{j, \theta} G_{r+1} + G_{r+2} \quad (3.51)$$

with j the variable index, θ , the applied threshold, G_{r+1}, G_{r+2} the Gini impurities for the two child nodes whose regions have index $r+1$ and $r+2$. We could also have taken the entropy for this minimisation problem. The process can stop either when no node impurity decrease is possible or when a specific criterion is met. For example, one can determine a maximum

depth, a minimum node impurity decrease, a minimal number of data points in a region to be split... Other criteria exist.

At the end, the outcome is determined by taking the proportion of positive instances in each leaf.

The problem of such model is overfitting [45]. Therefore, we introduce another algorithm: random forest.

From decision tree to random forest

The idea is to grow several trees and average their final answers. The yes-probability is obtained by taking the mean of all trees' yes-probabilities. Two additional techniques are used in order to improve performance. The first one is the bootstrap. It corresponds to growing each tree by only using part of the data. The samples are taken randomly with replacement. This reduces the variance [45]. The second one limits the number of features that can be considered for each split. Among p variables, m are chosen randomly at each node. It allows us to decorrelate the trees, which also reduces the variance [45].

Strengths and weaknesses of random forest algorithm

Strengths

- It reduces overfitting problem in comparison with decision trees [45]
- It works well with non-linear or complex problems [45]
- It can take complex features interactions into account.

Weaknesses

- Lack of interpretability in comparison with a single decision tree.
- Requires much training time and computational power. But it is parallelizable.
- When there are many irrelevant variables, it can perform poorly. Indeed, only m variables are considered at each split. As a consequence, the chance that a relevant variable is chosen is small [22].

Suitability of random forest for the conflict prediction problem

This model does not meet our needs due to the lack of interpretability. Indeed, policymakers are reluctant to use models they cannot understand. However, there exist techniques to solve this problem. We will introduce them in Chapter 4. It also requires a lot of training time, but the performances achieved were worth it, and it is feasible since the order of magnitude is a minute³. In our previous analysis, we have seen that the majority of the variables is

³Several tests should be conducted to derive precise time measures, the goal here is just to give an idea of the computation time magnitude order

relevant for the prediction. And the main advantage is that it can handle complex problems and generate good prediction performances as it can be seen in the results section 3.4.2. Therefore, random forest is adapted to our problem.

Hyperparameters tuning

By definition, the hyper-parameters need to be optimised separately from the training process. The chosen values are available in table 3.10. Using grid-search optimisation, we find a number of trees equal to 200. This number has been derived by first taking a broad range of values such as 100, 200, 500, 1000, 2000, and then refining the candidates range near the best values until no more improvement is observed. The same process is applied for the training sample size. Its value has been chosen among $n/4$, $n/3$, $n/2$, $2n/3$, $3n/4$ and n . The candidates for the maximal number of features m are 2,4,5,10 and 15. The best value, 4, corresponds to the rule proposed by Hastie et al. in [22]. They suggest taking $m = \lfloor \sqrt{p} \rfloor$. The choice between Gini or entropy for the impurity criterion does not influence the performances. Therefore, we let the default parameter which is Gini. Hastie et al. [22] recommend letting the tree depth free as it does not significantly improve the computing cost and results in less hyperparameter tuning. More generally, we do not set specific stopping criteria such as minimum impurity decrease, minimal samples split or minimal samples leaf. This implies that the trees grow as long as it is possible to reduce Gini impurities in certain regions by making new splits. Therefore, if there are no samples k and l such that $x_k = x_l$ but $y_k \neq y_l$, it is possible to classify all the training samples correctly. We have tried to limit trees growth to see whether it could prevent the random forest from overfitting and increase the tests sets performances, but it resulted in a decrease in all the metrics for the test and training sets.

During the hyperparameters tuning, we have observed that the most important one for performance improvement is the number of trees. For the other ones, most of the time, the default values were the good ones and changing them resulted in no improvement or even in a decrease of performances. That is why bootstrap is not used. We believe that the model has reached its maximum performances with the data it is provided with. However, we think that more data would improve its prediction power. Indeed, when we increased the number of external folds resulting in more samples for training, the performances on the test sets increased. This trend was confirmed over successive increasing of the number of folds. Of course, the improvement in the performance decreases each time since the increase of the training sample number is decreasing with the number of folds: when we go from 10 folds to 20 folds, the train samples number increase by $1/20$. When going from 20 to 40 folds, it only increases by $1/40$. For this reason, we believe that more data would improve the random forest performances. We are aware that the performances comparison is not totally accurate because the test sample sizes are not the same.

The detailed procedure for building a random forest is detailed in Algorithm 4. The procedure for using it is detailed in Algorithm 5.

Algorithm 4: building a random forest

```

1 Set  $N\_trees$ , the number of trees
2 Set  $m$ , the number of features considered at each split
3 for  $n\_tree = 1 : N\_tree$  do
4   create base_node and associate it region  $r_0 = \mathbb{R}^p$ 
5    $nodes \leftarrow \{base\_node\}$ 
6   while  $nodes \neq \emptyset$  do
7      $this\_node = nodes.pop()$ 
8      $M \leftarrow m$  values randomly chosen in  $\{1, \dots, p\}$ 
9      $r \leftarrow$  region associated with  $this\_node$ 
10     $N \leftarrow$  indices of the samples belonging to region  $r$ 
11     $G_r \leftarrow Gini(N)$ 
12    for  $j$  in  $M$  do
13      for  $i$  in  $N$  do
14         $Child\_1 \leftarrow \{k \in N \text{ such that } x_{kj} < x_{ij}\}$ 
15         $Child\_2 \leftarrow \{k \in N \text{ such that } x_{kj} \geq x_{ij}\}$ 
16         $G_{ij1} \leftarrow Gini(Child\_1)$ 
17         $G_{ij2} \leftarrow Gini(Child\_2)$ 
18        store  $G_{ij1}$  and  $G_{ij2}$ 
19      end
20    end
21     $k, l \leftarrow \underset{i,j}{\operatorname{argmin}} G_{ij1} + G_{ij2}$ 
22    if  $G_{kl1} < G_r$  or  $G_{kl2} < G_r$  then
23       $r$  gets split with threshold  $\theta = x_{kl}$  into  $r_1$  and  $r_2$ 
24      Create Child_Node_1 and Child_node_2
25      associate  $r_1$  to Child_Node_1 and  $r_2$  to Child_Node_2
26      append Child_Node_1 and Child_node_2 to  $nodes$ 
27    end
28  end
29  foreach leaf in the tree do
30    Compute  $\hat{p}$  by taking the proportion of positive instances at the leaf
31  end
32  Store tree
33 end

```

Hyperparameter	Value/option chosen
Size of training sample	n
Number of trees	200
Maximal number of features m	$\lfloor \sqrt{p} \rfloor = 4$
Impurity criterion	Gini
Maximum depth	None
Minimal samples split	2 (default)
Minimal samples leaf	1 (default)

Table 3.10: Random forest hyperparameters choice

Algorithm 5: using a random forest

```

1 set discriminating threshold  $t$ 
2 for  $i = 1 : n$  do
3   result  $\leftarrow 0$ 
4   for  $n\_tree = 1 : N\_tree$  do
5      $\hat{p}_i \leftarrow$  value obtained by feeding tree number  $n\_tree$  with instance  $i$ 
6     result  $\leftarrow$  result  $+$   $\hat{p}_i$ 
7   end
8    $\hat{p}_i \leftarrow$  result/ $N\_tree$ 
9   if  $\hat{p}_i < t$  then
10     $\hat{y}_i \leftarrow 0$ 
11  else
12     $\hat{y}_i \leftarrow 1$ 
13  end
14  store  $\hat{y}_i$  and  $\hat{p}_i$ 
15 end

```

3.4.2 Results obtained with random forest

The PR curves obtained are given in Figures 3.10a and 3.10b. The numerical results are available in Table 3.11. First, we observe that the classifier perfectly matches the training samples both on the plots and in the scores which are maximal for training sets, except the Brier Score, for all folds and repetitions as the interquartile ranges are 0. One could expect that it decreases the performances on the test set but apparently not. We have tested many hyperparameters values in order to reduce the performances on train set and increases predictive value on train set, but it always resulted in a decrease of the latter. The PR curve depicts better performances for the SN case and shows less variability. This is also reflected by the PRC AUC scores on test sets which are 0.962 and 0.972 with interquartile ranges of 0.023 and 0.015 for NP and SN cases respectively. The kappa scores on test sets are 0.894 and

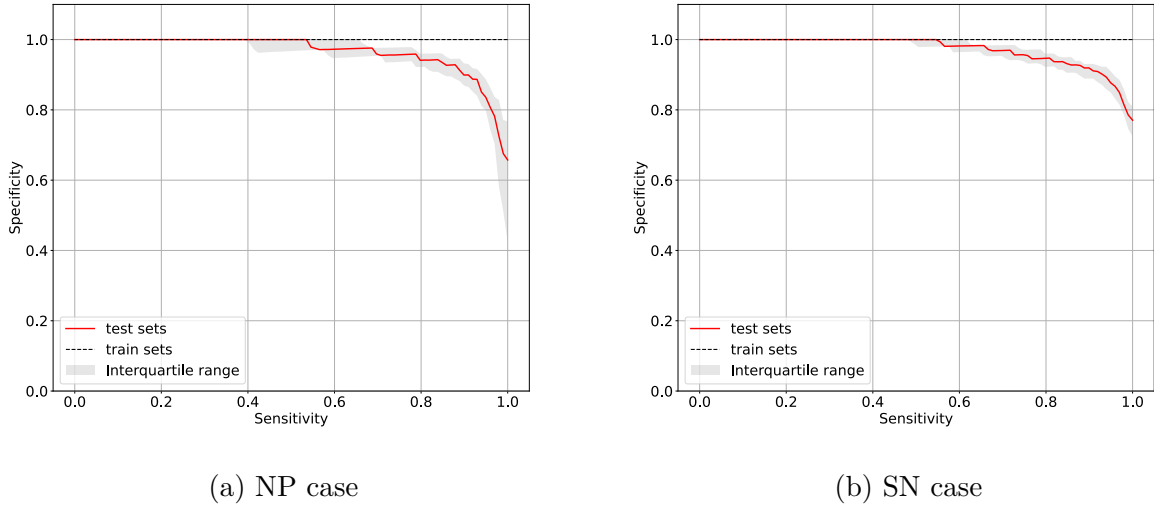


Figure 3.10: PR curves obtained with random forest classifier. The red line corresponds to the median of the test curves and the grey zone to their interquartile ranges. The dashed black line corresponds to the median train curve.

0.888 for NP and SN respectively. They are qualified as "almost perfect" according to Koch & Landis [52] (see Table 3.6). The ROC AUC being 0.989 and 0.99 for NP and SN cases respectively on tests sets are "outstanding discrimination" following Hosmer & Lemeshow [41] (see Table 3.4). The Brier Score is one order of magnitude smaller than the bounds of non-informativity derived in section 3.1.4.

The confusion matrix, gives us an error rate of 4.1% and 5.4% for NP and SN cases. Reported on the GCRI scale, it gives an error rate of 5.7 and 7.5 countries over 138. And that seems good. By changing t we could change these quantities. We also observe that for a sensitivity of 0.9, the specificity is 0.9 for both cases. For a sensitivity of 1, the specificity is 0.65 and 0.78. We believe that, due to the low error rate and the good sensitivity-sensibility trade-off, this model is adapted to provide accurate predictions in order to build an efficient foreign policy tool.

3.5 Comparison of the models and conclusion

Here we compare three models performances: the logistic regression built by Halkia et al. in [35], the logistic regression from section 3.3 and the random forest from section 3.4. The results are available in Figures 3.11a. Halkia et al. did not repeat the experiment several times. The median curve is derived from the 10 curves corresponding to the cross-validation folds. This is a limitation for the comparison since we repeated the experiment 10 times. The PRC median AUC are 0.778 and 0.821 with interquartile ranges of 0.062 and 0.066 for NP and SN

	NP		SN	
Median PRC AUC – test sets	0.962 (0.023)		0.971 (0.015)	
Median PRC AUC – train sets	1.0 (0.0)		1.0 (0.0)	
Median max kappa – test sets	0.894 (0.04)		0.888 (0.043)	
Median max kappa – train sets	1.0 (0.0)		1.0 (0.0)	
Median ROC AUC – test sets	0.989 (0.007)		0.99 (0.006)	
Median ROC AUC – train sets	1.0 (0.0)		1.0 (0.0)	
Median Brier score – test sets	0.035 (0.006)		0.044 (0.008)	
Median Brier-score – test sets	0.005 1.55E-04		0.006 1.82E-04	
Median confusion matrix for discriminating threshold $t = 0.5$ - test sets	52 (17.7%)	8 (2.7%)	81 (27.6%)	8 (2.7%)
	4 (1.4%)	230 (78.2%)	8 (2.7%)	169.5 (67%)

Table 3.11: Median results for random forest classification (the numbers in brackets are the interquartile ranges except for the confusion matrices where it corresponds to the percentages).

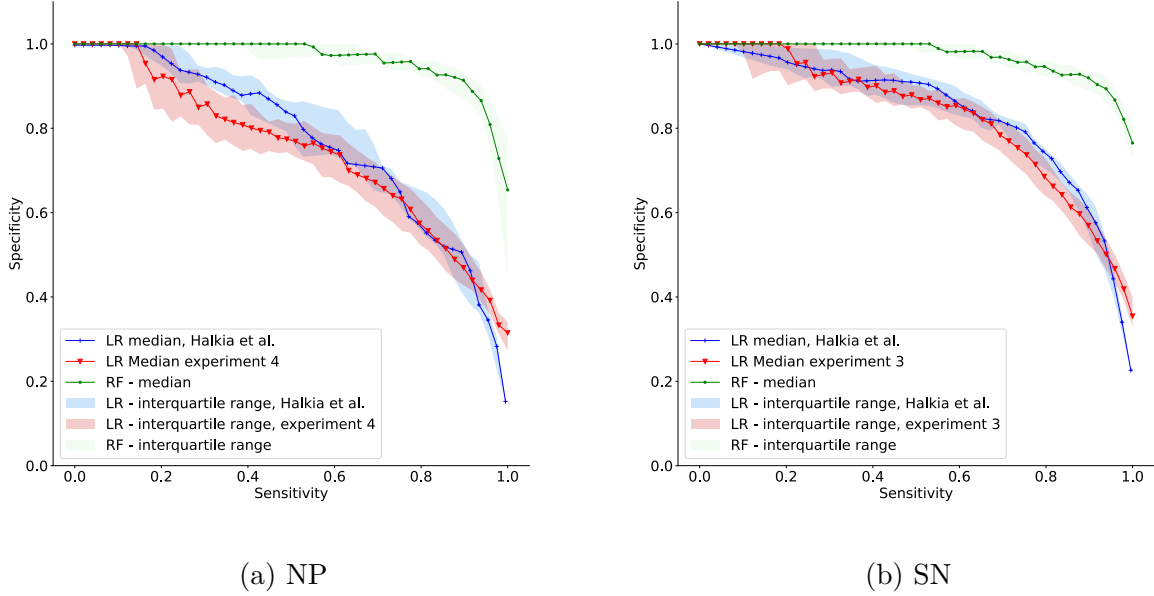


Figure 3.11: Comparison between the median PR curves for test sets obtained by Halkia et al. in [35] with logistic regression, by our implementation of logistic regression and by our implementation of random forest algorithm. The shaded area corresponds to the interquartile range. LR stands for logistic regression and RF for random forest.

respectively for the experiment realised by Halkia et al. [35]. Compared to our implementation of logistic regression, it scores better for NP but has bigger interquartile range. For SN, our implementation has better scores and shows lower variation. Our implementation of random forest scores better (0.962 and 0.971) and has lower interquartile ranges (0.023 and 0.015). On the plots, the distinction between our LR implementation and Halkia et al. one is not clear, the curves are near to each other. However, for a sensitivity of 1, our LR implementation shows a higher specificity both for SN and NP cases. Random forest has the best PR curve which is always above the other ones. It also has a lower interquartile range which depicts less variation in the performances. Halkia et al.'s median ROC AUC are 0.939 and 0.936 for NP and SN. This is better than our LR implementation but not than our RF implementation. The same applies for the Brier scores which are 0.062 (NP) and 0.078 (SN) for Halkia et al. We do not dispose of the Cohen's kappa score, so we cannot compare, but RF scores better than our LR implementation.

Based on the performances, random forest is the best model. It is the one we would choose among the three for an efficient foreign policymaking. However, the lack of interpretability is a real problem. That is why we will introduce an interpretable machine learning tool in Chapter 4.

Chapter 4

SHAP or how to make the black box white

In the conflict prevention field, one of the main challenges is to determine the most important variables to forecast conflict outbreaks. Additionally, policymakers need clear explanations about how the algorithm provided the answer they get. We would like to use a random forest algorithm since it performs better than logistic regression, but it does not meet the interpretability requirements as explained in section 3.4.1. We seek to know how $\hat{p}(x_i)$ depends on the features. Therefore, we need to develop a local explanation model g_i .

As a reminder, an explanation model is a function that provides results as close as possible to the original model, but that has a concise expression where the interaction between features and outcome are clearly highlighted. Typically, the probability $\hat{p}(x_i)$ outputted by the random forest algorithm has no concise and understandable expression as a function of the input variables x_{ij} , contrarily to the logistic regression. What we want to find is an expression that is as explicit and clear as the expression of logistic regression which is $\hat{p}(x_i) = 1/(1 + e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}})$, that would approximate $\hat{p}(x_i)$ for random forest.

Exemple 1 (continued)

We go back to our conceptual example with Corruption and Unemployment that are features 1 and 2, respectively. Two samples and their associated predicted probabilities are given in Table 4.1. The given values have already been scaled by removing their mean and dividing them by their standard deviation.

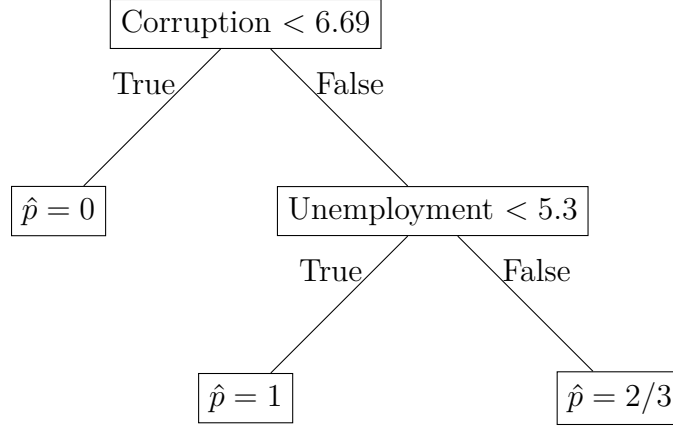


Figure 4.1: One of the trees of the conceptual example

	Country	Year	Corruption	Unemployment	Outcome (NP) y	$\hat{p}$
1	Canada	2008	0.16	4.93	0	0.234
2	Yemen	2007	6.69	7.43	1	0.95
3	Angola	2001	8.69	5.3	1	0.55
4	Angola	2002	7.76	5.2	0	0.54
5	RDC	2002	8.06	5.73	1	0.45
6	Pakistan	2010	7.6	4.38	1	0.89
7	Afghanistan	2003	8.76	4.33	1	0.55
...	...	...	...	...	...	...

Table 4.1: Toy dataset and predicted probability $\hat{p}$ for our conceptual example

We build a classification model based on a random forest algorithm with 20 trees of which the average result is given in column $\hat{p}$ of Table 4.1. An example of decision tree is given in Figure 4.1. However, the tree could be wider and deeper. For example, if all the True (or the False) branches are leaves, meaning that they have no child node, it can be as deep as the number of samples.

With 20 trees, the expression of the random forest function is nor concise nor easily understandable. It is difficult to highlight the important variables. Therefore, we need a function of the type :

$$g_i(x_i) = a_0 + a_1x_{i1} + a_2x_{i2} \quad (4.1)$$

Note that it is just an example, it could be non-linear or global. Here, we have chosen a local form as emphasised by the index i in g_i .

For tree-based models like random forests, it is possible to compute a specific feature importance score. It consists in computing the normalised impurity decrease for each feature. To achieve this, it is necessary to go through each node where the considered feature is responsible for the split and to compute the decrease in entropy or in the Gini impurity. We sum up these quantities and obtain a score for the feature. After repeating this process for each j , the scores are normalised. However, such computation tends to inflate the scores for features with high cardinalities [64] and they only provide global interpretation. In our dataset, some features are continuous and other ones are discrete, with less than ten values.

Feature permutation is another technique which consists in measuring the average increase in prediction error when a feature is permuted with another one. The procedure is repeated for many permutations and the average value is taken. However, this method explains the model globally. As stated above, we want to have a local explanation so that policymakers have a precise justification for each country. Additionally, these methods do not provide a clear relation between the model output and the variables. They simply give a normalised importance value.

For this reason, it is proposed in this section to examine this problematic of finding an explanation model in the light of the SHAP tool. It is based on Shapley values from game theory, and it provides both global and local analysis. Local analysis explains a particular prediction while global analysis explains the model generally over all predictions.

4.1 Method: SHAP values

4.1.1 Coalitional game and Shapley values

Before introducing SHAP values, we need to explain some basics of coalitional game theory, including Shapley values. All the information in this section comes from [48].

Coalitional game

Consider a game in the sense of game theory with N players. N can refer either to the number or to the set of players. S is a subset included in N . It is necessary to define the characteristic function v :

Definition 1. *Given a coalition game with N players, a characteristic function is a function of the form*

$$v : 2^N \rightarrow \mathbb{R} : S \rightarrow v(S) \quad (4.2)$$

with $v(\emptyset) = 0$, that assigns to each coalition $S \subseteq N$ the amount of transferable utility $v(S)$ it can achieve. It is called the worth of the coalition. [48]

From there, the coalitional game can be defined as :

Definition 2. A coalitional game is a pair (N, v) , where v is a characteristic function. [48]

Example 3

To get the intuition, we can imagine three cities that are near to each other and want to build a regional public transportation network. Each of them has a budget that corresponds exactly to the amount needed if they build the networks individually. By working together instead of individually, they could save on expenses. The three players are the cities 1, 2 and 3. From there, $N = \{1, 2, 3\}$. The amount of saved money is the characteristic function $v(S)$ and depends on the participation of the cities. Let us say that $(v(1, 2, 3) = 100, v(\{1, 2\}) = 50, v(\{2, 3\}) = 60, v(\{1, 3\}) = 50$ and $v(\{1\}) = v(\{2\}) = v(\{3\}) = 0$. They have to share the saved expenses in a satisfying way.

Shapley Values

More formally, the players have to share a payoff $v(S)$ corresponding to the value obtained for the coalition S . We want to find a payoff's distribution $\phi(N, v)$ satisfying the following axioms in which $\phi_i(N, v)$ is the amount attributed to player i :

Axiom 1. *Symmetry: two players contributing equally to each coalition deserve the same payoff: if $v(S \cup \{i\}) = v(S \cup \{j\})$ for all $S \subseteq N \setminus \{i, j\}$, then $\phi_i = \phi_j$ [48]*

Example 3 (continued)

Cities 2 and 3 contribute equally to each coalition. Because of symmetry property, they deserve the same payoff. In other words, $\phi_2 = \phi_3$

Axiom 2. *Support: Player i is a dummy if $\forall S \subset N$ with $i \notin S$, $v(S \cup i) = v(S) + v(\{i\})$. For any v , if i is a dummy, then $\phi_i(N, v) = v(\{i\})$. [48]*

This implies that a player that brings no value to the coalition gets a payoff of 0.

Example 3 (continued)

Suppose that there is a fourth city that does not save money by building its network with the others and that the other cities are not saving anything either when it participates in the network building, by support property, $\phi_4 = 0$.

Axiom 3. *Additivity: if v_1 and v_2 are any two characteristic functions, $\phi_i(N, v_1 + v_2) = \phi_i(N, v_1) + \phi_i(N, v_2)$ if $(v_1 + v_2)(S) = v_1(S) + v_2(S)$ [48]*

Example 3 (continued)

Back to the cities example, suppose that there are two distinct costs computed : the infrastructure and the vehicles. The money saved on the infrastructure is given by v_1 and by v_2 for the vehicles. The distribution can be done by considering these two costs separately or together, by additivity property, it produces the same final allocation.

Theorem 2. *Given any characteristic function v , there exists a unique function $\phi(N, v)$ satisfying the symmetry, support and additivity axioms which is given by ([48])*

$$\phi(N, v) = \sum_{S \subseteq N \setminus i} \frac{|S|!(|N| - |S| - 1)!}{|N|!} (v(S \cup \{i\}) - v(S)) \quad (4.3)$$

Intuitively, it corresponds to the weighted average of player's payoffs over all the possible orders of coming in the coalition, which is equal to the expected payoff if the players enter the coalition in a randomly chosen order.

Example 3 (continued)

Back to our cities example, it gives $\phi_1 = 30$ and $\phi_2 = \phi_3 = 35$. We observe that it obeys the symmetry axiom.

4.1.2 SHAP or how to determine features importance using Shapley Value

The idea introduced by Lundberg & Lee in [55] in 2016 is to apply the same principle to compute the features' importance in an individual prediction in case of a black box model. The players correspond to the features and the characteristic function is the prediction made based on a subset of features S as input of the model, up to a constant term. We will explain how the model makes a prediction based on a subset of features in a later section for the sake of understandability. For the moment, we consider it in an abstract way. The notations in this chapter are adapted from [62] and [55]. We define the coalition vector :

$$z \in \{0, 1\}^p \quad (4.4)$$

If feature j is in the coalition ($j \in S$), $z_j = 1$, otherwise ($j \in N \setminus S$), $z_j = 0$. We also define

$$x'_j = \begin{cases} 0 & \text{if feature } j \text{ is missing in the original vector } x_i \\ 1 & \text{otherwise} \end{cases} \quad 1 \leq j \leq p \quad (4.5)$$

Here, a missing feature refers to a feature for which there is no value in the original input vector x_i , contrarily to an absent feature which refer to a feature j where $j \in N \setminus S$.

Example 1 (continued)

If we look at our conceptual example, having no value in the dataset for Unemployment in Canada in 2008 as showed in Table 4.2 gives a missing feature whereas stating $S = \{1\}$ implies that, for this specific computation (not for this specific instance), feature 2 is absent of the coalition.

	Country	Year	Corruption	Unemployment	Outcome (NP) y	$\hat{p}$
1	Canada	2008	0.16	/	0	0.234
2	Yemen	2007	6.69	7.43	1	0.95
...	...	...	...	...	...	...

Table 4.2: Toy dataset and predicted probability $\hat{p}$ for our conceptual example

We now provide the explanation model :

$$g_i : \{0, 1\}^p \rightarrow \mathbb{R} : z \rightarrow g_i(z) \quad (4.6)$$

of which the explicit form is

$$g_i(z) = \phi_0 + \sum_{j=1}^p \phi_j^{(i)} z_j \quad (4.7)$$

with $\phi_j^{(i)}$ being the Shapley value for feature j when considering the instance i . A first observation is that the model is local, which is emphasised by the index i . It explains a single instance. For each sample, a specific model, with particular Shapley values, has to be computed. When the context is clear, we will omit the index i in g_i and $\phi_j^{(i)}$ for the sake of readability. A second observation that we can make is that g_i has an additive form (SHAP stands for Shapley Additive exPlanations). A third observation is that since z_j is binary, the feature value is "integrated" in the Shapley value. There is no multiplication between x_{ij} and a specific factor. We have said that SHAP provides both local and global explanations, and we only presented the local form. We will explain the global form in section 4.1.3.

Lundberg & Lee explain that this explanation model has three desirable properties.

Property 1. *Local accuracy:* if $\phi_0 = \mathbb{E}[\hat{p}(x)]$,

$$\hat{p}(x_i) = g_i(x') = \phi_0 + \sum_{j=1}^p \phi_j^{(i)} x'_j \quad (4.8)$$

Summing all the Shapley values gives $\hat{p}(x_i)$, up to a constant term ϕ_0 that is the same for all instances. The explanation model is not matched with $\hat{f}$ but with $\hat{p}$. This is needed since we must consider real values and not binary ones to compute Shapley values.

Property 2. Missingness:

$$x'_j = 0 \Rightarrow \phi_j^{(i)} = 0 \quad (4.9)$$

This implies that a missing feature must have a zero SHAP value.

For the next axiom, we need to explain how to deal with absent features as Shapley value's calculation depends on it. An absent feature is a feature that is not in coalition S . The technique consists in replacing absent features' values by values taken from randomly drawn instances, for the same features. In other words, if $j \in S$, then we keep x_{ij} , if $j \in N \setminus S$, we take x_{kj} , with k , a uniformly random integer between 1 and n , the number of samples. The randomly drawn instance is different for each absent feature. We now introduce a new notation :

$$h_i : \{0, 1\}^p \rightarrow \mathbb{R}^p : z \rightarrow h_i(z) \quad (4.10)$$

It is specific to each instance and maps the 1 to the original value x_{ij} and the 0 to the randomly drawn values x_{kj} . It is called the interventional approach since we modify the vector x_i .

Example 1 (continued)

Coming back to our conceptual example, let us imagine that we are now determining SHAP values for the prediction concerning Canada in 2008 which is instance 1. For this case, we have $x_1 = (0.16, 4.93)$ and $\hat{p}(x_1) = 0.234$. We then have for example $h_1((0, 1)) = (6.69, 4.93)$. The replacing value for Corruption is randomly taken from another instance which is Yemen in 2007 in this case. We could also have $h_1 = (8.69, 4.93)$ or another value for the first entry of $h_1((0, 1))$. Taking $h_1((1, 1))$ yields x_1 since no instance is missing.

We want to find $\phi_1^{(1)}$ and $\phi_2^{(1)}$ which are the importance given to Corruption and Unemployment for the difference between the model average prediction and $\hat{p}(x_1) = 0.234$.

Because of missingness property, these values must obey $\phi_1^{(1)} + \phi_2^{(1)} + \phi_0 = 0.234$.

Property 3. Consistency: We note $p_h(z) = \hat{p}(h_i(z))$, $z_{\setminus j}$ the vector z such that $z_j = 0$ ^[1]. For any $\hat{p}$ and $\hat{p}'$ that obey

$$p'_h(z) - p'_h(z_{\setminus j}) \geq p_h(z) - p_h(z_{\setminus j}) \quad \forall z \in \{0, 1\}^p, \quad (4.11)$$

then

$$\phi_j^{(i)}(\hat{p}', x) \geq \phi_j^{(i)}(\hat{p}, x) \quad (4.12)$$

¹We took the notation $z_{\setminus j}$ in [55] and kept it to avoid confusion.

Lundberg & Lee [55] provide us with the following result:

Theorem 3. *Only one possible explanation model g_i follows equation 4.3 and satisfies properties 1, 2, and 3.*

$$\phi_j^{(i)}(\hat{p}, x) = \sum_{z \in \{0,1\}^p} \frac{|z|!(p - |z| - 1)!}{p!} (p_h(z) - p_h(z_{\setminus j})) \quad (4.13)$$

where $|z|$ is the number of non-zero entries in z and $z_{\setminus j}$ is the vector z with entry j equal to zero.

Attention must be paid to SHAP values interpretation. They do not provide causal explanations between the variables and the predicted phenomena. They give the importance of each feature in the predicted probability of this phenomenon. More precisely, we can interpret the $\phi_j^{(i)}$ as the contribution of feature j to the difference between the average prediction ϕ_0 and the given prediction $\hat{p}(x_i)$ for instance i .

Example 1 (continued)

Back to our example, having a positive value for $\phi_2^{(5)}$ does not imply that a high unemployment rate generates conflicts. It only says that, given x_5 , the Unemployment has a positive effect on the predicted probability $\hat{p}(x_5)$

Discussion of absent features handling technique

One could argue about the way SHAP expresses features absences of the coalition. It would not be reasonable to retrain the model for each subset of variables since it would be too expensive in terms of computation time. However, tree-based models can handle absent features, and we could take advantage of this. It suffices, when a node criterion is based on an absent variable, to take both paths and compute the weighted average of the final leaves results. This is the same as taking the conditional expectation :

$$\mathbb{E}[\hat{p}(u)|x_{iS}] \quad (4.14)$$

where x_{iS} corresponds to the values of the features in the subset S . The expectation is taken over all instances in the training set. We have deliberately written u and not x to avoid confusion. This is called the conditional approach. Both the interventional and the conditional approaches have been proposed by Lundberg & Lee respectively in [55] and [57]. Comparing both approaches, it looks like it would be easier to take the conditional approach than sampling many vectors for each subset of variables.

Both techniques assume features independence. Nevertheless, neglecting this assumption has a different impact on the results. Janzing et al. have shown in [46] that the conditional approach leads to incorrect approximations of Shapley values in case of feature dependence.

Some features that do not have importance in the prediction can get non-zero SHAP values and this violates the Support axiom (2). In the dataset that we use, the features are correlated. Even though the interventional approach also assumes features independence, it leads to less biased results than the conditional approach and these last results are worth to be considered [46]. That is why we chose this technique.

If we assume linear relations between the variables, we could decorrelate the data in order to use the SHAP conditional approach since it provides exact Shapley values when the variables are independent [57]. Unfortunately, this would result in a serious decrease in interpretability. The new variables would be linear combinations of older ones, and it would be difficult to draw conclusions on the model.

SHAP values computation

Lundberg & Lee propose a method in [55] to estimate $\phi_j^{(i)}$. It consists in generating M random vectors $z_k \in \{0, 1\}^p$ with $k \in \{1, \dots, M\}$ for each sample. M is the budget for the computation. A high M results in better approximations but a bigger computation time. Then, we compute $p_h(z_k)$ and assign it a weight:

$$\pi(|z_k|) = \frac{p - 1}{\binom{p}{|z_k|} |z_k| (p - |z_k|)} \quad (4.15)$$

The weight depends exclusively on the total number of features p and the number of features considered (in the coalition set S), $|z_k|$. This means that we can save calculations by computing these weights beforehand and avoid repeating the operation for each instance. We then fit the linear system:

$$g_i(z) = \phi_0 + \sum_{j=1}^p \phi_j^{(i)} z_j \quad (4.16)$$

As explained earlier, ϕ_0 is the average prediction for the model. For decision trees, it is computed by taking the prediction for each leaf weighted by the number of instances in that same leaf. For random forest, it is the average of each decision tree. To obtain the $\phi_j^{(i)}$, we seek to minimise the following function, corresponding to a weighted sum of squared errors:

$$\min_{\phi_1^{(i)}, \dots, \phi_p^{(i)}} \sum_{k=1}^M (p_h(z_k) - g_i(z_k))^2 \pi(|z_k|) \quad (4.17)$$

It is surprising to see that Shapley values can be estimated using a linear regression. The computed coefficients provide estimates for the Shapley values for sample i . Since we have a limited budget to generate coalition samples, we can take advantage of the weighting scheme. It gives more importance to coalitions with few features or to coalitions with almost all features but less importance to middle-sized coalitions. Intuitively, one can learn about a feature's individual effect in small coalitions and about features interaction effects in large coalitions. Therefore, in order to reduce computation cost, we can begin by sampling vectors

with all but one feature and vectors with only one or two features [62]. Alternatively, we could truncate the sampling by making instances likelihood proportional to their weight and remove $\pi(|z_k|)$ from equation 4.17. Again, we can save computation time by keeping the same vectors z_k with $1 \leq k \leq M$ for all instances.

To effectively compute these values for the whole dataset, we apply algorithm 6 that summarises the above section. For the sake of simplicity, we assume that the dataset has already been rescaled by removing features' means and dividing these by their standard deviations.

Algorithm 6: Computing SHAP values

```

1  $\phi_0 \leftarrow \mathbb{E}_{X^{train}}[\hat{p}(u)]$ 
2 for  $m = 1 : p$  do
3   | compute  $\pi(m)$ 
4   | store  $\pi(m)$  as  $\pi_m$ 
5 end
6 for  $i = 1 : n$  do
7   | for  $k = 1 : M$  do
8   |   | choose  $z$  randomly in  $\{0, 1\}^p$ 
9   |   | store  $z$  as  $z_k$ 
10  |   | for  $j = 1 : p$  do
11  |   |   | if  $z_j = 0$  then
12  |   |   |   |  $h_j \leftarrow$  uniformly random chosen value among  $x_{.j}$ 
13  |   |   | else
14  |   |   |   |  $h_j \leftarrow x_{ij}$ 
15  |   |   | end
16  |   | end
17  |   |  $p \leftarrow \hat{p}(h)$ 
18  |   | store  $p$  as  $p_k$ 
19  | end
20  | solve  $\min_{\phi \in \mathbb{R}^p} \sum_{k=1}^M (p_k - \phi \cdot z_k - \phi_0)^2 \pi_{|z_k|}$ 
21  | store  $\phi$  as  $\phi^{(i)}$ 
22 end

```

4.1.3 SHAP : global interpretation

The global feature importance can be computed by taking the mean absolute SHAP values over all the instances:

$$I_j = \frac{1}{n} \sum_{i=1}^n |\phi_j^{(i)}| \quad (4.18)$$

with I_j , feature j 's global importance and $\phi_j^{(i)}$, SHAP value for feature j given instance i .

4.1.4 SHAP interaction values

The extension of SHAP values are SHAP interaction values introduced in [56]. With these, we can understand that the same feature value gets a different SHAP value for different instances. Unfortunately, the computation of interaction values has not been implemented yet in the SHAP toolbox [58] for interventional expectation. It has been done for conditional expectation that we cannot use as explained above. We discuss it here because we believe that it is relevant and should be part of future works.

SHAP interaction values are defined the following way for $i \neq j$:

$$\phi_{i,j} = \sum_{z \in \{0,1\}^p} \frac{|z|!(p - |z| - 2)!}{2(p-1)!} \delta_{ij}(z) \quad (4.19)$$

with

$$\delta_{ij}(z) = f_x(z) - f_x(z_{\setminus j}) - f_x(z_{\setminus i}) + f_x(z_{\setminus i,j}) \quad (4.20)$$

$$= f_x(z) - f_x(z_{\setminus i}) - \left(f_x(z_{\setminus j}) - f_x(z_{\setminus i,j}) \right) \quad (4.21)$$

knowing that $z_{\setminus i,j}$ is the vector z with entries i and j equal to zero. The same interpretation holds for $z_{\setminus i}$ and $z_{\setminus j}$. We can then compute

$$\phi_{ii} = \phi_i - \sum_{i \neq j} \phi_{ij} \quad (4.22)$$

This last quantity is called the main effect and can be interpreted as the "pure" effect of feature i , with all interaction values removed. In equation 4.19, the factor two at the denominator is there to equally split the interaction value between ϕ_{ij} and ϕ_{ji} which are consequently equal. For this reason, the total interaction effect is equal to $\phi_{ij} + \phi_{ji} = 2\phi_{ij} = 2\phi_{ji}$. The equation 4.20 can be decomposed in two parts as highlighted in equation 4.21. The first part is the feature i 's effect when feature j is in the coalition and the second part is feature i 's effect when feature j is absent of the coalition. The interaction effect is the difference between these two quantities. The interaction effects obey the same axioms as SHAP values [56]. It can be extended to more than two features by applying concepts developed by Fujimoto et al. in [23]. However, there is a limitation. Interaction effects are difficult to represent. For each instance, we have a $p \times p$ symmetric matrix $\mathcal{I}^{(i)}$. To get the global interaction values, we could possibly build $\mathcal{G}$ such that $\mathcal{G}_{kl} = \frac{1}{n} \sum_{i=1}^n \mathcal{I}_{kl}^{(i)}$ but that would be at the cost of local explanation.

4.1.5 SHAP: strenghts and weaknesses

Strenghts:

- It provides global and local explanations. A general trend is given for the predictor by taking the mean over all SHAP values for a given feature. Additionally, a local

interpretation is given. This means that we can analyse predictions that do not follow the global trends.

- SHAP is built through axiomatic derivation.

Weaknesses

- It is computationally expensive even with approximations
- It ignores features dependence. Replacement of values in the samples leads to unlikely or even unrealistic data.

4.1.6 Conclusion : is SHAP appropriate for our problem ?

We already discussed SHAP appropriateness in the introduction of this chapter. It has many advantages over the other methods. And its disadvantages are compensated by the fact that it provides both local and global explanations and is built through axiomatic derivation.

4.2 Results

We have computed the SHAP values for the 100 trained models (10 external folds for the 10 repetitions with data shuffling). It appeared that the global explanations showed similar results. The order of importance was the same up to two or three elements. Some local explanations varied from model to model. We suspect that it is due to the way random forests are built. As explained in Chapter 3, at each node, a fraction of the features is randomly sampled to be considered for the split. Since some of our variables are not independent, it is possible that for two different models, two variables carrying the same information do not have the same place and importance in the tree.

Our main goal being to provide explanations to policymakers about one model prediction, we select the best classifier according to precision-recall score. This approach is of course limited by the fact that this high score might be due to the validation fold composition. It could be that in this special case, all the instances in the test set have equivalent instances in the training set. By equivalent instances, we mean instances with many similarities such that the model has been trained for this type of case. This would imply that the model is not especially the best one. New data would be needed to assess this. The selected models have a PRC AUC of 0.989 and 0.995 for NP and SN cases respectively. The background dataset is the dataset from which we randomly draw instances for the replacement of absent features' values. We choose the original dataset as background dataset after removing of the means and division by the standard deviations. Then the SHAP values are computed using SHAP toolbox [58].

One can get the intuition behind SHAP by looking at Figure 4.2, It is the prediction for

Angola in 2002, for NP. The predicted probability is on the x -axis. The variables and the SHAP values in bracket are on the vertical axis. Starting from the bottom, with ϕ_0 , the model average prediction (equal to 0.183), we can follow the evolution of $\hat{p}$ resulting from the features values. Ethnic Power Change has a value of 0 and no impact on the prediction. Level of repression has a score of 7.5, and it increases $\hat{p}$ by 0.001, which is the SHAP value linked to this variable. $\hat{p}$ continues to increase for other features until we reach Regime Type whose score is 3.98. It has a negative SHAP value and decreases the predicted probability by 0.004. For the sake of readability, the features have been ordered following SHAP value (in absolute value). The five most important variables are in increasing order, Population size (negative effect), Structural Constraints (positive effect), Max value: NP and SN (positive effect), Income Inequality (negative effect) and Years Since HVC* (positive effect). Since the Structural Constraints are high (9), it seems normal that it increases $\hat{p}$. We observe that even a value of 5 for Max Value: NP And SN has an important impact on $\hat{p}$. The effect of high score for Income Inequality is surprising: we would expect that an unfair distribution of the wealth is positively related to conflict. The effect on $\hat{p}$ is due to the feature value but also sometimes to other features values, because of interaction effects. We believe that the interaction effect can lead to this counter-intuitive SHAP value for Income Inequality. Finally, we observe that the recent conflicts (Years Since HVC* =9) have a big impact on $\hat{p}$ and this seems logical since past conflicts are positively related with future ones. This plot is something that would increase the policymakers' trust in random forest prediction since it makes it more transparent. It would help for mutual understanding in interdisciplinary teams. Of course, SHAP is based on an approximation and relies on assumptions. But it gives a better idea of the prediction process.

In Figures 4.3 and 4.4, we display the global explanations for the whole dataset, train and test sets mixed. On each row, every dot is an instance and its colour corresponds to the feature value. The horizontal axis gives the SHAP value. By looking at it, we can see that the trend is clear for some features like Neighbours With HVC where the blue dots are on the left side and the red dots on the right side. For some other ones, like GDP Per Capita*, it is less clear: we observe from the colour mix that the feature value does not evolve monotonically with the SHAP value. For the 6 least important features, the SHAP values are small or equal to zero most of the time. We suppose that these features are not important but can have interaction effects with other ones for certain ranges of feature values. The five most important ones are the same for SN and NP situations up to one variable, Structural Constraints (NP) and Transnational Ethnic Bonds (SN). To gain a deeper insight, we draw the dependence plots in Figures 4.5 and 4.6. It consists in plotting all pairs $(x_{ij}, \phi_j^{(i)})$ for $1 \leq i \leq n$ and a fixed j . It shows the relation between the feature values and the SHAP values. The subplots are ranked according to feature global importance as defined in equation 4.18. For the sake of readability, the features on the plots correspond to their original values, not to the rescaled-ones. Two instances k and l with $x_{kj} = x_{lj}$ can get two different SHAP values for this feature: $\phi_j^{(k)} \neq \phi_j^{(l)}$. That is due to interaction effect with other features.

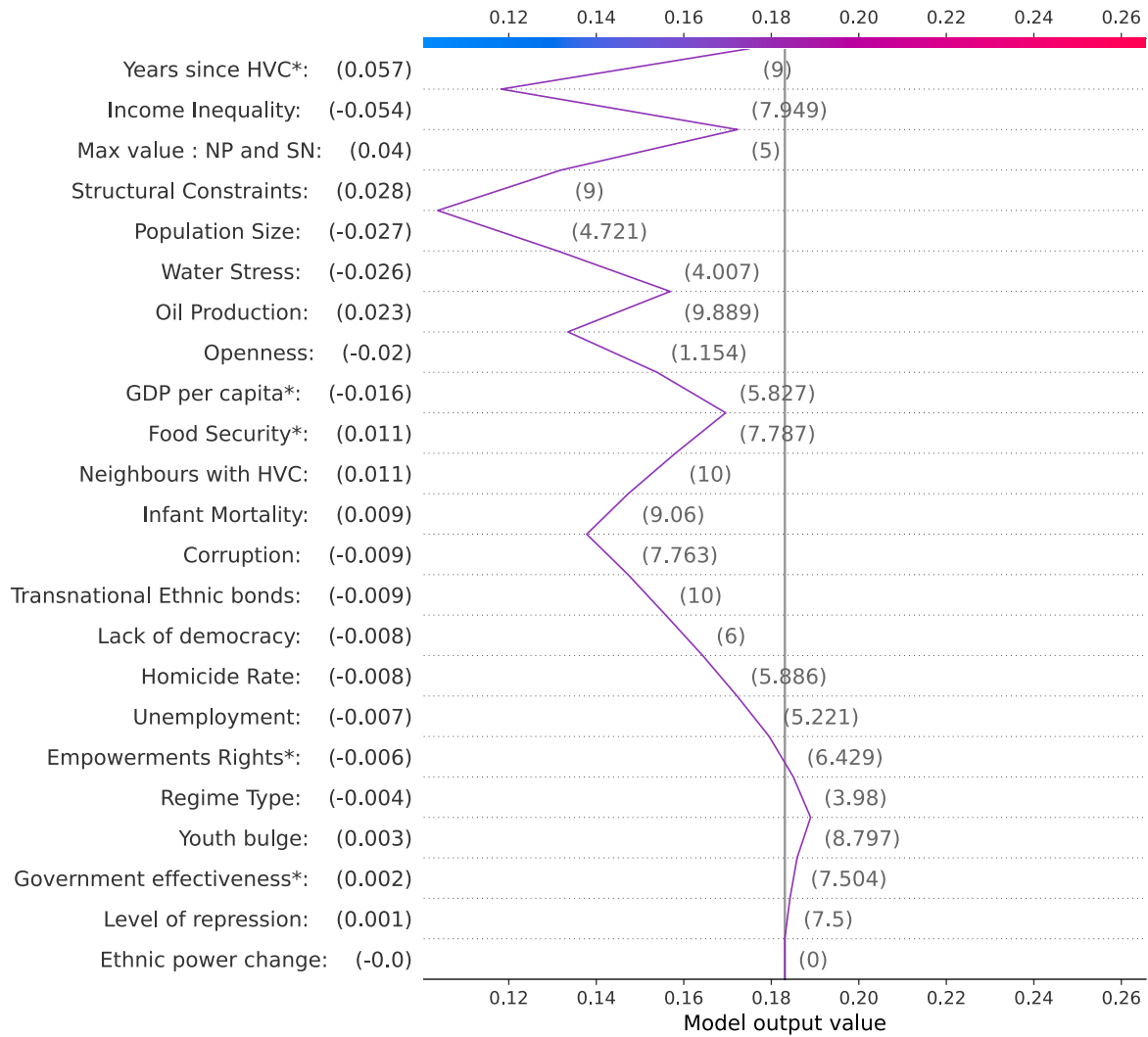


Figure 4.2: Decision plot for Angola in 2002, NP. On the left, the feature names are displayed with the SHAP values in brackets. The line starts from the expected value for the classifier which is 0.183. It then follows a trajectory according to the sum of the SHAP values displayed on the *y*-axis. The features' values (not rescaled) are given in brackets next to the line.

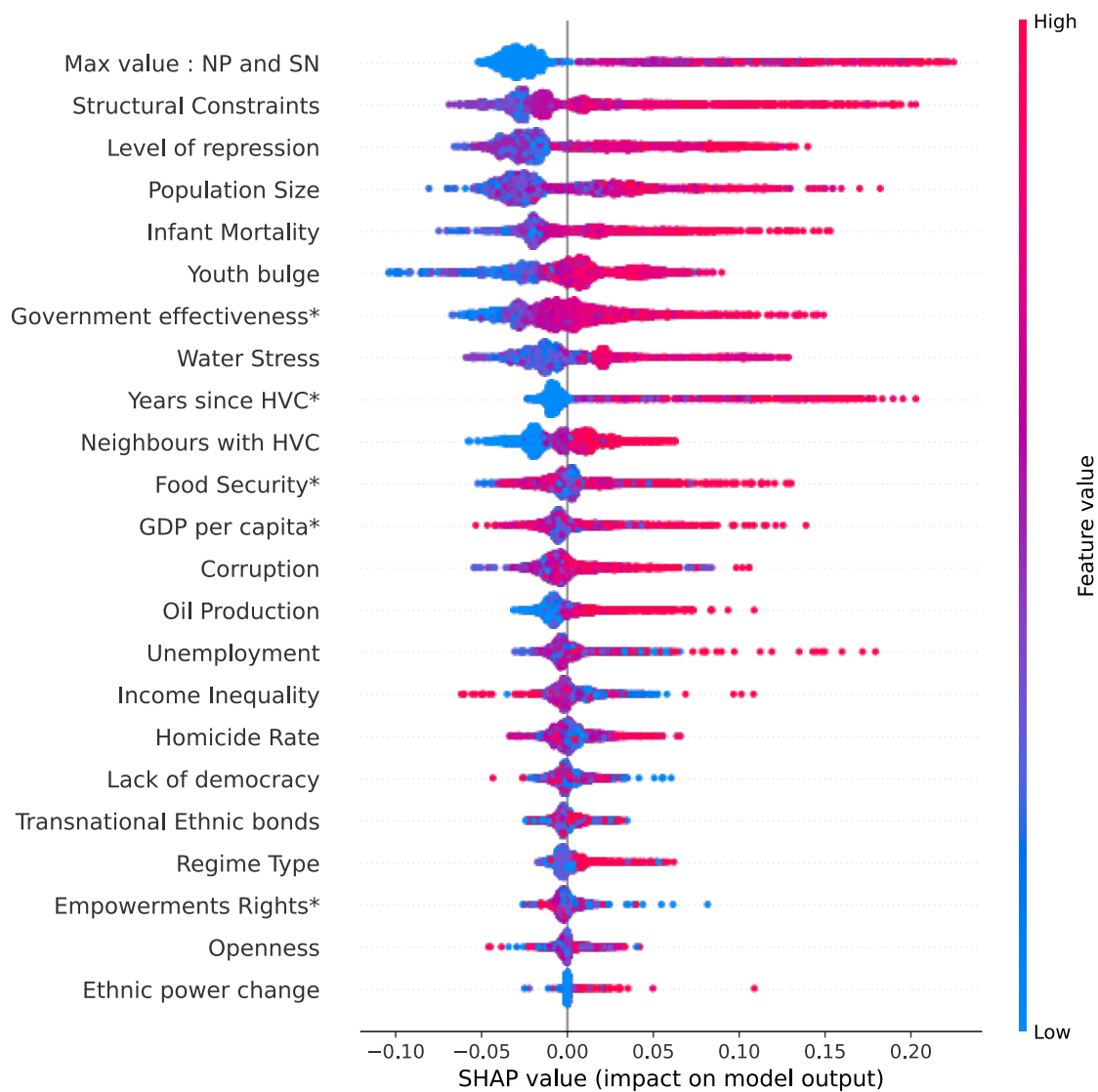


Figure 4.3: Summary plot for all the dataset instances (NP). It depicts the global explanation model. For each line, a dot corresponds to an instance and the colour to the feature value. The dot abscissa corresponds to the SHAP value for the feature displayed on the ordinate.

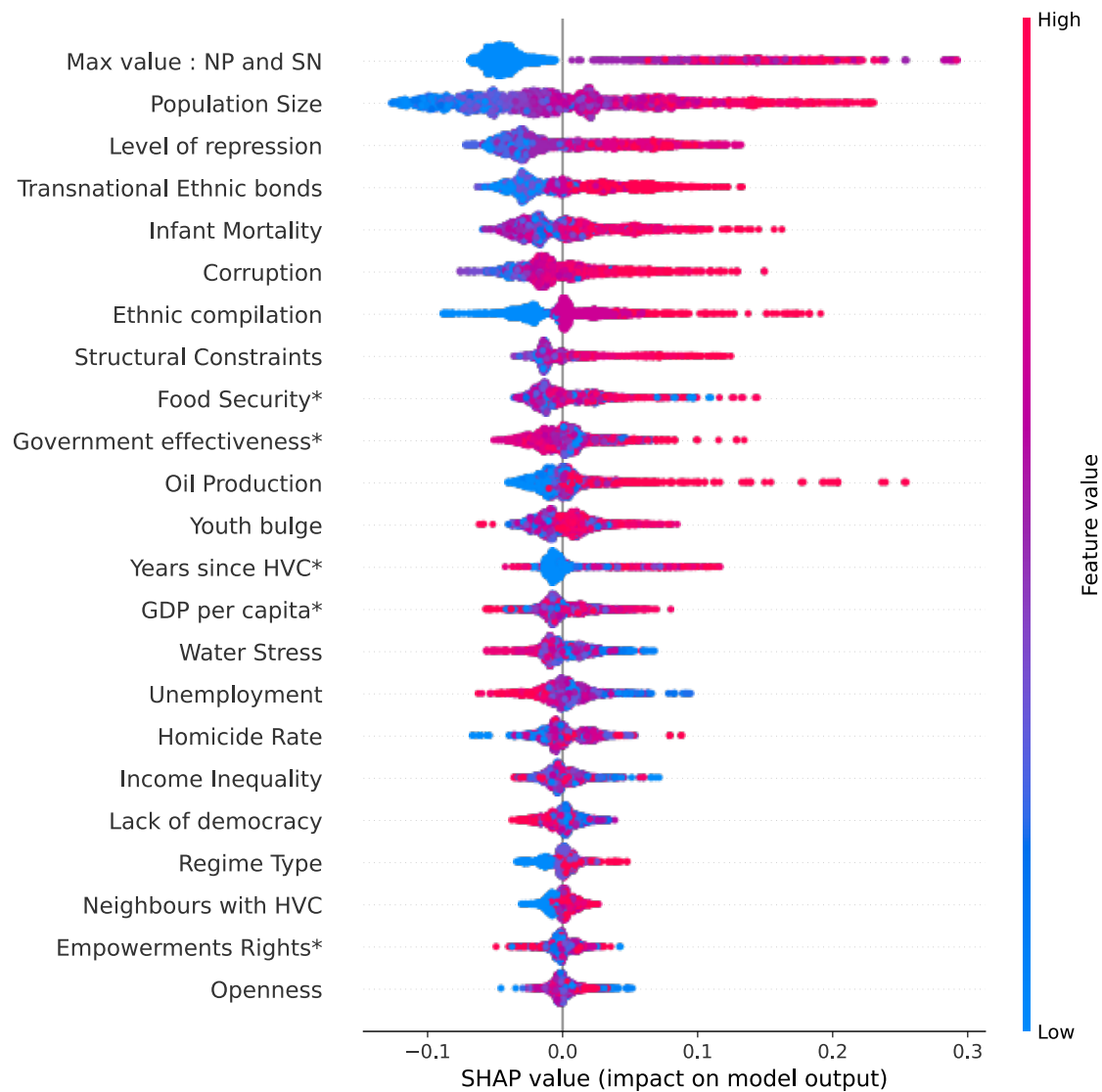
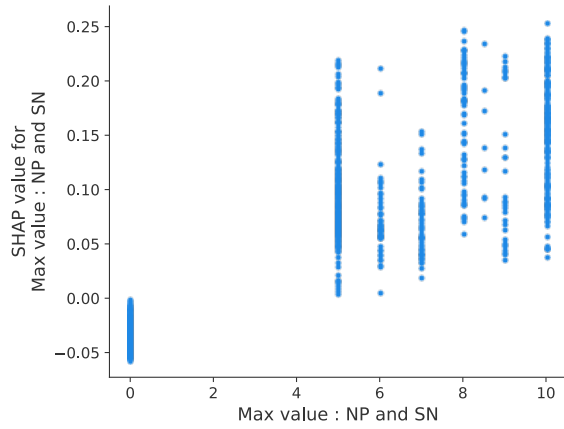
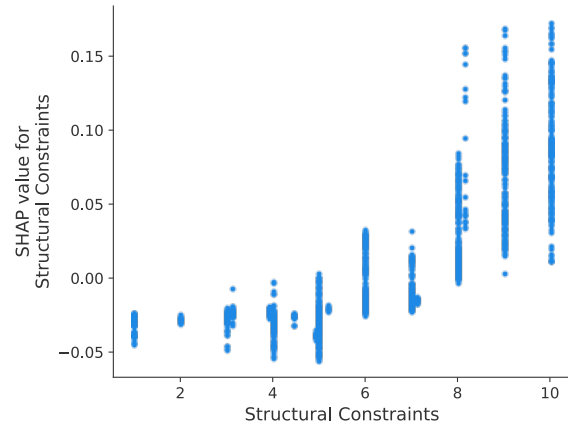


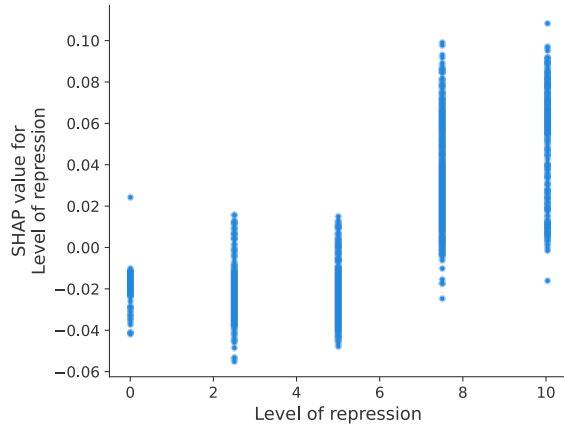
Figure 4.4: Summary plot for all the dataset instances (SN). For each line, a dot corresponds to an instance and the colour to the feature value. The dot abscissa corresponds to the SHAP value for the feature displayed on the ordinate.



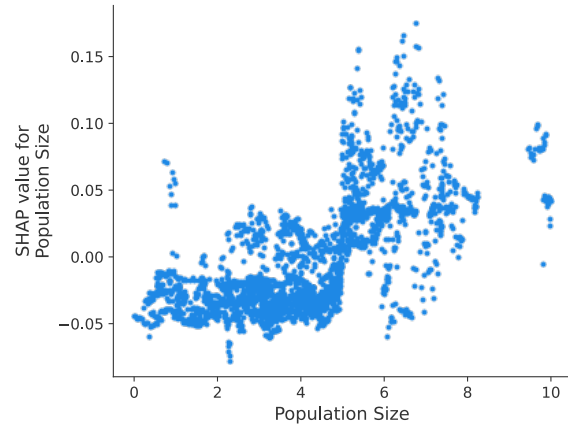
(a)



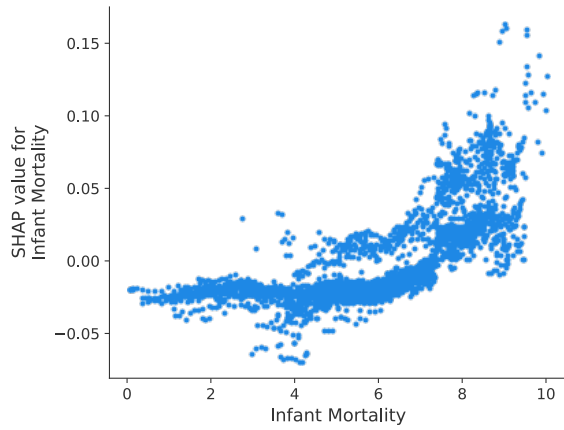
(b)



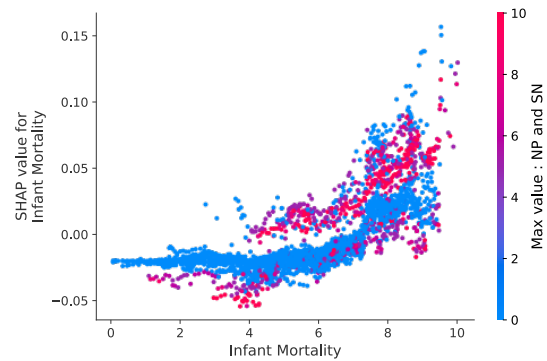
(c)



(d)

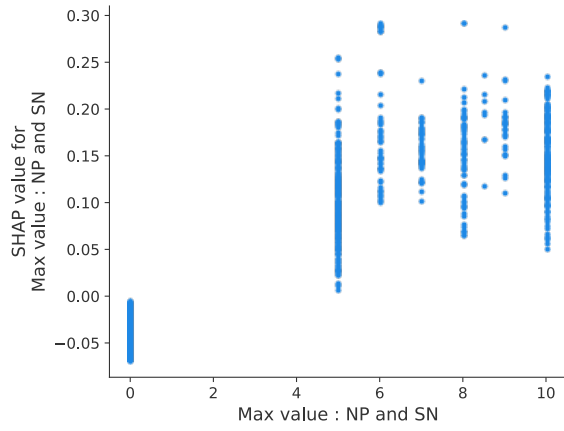


(e)

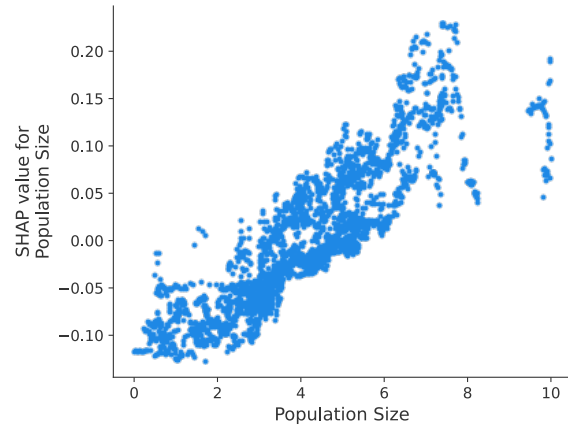


(f)

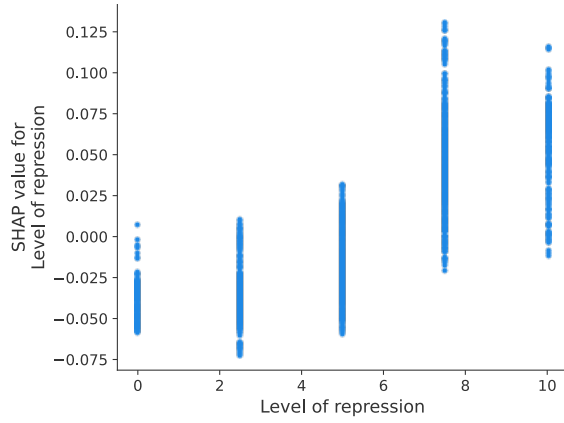
Figure 4.5: Dependence plots between SHAP and features values for the five most important variables (NP). It is a representation of the pairs $(x_{ij}, \phi_j^{(i)})$.



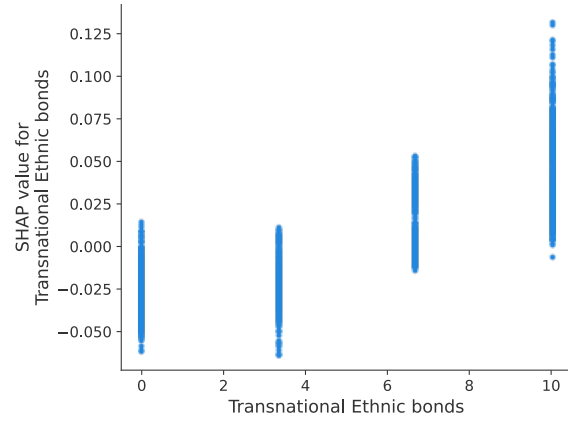
(a)



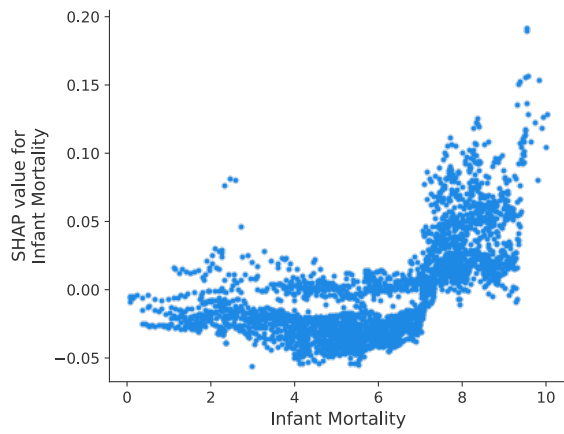
(b)



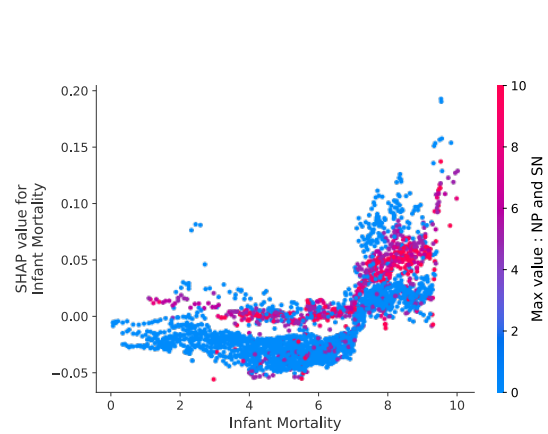
(c)



(d)



(e)



(f)

Figure 4.6: Dependence plots between SHAP and features values for the five most important variables (SN). It is a representation of the pairs $(x_{ij}, \phi_j^{(i)})$.

Analysis for NP fights

When Max Value: NP and SN equals 0, the effect is from -0.05 to 0 on $\hat{p}$. When it is above 5, no clear trend appears. It seems that there are other factors intervening in the SHAP values and that it results in high variability. Interaction effect computation would probably help understand this variability. For Structural Constraints, we observe a non-linear effect in Figure 4.5 with increasing return of probability as the feature value increases. Again, there is a high variability, which shows that there are interaction effects with other variables. Regarding the Level Of Repression, it seems that values above 6 have a clear increasing effect on the predicted probability and conversely for feature values below this threshold. Again, there is a high variability and there must be interaction effects. For Population Size, it seems that there are two major groups of country-years. The first one has a constant negative impact on $\hat{p}$ until it reaches a value of 5. From there it grows steeply until 0.15. This stops at a feature value of 5. The second group has a constant effect, near 0 for feature value from 2 to 6. From there, the variability grows, and the feature effect is not clear, going from -0.05 to 0.15. We cannot confirm that these two groups are separated, but it is plausible. This distinction and the variability within these groups is probably due to interaction values. A third minority group of country-years is isolated at a value of approximately 10. For this one, no clear effect is observable. For Infant Mortality, it seems again that there are two groups, following the same trends. The first one has a constant effect, just below 0 until the feature reaches 5 where the SHAP values begin to raise. The second group which begins at feature value of 4, is above the other one but follows the same non-linear trends. This becomes clear when we look at Figure 4.5f, the interaction plot between Infant Mortality and Max value: NP And SN. The dot's colour corresponds to Maximal NP and SN intensity. This value partly explains the belonging to one group or to the other and the ordinate variability for a same abscissa value. Such interaction plot does not show the interaction value as explained in section 4.1.4. It displays all the SHAP — feature values pairs and additionally colours the dots according to another feature. Such plots are very informative, but we unfortunately cannot show many of them. Just for the five most important features, there are 20 graphs possible.

Analysis for SN fights

For Max Value: NP And SN and Level of repression, we globally observe the same effects as for NP. For Population Size, except for the minority group we already observed for NP, we can see a linear effect on SHAP value with variability along the y -axis. This is different than for NP. For Infant mortality, it seems that there are two groups. The first one starts at zero with constant effect from 0 to 2 and a parabola shaped curve from 2 to 6, always in the negative values. After this, there is a steep grow around 7 and an increase in variability from 8 to 10 with SHAP values from 0 to 0.04 and 0.07 to 0.12. The second group is above the other one. It follows the same trends but with less variability at the end. The distinction is made clear in Figure 4.6f.

From chapter 3, we remember that GDP Per Capita* behaved in an unexpected way for

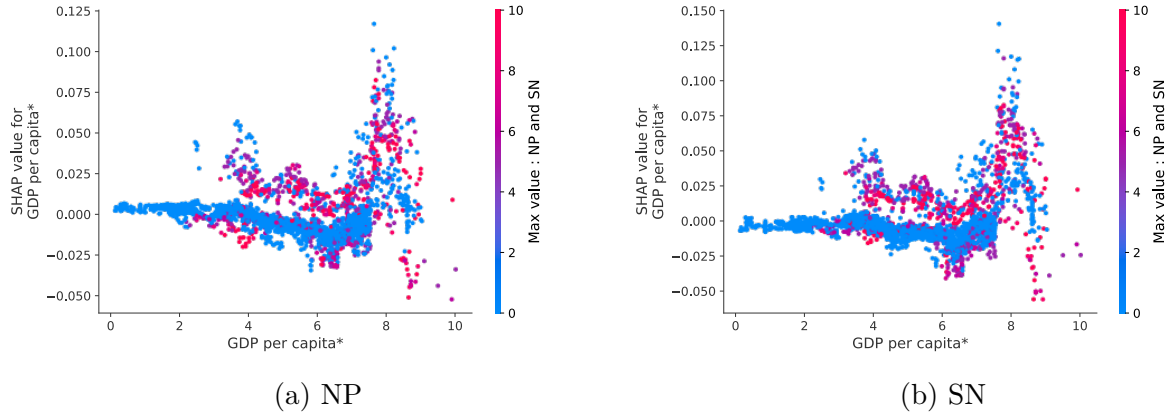


Figure 4.7: Interaction plots for GDP Per Capita* and Max Value: NP and SN. The dots represent the pairs $(x_{ij}, \phi_j^{(i)})$ for $1 \leq i \leq n$. On the abscissa, we have the GDP Per Capita* value and we have the SHAP value on the ordinate. The colour of the dots is related to the value of the current intensity of conflict for each instance. This figure allows us to understand interaction effects between these two variables.

logistic regression. It was supposed to have a decreasing effect on conflict. Let us observe how it behaves for the random forest model in Figures 4.7b and 4.7a. The trend is globally the same for NP and SN cases. We observe two main groups. The first one is composed of instances with low Max Value: NP And SN. It has no effect on $\hat{p}$ until it reaches a value 4. From there, there is a big mass of dots with negative effect on $\hat{p}$. From this, we could establish a link with the negative coefficient in logistic regression, but we cannot be certain. For values above 6, the effect starts to grow and becomes positive around 8. The second group is composed of instances with high Max Value: NP and SN. They follow a w-shaped curve with a strong increase at the end. There are some other groups of points that we do not analyse here. What could happen is that the main part of the data points are around the middle of the scale or just above and have slightly negative effect on $\hat{p}$. It could be that since they are numerous, they weighted more than “extreme” instances with high GDP Per Capita* value in the computing of LR weights.

4.3 SHAP to improve model performances

Now that we know which are the most important variables, we could remove the less important ones. This might improve random forest classifier performances by reducing possible overfitting. We tried to remove the 10 and then the 15 less important variables for the 10 repetitions of 10 cross-fold validation. One could argue that the importance of the features is based on one specific model and does not apply to all the 100. This critic is not founded since we have observed that the feature importance order is the same for all models, up to two or three swaps. This experiment resulted in a decrease of performance on the training

sets but also in a decrease of performance on the test sets. That is why we believe that a deeper analysis of the overfitting phenomenon in light of SHAP is needed, and we perform it in the following section.

4.4 SHAP to understand model failures

It is possible to analyse SHAP values for misclassified instances and observe how feature importance is different for these. In this section, we still consider the best performing classifier whose performances are given in section 4.1.3. The misclassified instances are detailed in appendix C. After a thorough analysis, we realised that it was mostly due to an upcoming (in five years) or recent conflict (that finishes the considered year). This is the case for all NP errors. For SN, it is the case for all of them except Indonesia and Sri Lanka. We suppose this two latter have specific patterns. From the decision plots given in Figures C.2 and C.4, we observe that the features' importance order is globally the same as for the general case. For this reason, we suppose that these errors are mainly due to instances for which $\hat{p}$ is near the discriminating threshold t , resulting from ambiguous circumstances.

In a second phase, we can use SHAP on the test set samples in order to determine which features have a good predictive power, compare it to the overall results and deduce which feature might be responsible for overfitting. The summary plots are given in Figures C.1 and C.3 in appendix C. We can observe that the feature importance order does not differ (or only slightly) to the order for all the samples given in Figures 4.3 and 4.4. We deduce that the difference in performance is not due to a specific overfitting phenomenon. We support that they are due to few samples that are in the grey zone and two samples with uncommon patterns. This suggests that there is no need to worry about overfitting. Furthermore, as said previously, we postulate that more data would reduce the number of misclassified instances.

4.5 Chapter Conclusion

SHAP values give us a deep insight into the prediction process. From a black box model, we derived an explanation model which explains precisely the features contributions in the output. It allows us to understand which are the most important features but also to analyse the model failures. From there, we derived the 5 most important variables which are, in decreasing order, Max Value : NP And SN, Structural Constraints, Level Of Repression, Population Size (NP) or Transnational Ethnic Bonds (SN) and Infant Mortality. We have showed that looking at these variables individually is not enough. Interaction effects have an important role. Additionally, we have highlighted that there can be some variations in the most important features depending on the chosen model, for the same dataset. The same information is contained in a set of features. Due to the random part of the random forest algorithm, one feature can have more importance than another depending on the run, but this happens mainly because there is common information in several variables which is

established in Chapter 2. We note that there is a limitation to SHAP computations : it is an estimation for Shapley values and not the exact ones. We believe that the SHAP values contains precious information for the specialists analysing conflict. What could be done in the future is to provide the interaction plots. This is however a time-consuming task since one has to analyse $2^{\binom{23}{2}} = 506$ figures and keep only the most interesting ones. Another improvement is to implement the SHAP interaction values for interventional approach. From this, we could derive interaction matrices. At last, it is possible to cluster the instances based on their SHAP values, it creates groups of x_i whose prediction has been done similarly. This helps detecting cases with unexpected characteristics whose prediction might be less reliable.

Chapter 5

Final discussion and future work

5.1 General discussion and limitations

First, we summarise our contributions and how they followed each other throughout the chapters. In Chapter 2, we highlighted important variables based on the ANOVA test and mutual information. In Chapter 3, we predicted the outcome variable $\mathbf{y}$ based on the input $\mathbf{X}$ using logistic regression and random forest. In Chapter 4, we analysed the variables importance in the random forest prediction model based on SHAP. The core focus of this thesis is the classification technique to predict conflict outbreaks. ANOVA and MI are applied upstream of classification in order to perform this core task in the best way. SHAP is a meta-technique used on the classification model in order to understand how it works and to improve its performances. We will provide here a discussion of all the applied techniques.

In Chapter 2, we evaluated the strength of the links between input variables pairwise or between each input variable and the outcome. The provided techniques do not aim at predicting the outcome, contrarily to the ones used in Chapter 3. There is however a similarity in the sense that, in both chapters, the first method assumes linearity while the second one can handle non-linear effects. In Chapter 3, we also considered feature importance with the logistic regression weights. For all the models, the current intensity of conflict is in the five most important variables. This gives credit to our models since a good predictor for future quarrels is whether there is already a conflict in the country. For NP, we observe that Structural Constraints is always in the five most important features, except for mutual information which generally shows different results than the rest. The Level Of Repression is in the top 5 for all models, both for SN and NP except for LR coefficient (NP) and MI (SN and NP). This is particularly important since this type of indicator is used by other countries to decide on and justify intervention. We see that our models use this predictor intensively. For SN cases, Population Size is in all top five's except for MI. It is particularly interesting to see that the ANOVA and SHAP have the same 5 most important features up to one element. This stresses out the fact, that even if no assumption has been respected, the ANOVA provided us, in this special case, with reliable results that we could cross-check with another model. The most different results are obtained with mutual information. As discussed earlier, this could be due to bad estimations since it is difficult to approximate mutual information in

some cases as explained in section 2.2.2.

Another main difference between the methods arises if we want to do feature selection. If we take the five most significant variables given by ANOVA test or the five variables with the highest mutual information value with the outcome, we do not know whether the information that they contain about the outcome is redundant or not. It is possible that one can obtain a better model with less significant variables that do not carry the same information as the ones in the top five. This happens with ANOVA and mutual information since each variable is considered individually. In logistic regression, the variables are considered simultaneously. Keeping only the variables with highest coefficient (in absolute value) makes more sense than keeping most significant ones or the ones with the highest mutual information with the outcome when there are strong correlations between the variables. It is a bit more complicated with random forest. This is a consequence of the technique consisting in choosing the feature to make the split among a subset of all the features. For this reason, two variables carrying the same information can have equal importance at the end whereas only keeping one of them would not decrease the classifier performances.

The Level Of Repression and the Lack of Democracy are often used to justify international interventions by foreign countries [17]. This happens when some countries want to spread democracy around in non-democratic states and by this way, prevent future conflicts to happen in targetted states, among other things. We observe that the first indicator, the Level of Repression, is one of the most important variables for the random forest model and ANOVA, and in the top ten of most important variables for mutual information and the logistic regression. This means that it is a determining variable to predict future onset. The second variable, the Lack Of Democracy, is often one of the less important variables, except for the logistic regression in case of SN fights. We see that our models do not use it extensively and that it does not bring much information to determine the outcome. If the international intervention is moved by the concern of preventing civil casualties, it appears that the use of the Level of Repression is legitimate. However, it seems that the Lack Of Democracy is not a good indicator for future casualties in a country. The use of this variable to justify international interventions seems questionable to us.

In the light of what we learned from our results, we would recommend using the random forest model which achieves the best performances. For a threshold of $t = 0.5$, the error rate is equal to 4.1% and 5.4% for SN and NP cases. For a sensitivity of 1, the specificity is 0.65 and 0.78. After this, the user can check the results for each country using SHAP values with a decision plot as shown in Chapter 4. This would output a risk map for the world that could be used by policymakers and there would be reliable justifications for each decision. We do not pretend to be qualified in political sciences. This is a trade-off proposed for the domain problems that we understood, based on our analysis. We must keep in mind that such models can only be used as early warning system. It cannot replace people in the analysis of complex phenomena like conflicts. Once the results are computed, the human

analysis can begin and the developed model enables possibly limited teams to focus on most important problems and to highlight cases that might have been unseen.

In the field of conflict prevention, the numbers are often instrumentalised to justify external intervention. Decisions are often taken with cognitive biases resulting from past lived or transmitted experiences. The model has no feelings nor any past trauma. It can be useful in case of highly sensitive cases where it becomes difficult to remain rational. One could argue that the algorithms, because they have no emotions, are unable to predict human behaviour which is often irrational. Though, the model makes few mistakes.

It is not clear whether the results given by the algorithm can be used to suggest prevention policies. We cannot establish a causal relation between the input variables and the predicted phenomenon. It is possible that conflict is the cause of some feature values, that it is a consequence of some other feature values or even that conflicts and the phenomena described by the features have a common cause that is unexpressed.

Our model performs well as it can predict onset in a country in the following four years with high accuracy (given above) but we must keep its limitations in mind. The first one comes from the way the classification task is modelled. It only considers two types of conflicts: fight for national power and subnational quarrel. In these, it only makes the distinction between peace and disputes with more than 25 casualties. This is a crude distinction. A civil war with more than a thousand deceased people has not the same consequences as riots after elections. The second limitation is the detection of anomalies. Some situations are not common or diverge from the rest of the samples. The algorithm has not been trained with such instances and cannot handle them. The prediction it provides for such situations are less reliable. Of course, it has been trained on many samples and has the advantage over the human brain to be able to consider huge quantities of data and to detect patterns that are too complex for a single person, but it is improbable that it can handle well a sample that diverges from the training data it has been fed with.

More generally, our model is limited in terms of interactions and global considerations. Except for Neighbours With HVC and Transnational Ethnic Bonds, there is no consideration for interconnections with other countries. It seems crucial to consider other countries than the neighbouring ones in a globalised world. The events in Europe have an impact on African countries' stability. Moreover, civil wars tend to override national borders [29, 88].

One could argue that our model is limited in terms of resolution, both spatial and temporal. It could possibly be interesting to consider regional resolution instead of a national one. Moreover, considering predictions over the four next years may seem coarse. However, as explained in the historical overview (section 1.1) few models have been able to predict

conflicts precisely on the timescale. Nevertheless, the model is well defined for achieving its goal: building the EU long term foreign policy.

Finally, our implementation is limited in terms of long term performances. It is probable that the conflict patterns change in the 20 years to come. History is not repeating itself perfectly. Since it was trained with samples from 1989, it might be outdated in the future. Therefore, there is a need for an interdisciplinary team that maintains the dataset by adding samples each year and retrain the algorithm over the whole set. It is probable that, if new patterns of conflict appear, the algorithm does not predict them correctly the first times it encounters them. After some training on samples depicting these situations, this can be rectified.

5.2 Future work

The possible future work is considerable: we are only at the beginning and a lot of improvements are possible.

The first one is certainly to better define conflicts. One could imagine having several models. Another idea is to apply regression methods to determine the number of casualties. One might also use classification in order to determine the type of conflict: genocide, civil war, coup. One could also distinguish the conflicts based on their consequences: types of buildings and infrastructure destructed, resulting instabilities... More generally, one may develop different models for the starting time, the duration and the geographical spread. One could consider emerging conflict differently from lasting conflict. Alternatively, one could consider several years of data for the prediction to grasp the evolution of the variables and not only their values. Obviously, more data is needed. It exists but is not always accessible or needs a big preparation time.

In order to assess prediction performance on a new sample, one can perform anomalies detection. It consists in determining what is the probability of arising of the considered sample x_i , respectively to the background data $\mathbf{X}$ we dispose of. It must be clearly distinguished from the classification task where we estimate a probability for the outcome y_i given input sample x_i . Here, we consider that x_i is a random vector drawn from an unknown probability distribution from which we have n realisations given in matrix $\mathbf{X}$. From $\mathbf{X}$, we estimate the joint probability distribution, and we derive a bound for probability for x_i to be drawn. If the probability is large, then the new sample is similar to the data the model has been trained with, and we can reasonably trust the prediction. If the probability is low, then x_i is probably not similar to the training data, and we must consider the prediction carefully.

One should certainly better consider the interactions between the variables. It can be done using multivariate ANOVA to test for joint statistical significance. We also recommend computing SHAP interaction values and draw interaction plots. At last, clustering instances

based on SHAP values would help to distinguish groups of samples that are predicted in the same way.

Concerning data, we have observed in Chapter 3 that more samples would increase our model performances. The extension to next years of the dataset we use exists and is already prepared but the GCRI team which we have contacted does not want to release it.

As explained in the variable discussion, new indicators could improve the prediction: population density (over usable land) is related to conflict. The type of past violence, how and whether it has been solved also gives insight into future quarrels. The prediction is now based solely on one year of data. One could take the evolution in account and extend vector x to have an input of several years.

We would be interested into the development of new tools using graph theory. The pertinence of networks used for the modelling of international relations, which have an impact on conflicts has been established in [33, 60]. Furthermore, it has been observed that many civil wars have transnational dimensions [29, 88]. Graph neural network (GNN) could provide better predictions since it takes globalised interactions into account. This technique starts from a network with attributes for each node and embeds it in order to provide a prediction for each node using a neural network. There are, however, two main difficulties. One needs to build a worldwide interaction network and to quantify the link between each pair of countries, which is a considerable amount of work, far from easy to perform. The second problem is that it is not clear whether there is enough data for the graph neural network to perform well.

5.3 Conclusion

We now review the thesis as a whole and provide the reader with our final thoughts. We wanted to predict unrest or peace in the next four years to come based on yearly measured national indicators. First, we computed variables importance upstream of the prediction model. Then, we established the prediction models and compared them. Last, we analysed the output of the prediction models, downstream, with a meta-tool.

With this thesis, we showed that there are machine learning tools available to answer both the need for accurate and interpretable predictions that are the key for efficient future long-term foreign policies. Although we only highlighted the tools in machine learning, the future work is promising and there are many more techniques in applied mathematics that we did not talk about.

In addition, we observed that some of the indicators that are used to justify international intervention by a foreign country are used extensively by our models (Level of Repression)

while others are not (Lack of Democracy).

Next to the technical analysis performed in this thesis, we have realised that the future of this research field depends on the building of interdisciplinary teams. Until now, the attempt to make various profiles gather are seldom. There is a need for engineers, social scientists and policymakers to work together on such projects. The success of this challenging undertaking depends on the commitment of the three profiles. Furthermore, it is not enough to create these teams. One has to make them work well together which is far from easy. There is a need for pedagogic explanations in order to build mutual understanding. This might be the difficult part but, in our opinion, it is the one that makes sense and that is the most interesting since we can learn from each other and think with new paradigms.

Appendix A

Variables origins and details

Here, more details about the variables is available.

N.B: Ethnic power change is only used for national power and ethnic compilation is only used for subnational fights.

The conflict datasets used:

- The battle related death dataset (BRD) [51]
- The one-sided violence dataset (OSV) [66] and [18]
- The non-state conflict dataset (NSC) [66] and [83]

The other indicators datasets:

- POLITY2 database [32]
- World Bank's Worldwide Governance Indicators [85]
- Databases from Amnesty International, Human Right Watch and US state department of political scale terror [27] and [37]
- Cingranelli and Richards (CIRI) Human Rights Data Project [11] and [10]
- The conflict barometer [4]
- Ethnic Power Relations (EPR) Core Dataset [92] and [91]
- Minorities at Risk Dataset [61]
- Standardized World Income Inequality Database [80]
- Food Security Indicators database [21]
- Aqueduct Country and River Basin Rankings [26]
- Bertelsmann Stiftung's Transformation Index [44]
- UN DESA's World Population Prospects [16]

Variable	Variable name in the code	Source	Name of dataset	Name of original indicator
Regime type	REG_U	Center for systemic peace	Polity IV Annual Time-Series, 1800-2015	PARCOM, EXREC
Lack of Democracy	REG_P2	Center for systemic peace	Polity IV Annual Time-Series, 1800-2015	POLITY2
Government effectiveness	GOV_EFF	World bank	Government Effectiveness: Estimate	GE.EST
Level of repression	REPRESS	Political terror scale	PTS Data	Highest of the three indicators in the set (PTS_A, PTS_H, PTS_S)
Empowerment Rights	EMPOWER	CIRI Human Righths Data Project	CIRI Data	NEW_EMPINX
Recent internal conflict	CON_INT	HIIK; UCDP/PRIO	Battle related deaths, One-sided violence, Non-state conflict, Conflict Barometer 2016	Highest casualty estimates
Neighbours with highly violent conflict	CON_NB	HIIK; UCDP/PRIO	Battle related deaths, One-sided violence, Non-state conflict, Conflict Barometer 2016	Highest casualty estimates
Years since highly violent conflict	YRS_HVC	HIIK; UCDP/PRIO	Armed Conflict Dataset, Conflict Barometer 2016	Conflicts of intensity level 2
Corruption	CORRUPT	World Bank	Control of Corruption: Estimate	CC.EST
Ethnic Power Change	ETHNIC_NP	ETH Zurich	EPR Core Dataset	Recording of dataset
Ethnic Compilation	ETHNIC_SN	ETH Zurich	EPR Core Dataset	Recording of dataset,
Transnational Ethnic Bonds	DISPER	CIDCM Center for International Development & Conflict Management	Marupdate20042006	GC10
Homicide Rate	HOMIC	World Bank	World Development Indicators	Intentional homicides (per 100,000 people)
Infant Mortality	MORT	World Bank	World Development Indicators	Mortality rate, under-5 (per 1,000 live births)
GDP per capita	GDP	World Bank	World Development Indicators	GDP per capita, PPP (constant 2011 international \$)
Income Inequality	INEQ_SWIID	Harvard Dataverse Network	The Standardized World Income Inequality Database	Net inequality
Economic openness	ECON_ISO	World Bank	World Development Indicators	Foreign direct investment, net inflows (BoP, current US\$)
Food Security	FOOD	FAO	Food security indicators	Foreign direct investment, net inflows (% of GDP)
Unemployment	UNEMP	World Bank	World Development Indicators	Exports of goods and services (% of GDP)
Water Stres	WATER	World Resources Institute	Aqueduct Country and River Basin Rankings (Raw country scores)	tdefm
Oil Production	FUEL_EXP	World Bank	World Development Indicators	Fuel exports (% of merchandise exports)
Structural Constraints	STRUCT	BTI: The Bertelsmann Stiftung	BTI 2016	Structural constrains (Q13.1)
Population Size	POP	UN DESA/ Population Division	Annual population by single age - Both Sexes.	Sum of all ages
Youth Bulge	YOUTHBOTH	UN DESA/ Population Division	Annual population by single age - Both Sexes.	Sum of ages 15-24 divided by sum of ages 25+

Table A.1: Data origin of each variable

For detailed information about the treatment of the variables, please refer to [36] and [35].

In Table A.2, the countries that do not cover the whole period are transcribed. In the third column, the reader can find the number of exploitable years which corresponds to the full range minus the four last years. Since the prediction goes over four years, no conclusion can be taken from 2011 to 2014.

Country	Begin	End	Exploitable years	Country	Begin	End	Exploitable years
Azerbaïdjan	1991	2014	20	Montenegro	2006	2014	5
Belarus	1991	2014	20	Namibia	1990	2014	21
Bosnia and Herzegovina	1992	2014	19	Russian Federation	1991	2014	20
Eritrea	1993	2014	18	Serbia	2006	2014	5
Georgia	1991	2014	20	Timor-Leste	2002	2014	9
Kazakhstan	1991	2014	20	Turkmenistan	1991	2014	20
Kyrgyzstan	1991	2014	20	Ukraine	1991	2014	20
Macedonia, the former Yugoslav Republic of	1991	2014	20	Uzbekistan	1991	2014	20
Moldova, Republic of	1991	2014	20				

Table A.2: Years of data for countries in the dataset

Appendix B

Algorithms of Chapter 2

In this appendix, we display the algorithm for performing the normality (Algorithm 7) and homoscedasticity (Algorithm 8) tests.

Algorithm 7: Shapiro-Wilk

input : the vector $u = (u_1, \dots, u_n)$ for which we want to test normality

output: the p -value given by the Shapiro-Wilk test

- 1 Define significance level α
 - 2 $\bar{u} \leftarrow$ the mean of u
 - 3 $u' \leftarrow$ ordered u by putting its elements in increasing order
 - 4 Compute m which is equal to the element-wise expectation of an ordered vector of size n which elements have been drawn from $\mathcal{N}(0, 1)$.
 - 5 Compute V , the covariance matrix of an ordered vector of size n which elements have been drawn from $\mathcal{N}(0, 1)$
 - 6 $a \leftarrow \frac{m^T V^{-1} m}{\|V^{-1} m\|_2}$
 - 7 $W \leftarrow \frac{(\sum_{i=1}^n a_i u'_i)^2}{\sum_{i=1}^n (u_i - \bar{u})^2}$
 - 8 Compute p -values using Monte-Carlo simulation from test statistic W
 - 9 **if** $p \leq \alpha$ **then**
 - 10 | reject H_0 , this implies that normality is rejected
 - 11 **else**
 - 12 | reject H_a
 - 13 **end**
-

Algorithm 8: Brown-Forsythe

input : the vectors $x_{\cdot j}^{(0)}$ and $x_{\cdot j}^{(1)}$ of size n_0 and n_1 , respectively

output: The p -value given by the Brown-Forsythe test

```
1 Define significance level  $\alpha$ 
2 Define  $v_{\cdot j}^{(1)}$  and  $v_{\cdot j}^{(1)}$ , two vectors of size  $n_0$  and  $n_1$ , respectively
3  $K \leftarrow 2$ 
4  $\tilde{x}_j^{(0)} \leftarrow$  the median of  $x_{\cdot j}^{(0)}$ 
5  $\tilde{x}_j^{(1)} \leftarrow$  the median of  $x_{\cdot j}^{(1)}$ 
6 for  $k = 0 : K - 1$  do
7   for  $i = 1 : n_k$  do
8      $v_{ij}^{(k)} = |x_{ij}^{(k)} - \tilde{x}_j^{(k)}|$ 
9   end
10 end
11  $F \leftarrow \frac{n-K}{K-1} \frac{\sum_{k=0}^{K-1} n_k (\bar{v}_j^{(k)} - \bar{v}_j)^2}{\sum_{k=0}^{K-1} \sum_{i=1}^{n_k} (v_{ij}^{(k)} - \bar{v}_j^{(k)})^2}$ 
12 compute the  $p$ -value from  $F$ 
13 if  $p \leq \alpha$  then
14   | reject  $H_0$ , this implies that homoscedaticity is rejected
15 else
16   | reject  $H_a$ ,
17 end
```

Appendix C

Additional figures for Chapter 4

In this appendix, the additional plots related to chapter 4 are available. On Figures C.2 and C.4, the decision plots for misclassified instances are given for NP and SN, respectively.

For NP, from lowest predicted probability to the highest the concerned countries-years are:

- the former Yugoslav republic of Macedonia - 1997: as $\hat{p} = 0.215$, this case is a false negative
- Ivory Coast - 2007: as $\hat{p} = 0.499$, this case is a false negative.
- Congo (Brazzaville) - 2002: as $\hat{p} = 0.56$, this case is a false positive
- Mexico - 1997: as $\hat{p} = 0.615$, it is a false positive.

For SN, from lowest predicted probability to highest, the concerned country years are:

- Zimbabwe - 2004: as $\hat{p} = 0.365$, it is a false negative.
- Sri Lanka - 2002: as $\hat{p} = 0.445$, it is a false negative.
- Central African Republic - 2002: as $\hat{p} = 0.45$, it is a false negative.
- Indonesia - 1998: as $\hat{p} = 0.455$, it is a false negative.
- Uganda - 1992: as $\hat{p} = 0.505$, it is a false positive.
- Burundi - 2008: as $\hat{p} = 0.535$, it is a false positive.
- Papua New Guinea - 2007: as $\hat{p} = 0.59$, it is a false positive.
- Canada - 1998: as $\hat{p} = 0.64$, it is a false positive.

On Figures C.1 and C.3, the summary plots for test sets are given for NP and SN, respectively.

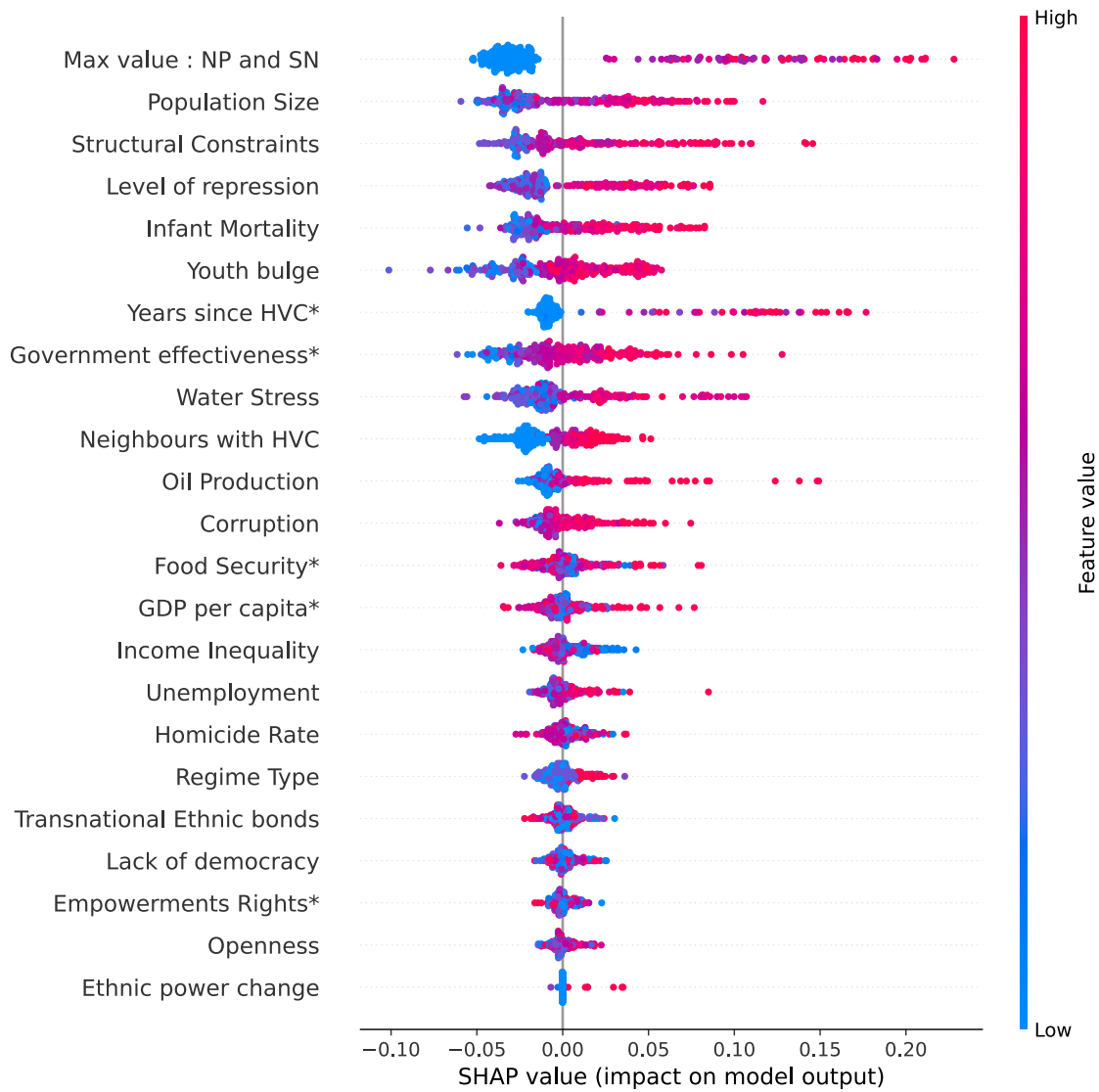


Figure C.1: Summary plot for NP test set. For each line, a dot corresponds to an instance and the colour to the feature value. The dot abscissa corresponds to the SHAP value for the feature displayed on the ordinate.

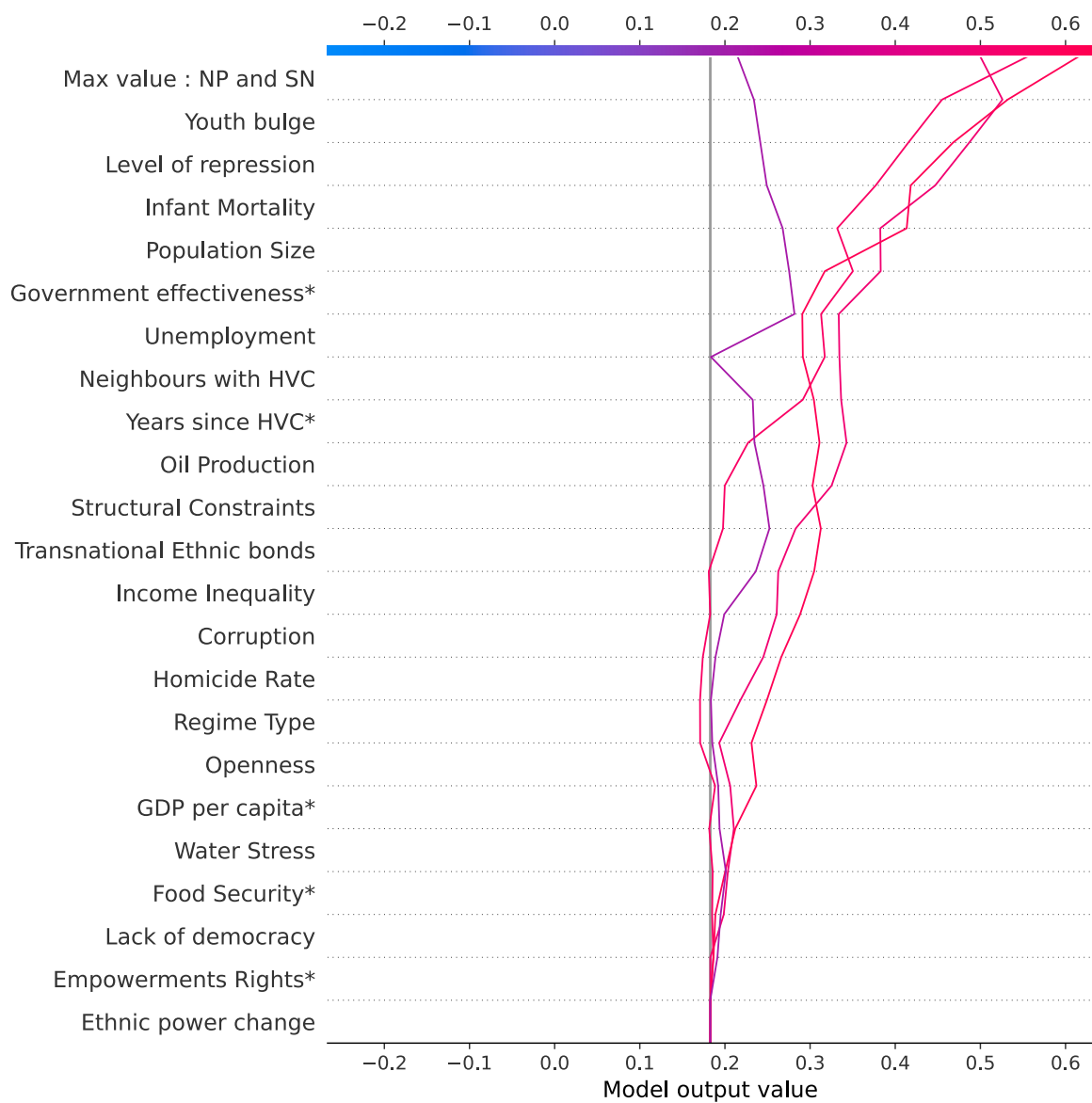


Figure C.2: Decision plot for misclassified instances (NP). The line starts from the expected value for the classifier which is 0.183. It then follows a trajectory according to the sum of the SHAP values displayed on the y -axis.

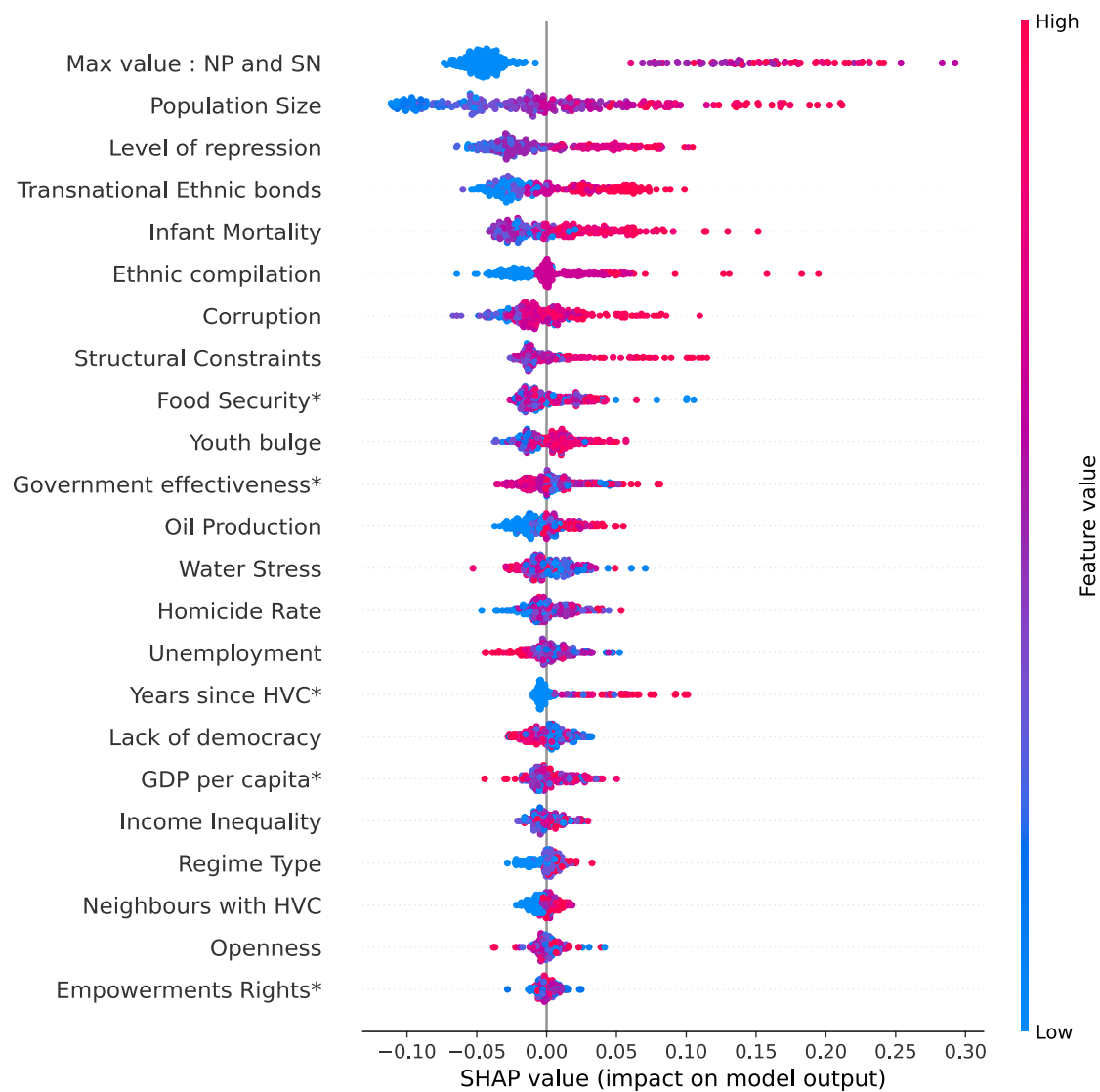


Figure C.3: Summary plot for SN test set. For each line, a dot corresponds to an instance and the colour to the feature value. The dot abscissa corresponds to the SHAP value for the feature displayed on the ordinate.

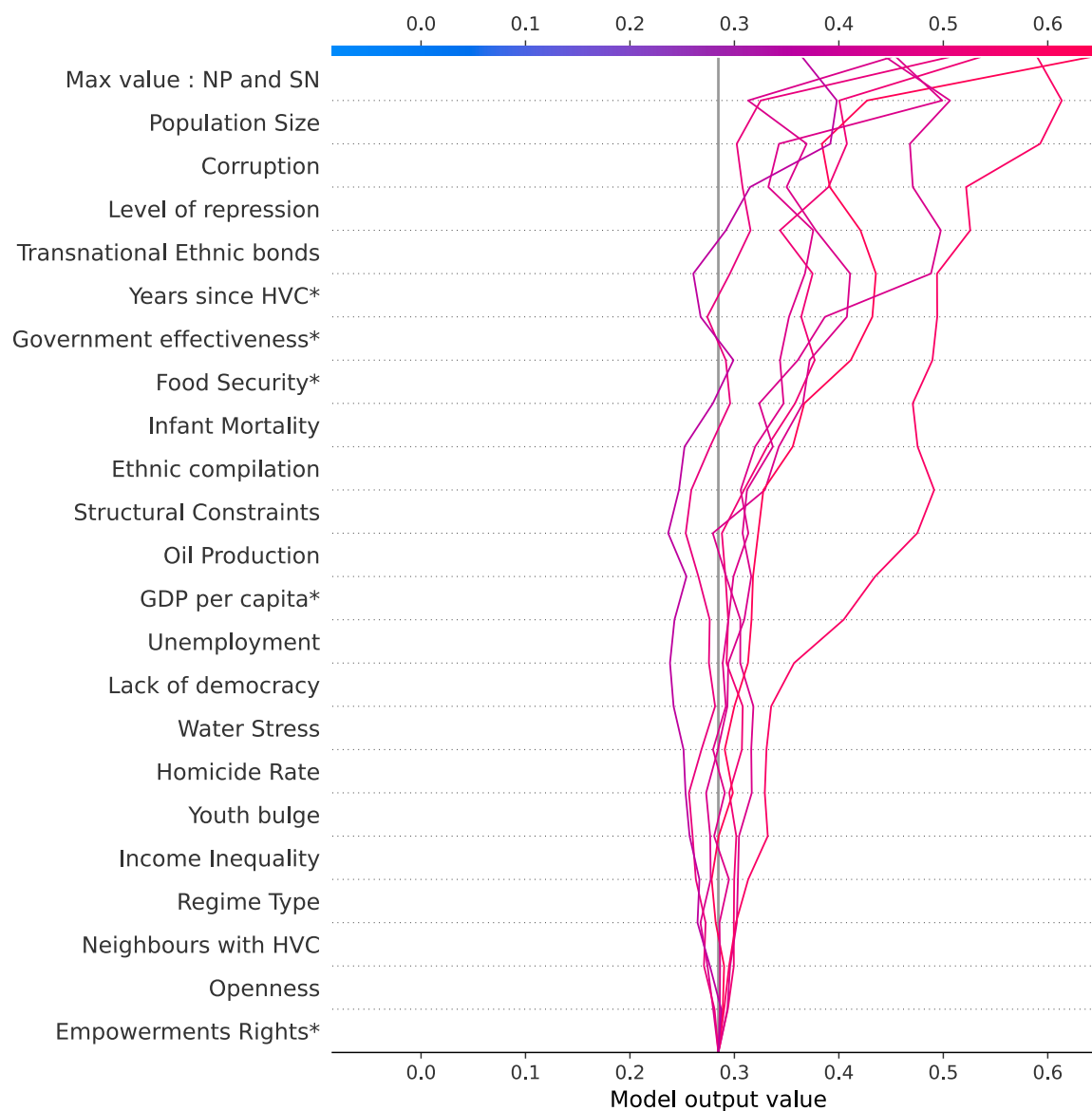


Figure C.4: Decision plot for misclassified instances (SN). The line starts from the expected value for the classifier which is 0.28. It then follows a trajectory according to the sum of the SHAP values displayed on the *y*-axis.

References

- [1] P. D. Allison, ‘Convergence failures in logistic regression,’ in *SAS Global Forum*, vol. 360, 2008, pp. 1–11 (cit. on p. 54).
- [2] B. R. Amin, ‘Cyprus conflict: How could be resolved and reunified?’ *Modern Diplomacy*, May 2021 (cit. on p. 11).
- [3] S. Arlot, A. Celisse *et al.*, ‘A survey of cross-validation procedures for model selection,’ *Statistics surveys*, vol. 4, pp. 40–79, 2010 (cit. on p. 48).
- [4] C. Barometer, ‘Disputes non-violent crises violent crises limited wars wars, 26,’ *Heidelberg Institute for International Conflict Research, Heidelberg*, 2018 (cit. on p. 108).
- [5] N. Beck, G. King and L. Zeng, ‘Improving quantitative studies of international conflict: A conjecture,’ *American Political science review*, pp. 21–35, 2000 (cit. on p. 6).
- [6] C. Blattman, O. Dube, S. Bazzi, M. Gudgeon, R. Peck and R. Blair, ‘The promise and pitfalls of conflict prediction: Evidence from colombia and indonesia,’ 2019 (cit. on p. 6).
- [7] M. Borda, *Fundamentals in information theory and coding*. Springer Science & Business Media, 2011 (cit. on p. 30).
- [8] R. H. Byrd, P. Lu, J. Nocedal and C. Zhu, ‘A limited memory algorithm for bound constrained optimization,’ *SIAM Journal on scientific computing*, vol. 16, no. 5, pp. 1190–1208, 1995 (cit. on p. 58).
- [9] L.-E. Cederman and N. B. Weidmann, ‘Predicting armed conflict: Time to adjust our expectations?’ *Science*, vol. 355, no. 6324, pp. 474–476, 2017 (cit. on p. 6).
- [10] D. L. Cingranelli, ‘Short variable descriptions for indicators in the cingranelli-richards (ciri) human rights dataset,’ *Document ver*, vol. 5, p. 14, 2014 (cit. on p. 108).
- [11] D. L. Cingranelli, D. L. Richards and K. C. Clay, ‘The ciri human rights dataset,’ 2014 (cit. on p. 108).
- [12] J. Cohen, ‘A coefficient of agreement for nominal scales,’ *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960. DOI: [10.1177/001316446002000104](https://doi.org/10.1177/001316446002000104).

- eprint: <https://doi.org/10.1177/001316446002000104>. [Online]. Available: <https://doi.org/10.1177/001316446002000104> (cit. on p. 46).
- [13] C. Croarkin, P. Tobias, J. Filliben, B. Hembree, W. Guthrie *et al.*, ‘Nist/sematech e-handbook of statistical methods,’ *NIST/SEMATECH, July*. Available online: <http://www.itl.nist.gov/div898/handbook>, 2006 (cit. on p. 24).
 - [14] T. De Groeve, H. Peter and L. Vernaccini, ‘The global conflict risk index (gcri) regression model: Concept and methodology,’ *Luxembourg: Publications Office of the European Union*, 2014 (cit. on p. 16).
 - [15] B. Derrick, A. Ruck, D. Toher and P. White, ‘Tests for equality of variances between two samples which contain both paired observations and independent observations,’ *Journal of Applied Quantitative Methods*, vol. 13, no. 2, pp. 36–47, 2018 (cit. on p. 24).
 - [16] U. DESA, ‘World population prospects 2019. online edition, rev 1, united nations, department of economic and social affairs,’ *Population Division*, 2019 (cit. on p. 108).
 - [17] A. B. Downes and J. Monten, ‘Forced to be free?: Why foreign-imposed regime change rarely leads to democratization,’ *International Security*, vol. 37, no. 4, pp. 90–131, 2013 (cit. on p. 102).
 - [18] K. Eck and L. Hultman, ‘One-sided violence against civilians in war: Insights from new fatality data,’ *Journal of Peace Research*, vol. 44, no. 2, pp. 233–246, 2007 (cit. on p. 108).
 - [19] R. A. Fisher, ‘Xv.—the correlation between relatives on the supposition of mendelian inheritance.,’ *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, vol. 52, no. 2, pp. 399–433, 1919 (cit. on p. 22).
 - [20] R. A. Fisher *et al.*, ‘Statistical methods for research workers.,’ *Statistical methods for research workers.*, no. 5th Ed, 1934 (cit. on p. 23).
 - [21] *Food security indicators*, <http://www.fao.org/economic/ess/ess-fs/ess-fadata/en/> (cit. on p. 108).
 - [22] J. Friedman, T. Hastie, R. Tibshirani *et al.*, *The elements of statistical learning*, 10. Springer series in statistics New York, 2001, vol. 1 (cit. on pp. 71, 72).
 - [23] K. Fujimoto, I. Kojadinovic and J.-L. Marichal, ‘Axiomatic characterizations of probabilistic and cardinal-probabilistic interaction indices,’ *Games and Economic Behavior*, vol. 55, no. 1, pp. 72–99, 2006 (cit. on p. 89).
 - [24] J. Fürnkranz, J. Petrak, R. Trappl and J. Bercovitch, ‘Machine learning methods for international conflict databases: A case study in predicting mediation outcome,’ 1994 (cit. on p. 7).

- [25] W. Gao, S. Kannan, S. Oh and P. Viswanath, ‘Estimating mutual information for discrete-continuous mixtures,’ *arXiv preprint arXiv:1709.06212*, 2017 (cit. on p. 33).
- [26] F. Gassert, P. Reig, T. Luo and A. Maddocks, ‘A weighted aggregation of spatially distinct hydrological indicators,’ *World Resources Institute*, 2013 (cit. on p. 108).
- [27] M. Gibney, L. Cornett, R. Wood, P. Haschke, D. Arnon, A. Pisanò and G. Barrett, ‘The political terror scale 1976-2018,’ *Date Retrieved, from the Political Terror Scale website (<http://www.politicalterrorscale.org/>)*, 2019 (cit. on p. 108).
- [28] C. Gini, *Variabilità e mutabilità*. 1912 (cit. on p. 15).
- [29] K. S. Gleditsch, ‘Transnational dimensions of civil war,’ *Journal of peace research*, vol. 44, no. 3, pp. 293–309, 2007 (cit. on pp. 103, 105).
- [30] K. S. Gleditsch and M. D. Ward, ‘Forecasting is difficult, especially about the future: Using contentious issues to forecast interstate disputes,’ *Journal of Peace Research*, vol. 50, no. 1, pp. 17–31, 2013 (cit. on p. 6).
- [31] R. M. Gray, *Entropy and information theory*. Springer Science & Business Media, 2011 (cit. on p. 30).
- [32] T. R. Gurr, M. G. Marshall and K. Jagers, ‘Polity v project: Political regime characteristics and transitions, 1800-2018, dataset users,’ *Manual*. Center for Systemic Peace, 2020 (cit. on p. 108).
- [33] E. M. Hafner-Burton, M. Kahler and A. H. Montgomery, ‘Network analysis for international relations,’ *International organization*, pp. 559–592, 2009 (cit. on p. 105).
- [34] M. Halkia, S. Ferri, I. Joubert-Boitat and F. Saporiti, ‘Conflict risk indicators: Significance and data management in the gcric,’ *Luxembourg: Publications Office of the European Union*, 2017 (cit. on pp. 12, 14, 16).
- [35] M. Halkia, S. Ferri, M. K. Schellens, M. Papazoglou and D. Thomakos, ‘The global conflict risk index: A quantitative tool for policy support on conflict prevention,’ *Progress in Disaster Science*, p. 100 069, 2020 (cit. on pp. 6, 8, 11, 12, 17, 29, 46, 47, 58, 75, 77, 110).
- [36] S. Halkia, S. Ferri, I. Joubert-Boitat, F. Saporiti and M. Kauffmann, ‘The global conflict risk index (gcric) regression model: Data ingestion, processing, and output methods,’ *Luxembourg: Publications Office of the European Union*, 2017 (cit. on pp. 12, 110).
- [37] P. Haschke, ‘The political terror scale (pts) codebook,’ *University Of North Carolina, Asheville. Gathered*, vol. 25, 2017 (cit. on p. 108).

- [38] H. Hegre, M. Allansson, M. Basedau, M. Colaresi, M. Croicu, H. Fjelde, F. Hoyles, L. Hultman, S. Höglbladh, R. Jansen *et al.*, ‘Views: A political violence early-warning system,’ *Journal of peace research*, vol. 56, no. 2, pp. 155–174, 2019 (cit. on p. 6).
- [39] H. Hegre, N. W. Metternich, H. M. Nygård and J. Wucherpfennig, *Introduction: Forecasting in peace research*, 2017 (cit. on p. 5).
- [40] J. I. Hoffman, *Biostatistics for medical and biomedical practitioners*. Academic press, 2015 (cit. on p. 57).
- [41] D. W. Hosmer Jr, S. Lemeshow and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013, vol. 398 (cit. on pp. 42, 59, 75).
- [42] H. W. Ian and F. Eibe, *Data mining: Practical machine learning tools and techniques*, 2005 (cit. on p. 31).
- [43] C. Iceland, ‘How to solve water-related conflicts,’ (cit. on p. 34).
- [44] B. T. Index, *Bertelmannstiftung 2018* (cit. on p. 108).
- [45] G. James, D. Witten, T. Hastie and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013, vol. 112 (cit. on pp. 9, 71).
- [46] D. Janzing, L. Minorics and P. Blöbaum, ‘Feature relevance quantification in explainable ai: A causal problem,’ in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2020, pp. 2907–2916 (cit. on pp. 86, 87).
- [47] L. A. Jeni, J. F. Cohn and F. De La Torre, ‘Facing imbalanced data—recommendations for the use of performance metrics,’ in *2013 Humaine association conference on affective computing and intelligent interaction*, IEEE, 2013, pp. 245–251 (cit. on pp. 43, 46).
- [48] R. Jungers, *Linma2345 - game theory*, Apr. 2021 (cit. on pp. 81–83).
- [49] J. KORTLANDT, J. JULIA KARLYSHEVA, M. TIGNINO and C. PELLATON, ‘Water for peace: Preventing conflict related to water and wetlands,’ *Development for Peace*, Aug. 2020 (cit. on p. 34).
- [50] A. Kraskov, H. Stögbauer and P. Grassberger, ‘Estimating mutual information,’ *Physical review E*, vol. 69, no. 6, p. 066 138, 2004 (cit. on p. 33).
- [51] B. Lacina and N. P. Gleditsch, ‘Monitoring trends in global combat: A new dataset of battle deaths,’ *European Journal of Population/Revue Européenne de Démographie*, vol. 21, no. 2, pp. 145–166, 2005 (cit. on p. 108).
- [52] J. R. Landis and G. G. Koch, ‘The measurement of observer agreement for categorical data,’ *biometrics*, pp. 159–174, 1977 (cit. on pp. 46, 47, 59, 75).

- [53] J. L. LAWLESS, ‘Explainer: What is behind the latest unrest in n ireland?’ *AP News*, Apr. 2021 (cit. on p. 11).
- [54] M. van Leeuwen, A. Ansoms, E. M. Mudinga, A. N. Bisoka, R.-C. Niyonkuru, J. Shaw and G. van der Haar, ‘Promoting land tenure security for sustainable peace—lessons on the politics of transformation,’ *Current Opinion in Environmental Sustainability*, vol. 49, pp. 57–65, 2021 (cit. on p. 12).
- [55] S. Lundberg and S.-I. Lee, ‘A unified approach to interpreting model predictions,’ *CoRR*, vol. abs/1705.07874, 2017. arXiv: 1705.07874. [Online]. Available: <http://arxiv.org/abs/1705.07874> (cit. on pp. 83, 85–87).
- [56] S. M. Lundberg, G. G. Erion and S.-I. Lee, ‘Consistent individualized feature attribution for tree ensembles,’ *arXiv preprint arXiv:1802.03888*, 2018 (cit. on p. 89).
- [57] S. M. Lundberg and S.-I. Lee, ‘Consistent feature attribution for tree ensembles,’ *arXiv preprint arXiv:1706.06060*, 2017 (cit. on pp. 86, 87).
- [58] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal and S.-I. Lee, ‘From local explanations to global understanding with explainable ai for trees,’ *Nature Machine Intelligence*, vol. 2, no. 1, pp. 2522–5839, 2020 (cit. on pp. 89, 90).
- [59] *Machine learning tutorials*, <https://www.mltut.com/k-fold-cross-validation-in-machine-learning-how-does-k-fold-work/> (cit. on p. 49).
- [60] Z. Maoz, L. G. Terris, R. D. Kuperman and I. Talmud, ‘International relations: A network approach,’ *New directions for international relations: confronting the method-of-analysis problem*, pp. 35–64, 2005 (cit. on p. 105).
- [61] *Minorities at risk dataset*, [http://www.mar.umd.edu/#:~:text=The%20Minorities%20at%20Risk%20\(MAR,population%20of%20at%20least%20500%2C000](http://www.mar.umd.edu/#:~:text=The%20Minorities%20at%20Risk%20(MAR,population%20of%20at%20least%20500%2C000). (cit. on p. 108).
- [62] C. Molnar, *Interpretable Machine Learning, A Guide for Making Black Box Models Explainable*. 2019, <https://christophm.github.io/interpretable-ml-book/> (cit. on pp. 83, 88).
- [63] A. Ng, ‘Cs229 lecture notes,’ *CS229 Lecture notes*, vol. 1, no. 1, pp. 1–3, 2000 (cit. on p. 52).
- [64] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, ‘Scikit-learn: Machine learning in Python,’ *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011 (cit. on pp. 33, 81).

- [65] C. Perry, ‘Machine learning and conflict prediction: A use case,’ *Stability: International Journal of Security and Development*, vol. 2, no. 3, p. 56, 2013 (cit. on p. 6).
- [66] T. Pettersson and M. Öberg, ‘Organized violence, 1989–2019,’ *Journal of peace research*, vol. 57, no. 4, pp. 597–613, 2020 (cit. on p. 108).
- [67] H. Pierre, *La revanche des passions. métamorphoses de la violence et crises du politique*, 2015 (cit. on p. 1).
- [68] J. P. Pluim, J. A. Maintz and M. A. Viergever, ‘Mutual-information-based registration of medical images: A survey,’ *IEEE transactions on medical imaging*, vol. 22, no. 8, pp. 986–1004, 2003 (cit. on p. 30).
- [69] P. Ranganathan, C. Pramesh and R. Aggarwal, ‘Common pitfalls in statistical analysis: Logistic regression,’ *Perspectives in clinical research*, vol. 8, no. 3, p. 148, 2017 (cit. on p. 57).
- [70] V. Rosoux, ‘Réconciliation post conflit: À la recherche d’une typologie,’ *Revue internationale de politique comparée*, vol. 22, no. 4, pp. 557–577, 2015 (cit. on p. 12).
- [71] —, ‘Memory, cultural heritage, and legacies of wars,’ in *Routledge Handbook of Peace, Security and Development*, Routledge, 2020, pp. 226–238 (cit. on p. 1).
- [72] V. Rosoux, ‘Time and reconciliation: Dealing with festering wounds,’ in *Negotiating Reconciliation in Peacemaking*, Springer, 2017, pp. 69–91 (cit. on p. 1).
- [73] B. C. Ross, ‘Mutual information between discrete and continuous data sets,’ *PloS one*, vol. 9, no. 2, e87357, 2014 (cit. on p. 33).
- [74] T. Roughgarden and G. Valiant, ‘Cs168: The modern algorithmic toolbox lecture# 6: Regularization,’ 2017 (cit. on p. 55).
- [75] D. Sabine, H. Hartmann, R. Schwarz and S. Steinkamp, *Bertelsmann Stiftung Transformation Index, codebook*. BTI, 2020 (cit. on p. 16).
- [76] H. Sahai, M. I. Ageel and A. Khurshid, *Analysis of variance model, effects of departures from assumptions underlying*. 2011 (cit. on p. 26).
- [77] T. Saito and M. Rehmsmeier, ‘The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets,’ *PloS one*, vol. 10, no. 3, e0118432, 2015 (cit. on p. 45).
- [78] F. des Sciences de l’université de Namur, *Smatm102, analyse multivariée et logiciels statistiques*, 2016-2017 (cit. on pp. 23, 26).
- [79] S. S. Shapiro and M. B. Wilk, ‘An analysis of variance test for normality (complete samples),’ *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965 (cit. on p. 23).

- [80] F. Solt, ‘The standardized world income inequality database,’ *Social science quarterly*, vol. 97, no. 5, pp. 1267–1281, 2016 (cit. on p. 108).
- [81] H. Strand, S. A. Rustad, H. Urdal and H. M. Nygård, ‘Trends in armed conflict, 1946–2018,’ *Conflict trends*, vol. 3, p. 2019, 2019 (cit. on p. 1).
- [82] H. Strand, L. Wilhelmsen and N. P. Gleditsch, ‘Armed conflict dataset codebook,’ *PRIO: Oslo*, 2003 (cit. on p. 8).
- [83] R. Sundberg, K. Eck and J. Kreutz, ‘Introducing the ucdp non-state conflict dataset,’ *Journal of peace research*, vol. 49, no. 2, pp. 351–362, 2012 (cit. on p. 108).
- [84] R. Sundberg, K. Eck, J. Kreutz *et al.*, ‘Ucdp non-state conflict codebook version 20.1,’ 2012 (cit. on p. 8).
- [85] *The world bank. worldwide governance indicators. data cat 2019*. <https://datacatalog.worldbank.org/dataset/worldwide-governance-indicators> (cit. on pp. 14, 108).
- [86] R. Trappl, *Programming for peace: computer-aided methods for international conflict resolution and prevention*. Springer Science & Business Media, 2006, vol. 2 (cit. on p. 6).
- [87] R. Trappl, J. Fürnkranz and J. Petrak, ‘Digging for peace: Using machine learning methods for assessing international conflict databases,’ in *ECAI*, PITMAN, 1996, pp. 453–457 (cit. on pp. 5, 6, 8).
- [88] N. Twagiramungu, A. Duursma, M. G. Berhe and A. de Waal, ‘Re-describing transnational conflict in africa,’ *The Journal of Modern African Studies*, vol. 57, pp. 377–391, 2019 (cit. on pp. 103, 105).
- [89] ‘Uppsala conflict data program,’ *UCDP Conflict Encyclopedia*, 2021 (cit. on p. 8).
- [90] J. R. Vergara and P. A. Estévez, ‘A review of feature selection methods based on mutual information,’ *Neural computing and applications*, vol. 24, no. 1, pp. 175–186, 2014 (cit. on p. 32).
- [91] M. Vogt, *The ethnic power relations (epr) core dataset 2014: Codebook*, 2014 (cit. on p. 108).
- [92] M. Vogt, N.-C. Bormann, S. Rüegger, L.-E. Cederman, P. Hunziker and L. Girardin, ‘Integrating data on ethnicity, geography, and conflict: The ethnic power relations data set family,’ *Journal of Conflict Resolution*, vol. 59, no. 7, pp. 1327–1342, 2015 (cit. on p. 108).
- [93] J. R. Vreeland, ‘The effect of political regime on civil war: Unpacking anocracy,’ *Journal of conflict Resolution*, vol. 52, no. 3, pp. 401–425, 2008 (cit. on p. 15).

- [94] S. J. Wright, ‘Coordinate descent algorithms,’ *Mathematical Programming*, vol. 151, no. 1, pp. 3–34, 2015 (cit. on p. [60](#)).
- [95] B. W. Yap and C. H. Sim, ‘Comparisons of various types of normality tests,’ *Journal of Statistical Computation and Simulation*, vol. 81, no. 12, pp. 2141–2155, 2011 (cit. on p. [23](#)).

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

École polytechnique de Louvain

Rue Archimède, 1 bte L6.11.01, 1348 Louvain-la-Neuve, Belgique | www.uclouvain.be/epl