Analysis of ethnic groups characteristics and their relation

Weijie Jin, Martina Soldini, Yu Yang

May 2019

Contents

1 Introduction

The purpose of this work is to analyse some aspects of different ethnic groups through machine learning methods and, more in detail, to find the important attributes of these groups and how those can influence each other. Therefore, our aim is to find a method capable to evidence some non-trivial relationships between different characteristics, which may help to better understand how some aspects of the group can or can not affect its well being or status.
Here, with the expression "ethnic group" we refer to a group of people linked and unified by a shared cultural background, such as religion or language. The definition of "ethnic group", provided in the description of the data set that we will be using to perform this analysis, is quoted here:

> "An ethnic group is defined as an identity group that defines itself or is defined by others along linguistic, religious or racial characteristics."[1]

The interest is to extend this analysis to all the regions in the World, so that we might achieve a partial insight of a generally valid behaviour, instead of limiting the research to a smaller area that might display more homogeneous characteristics and therefore more locally-valid results. Moreover, the analysis is also extended in time, at least in the first steps, since we will take into account different years in our sample, also in the attempt to achieve more generality.

In this work, we have chosen to deal with a supervised learning task, since the complexity of the topic would require an even deeper knowledge of the social mechanisms concerning those groups in order to perform an unsupervised task and to comment on its results. The approach that we followed is the one briefly summarised here:

1. Choose the most relevant features, in the attempt of keeping the ones that are carrying the most information. In order to do so we used two different approaches: K-means clustering and simulated annealing algorithm;
2. Use different supervised machine learning models to find the relation between those features and a target;
3. Study and improve those models;
4. Eventually, try to understand the results;

As a target we wanted to chose a variable that could exemplify the social and economical status of the group, and in particular we selected as target feature the variable "imr_mean_mean" (infant mortality rate), which will be described in the next section together with the data set. Even if this choice is quite specific, the aim of this work is to elaborate the method instead of focusing the attention on the specific variable, in principle the same procedure could be applied to describe other variables.

[1] https://growup.ethz.ch/about

2 Data set

2.1 Data collecting: the GROWup data set

The source of the data used in this work is the "GROWup" data set (Geographical Research on War, Unified Platform) [1]. Here a list of ethnic groups is provided with a set of features related to them, distinguished by year. The features concern various aspects of the groups considered, such as the political relevance, the status of the ongoing and past conflicts, the availability of natural resources and so on. Here we quote the brief description of the data set:

> "The GROWup platform offers a visualisation of settlement patterns of politically active ethnic groups around the world from 1946-2017. Additionally, it provides information about ethnic groups' access to executive government power, their involvement in civil war, federal administrative units, physical elevation, nightlight data, as well as population and GDP data by area."[2]

The fact that the provided features belong to different spheres (i.e. economy, politics, religion, language etc.) is a challenge, since generally they do not display strong correlation among themselves, unless they are part of a same ensemble describing the same aspect through different points of view.

2.1.1 Target variable

The target variable we have chosen is denominated "imr_mean_mean". This feature belongs to the PRIO-GRID data set, and the description of the acronym provided in the codebook is the following:

> "imr measures infant mortality rate, based on raster data from the SEDAC Global Poverty Mapping project. The original pixel value is the number of children per 10,000 live births that die before reaching their first birthday. This indicator is a snapshot for the year 2000 only.
> imr_mean gives the average infant mortality rate within the grid cell." [2]

This variable is then reporting the mortality rate of children, mediated and summed over the group considered for only the year 2000.

We will be now studying the data set in more detail.

[2] https://growup.ethz.ch/about

2.2 Data preprocessing and analysis

2.2.1 An overview of the shape of our data set:

The initial shape of the data set was (51715, 412), namely around fifty thousands ethnic groups distinguished per year considered and about four hundreds features.
To have a first insight of how some of the features are distributed we plotted the histograms counting how many times each value of the feature appears. For example, we can see how the number of groups considered for each year is constantly increasing and does not have a strong variation, Fig. 1.

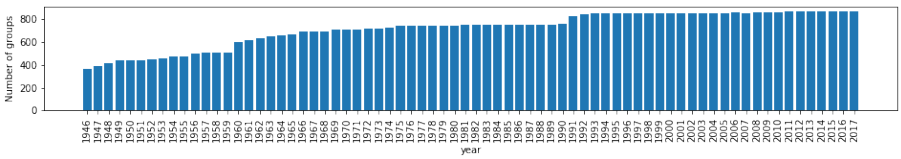

Figure 1: Years distribution

The number of times each Country appears is also of the order of few hundreds, with some exceptions (Russia: 3639, China: 2113, India: 1443 are the highest peaks), the histogram related is plotted in Fig. 2.

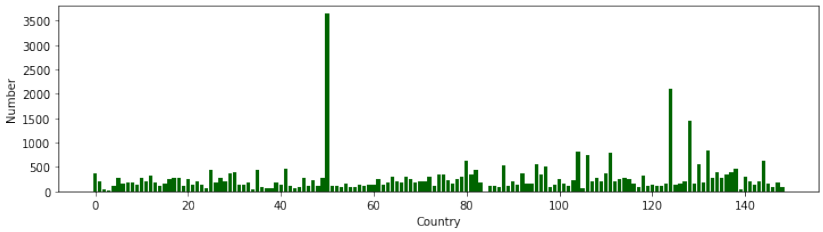

Figure 2: Countries distribution.

Another interesting feature is the "status_pwrrank", indicating the political status of the group, rated with an integer that goes from 2 (discriminated) to 7 (monopoly), Fig. 3.

Another relevant characteristic of the data set is that in general the values of the features do not vary considerably for the same group considered in different years, i.e. features are normally constant as the time passes. This low variance represent an obstacle for our models but, at the same time, it is already an interesting aspect of our sample that is part of the groups' behaviour analysis. Since we verified this flatness in time of our features, we decided to limit our analysis to one year only: we chose the year 2000 to be consistent with the choice of the target variable, "imr_mean_mean", that is related to this year.

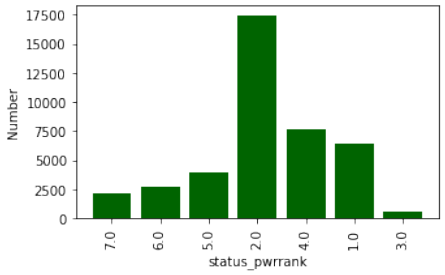

Figure 3: Power ranks distribution

2.2.2 Data preprocessing:

To prepare the set for our models it was necessary to perform a series of operations.
Firstly, we dropped all the string type variables, since for each of those the data set provides an encoded copy of their value: a column reporting that feature translated in a numeric code was already present in the set. This is for example the case for the variable "countryname" (i.e. name of the country) that has an encoded version, "countries_gwid" (each country correspond to a different integer).
After this first step, we dropped all the multiple features describing the same quantity, keeping just one value and avoiding therefore to have many features carrying the same substantial information. This was necessary in particular for the second half of the data set where quantities related to natural resources are displayed: for each of those the data set provides the sum, mean and standard deviation of the value. We decided to keep only the "sum" variables, since they can be thought as an enough representative description of the value. [3]

3 Probelm solution

3.1 Evaluation of the models

To evaluate the success of the models we used two main estimator: the score and the root mean squared error (RMSE).
The scoring is evaluated through built-in functions of the models, and it is

[3]These features mainly belong to the PRIO-GRID data set: http://grid.prio.org/#/

calculated through the expression:

$$R^2 = 1 - \frac{S_{res}}{S_{tot}} \tag{1}$$

where:

$$S_{tot} = \sum_{i=1}^{N} (\bar{y} - y_i)^2; \qquad S_{res} = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{2}$$

here, y_i indicates the data value, $\bar{y}$ the mean value of y_i, $\hat{y}_i$ the predicted value and N the total number of data y_i. The best the model is, the closer the score will be to 1.
The second evaluation method is the RMSE, root mean squared error. To calculate this quantity, first we compute the mean squared error (MSE)[4] and then we apply the square root to it:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2} \tag{3}$$

The better the model is, the closer the RMSE will be to the value of zero.

3.2 Data splitting:

To prepare the data set for the next elaboration, we split the lines in a training, validation and testing sections (respectively 80%, 10% and 10% of the whole data set) and we normalised them.

At this point we decided to study an algorithm to reduce the number of the features and to extract the more interesting ones in terms of correlation with the other features. The approaches that we followed are the K-means classification and a simulated annealing algorithm.

3.3 Features selection

3.3.1 K-means to classify the features

At this point we have more than a hundred features in our data set and we are aware that not all of them are carrying substantial information. Therefore, it is meaningless or even counterproductive to use all the features to predict our target. In order to solve this problem, we applied a classification method to our features: in this way we cluster them in different classes, and we can choose one feature in each separated group. In our case, the K-means method is chosen to

[4] In our case we used the sklearn function: sklearn.metrics.mean_squared_error, documentation link: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html

realise the classification.
Let's assume that we want to separate the features in k different classes, and we use $|1 - correlation(i, j)|$ as the distance between two features (the closer the correlation is to 1, the smaller will be the distance). Our goal is to minimise this equation:

$$\sum_{i=0}^{k} \sum_{p,q \in S_i} (1 - |correlation(p, q)|) \tag{4}$$

Where S_i represents the i^{th} set, and p,q are two features.
First, we select k random features as centres of the sets. Then, we calculate the correlation of every remaining feature with each centre: if the feature p has the biggest correlation with the i^{th} centre feature, then p will be assigned to the set i. When all the features have been assigned to a class, the next step is to choose new centre features for each set. For this, we pick the feature which fulfil:

$$max\{D_q | D_q = \sum_{p \in S_i} correlation(q, p),\ q \in S_i\} \tag{5}$$

D_q is the sum of the correlations between q^{th} feature and other features in the same set. We pick the one which has the biggest D_q as the new centre of this set. Repeating this choosing and classifying steps for several times, eventually the classification converges and gives the result.

3.3.2 Result of K-means

We set the number of the classes to 15, and the result of the classification for the first set is the one reported here (the other classes can be seen in the program):

['gdp05_total', 'pop90_total', 'pop90_corr', 'pop00_total', 'pop00_corr', 'pop10_total', 'pop10_corr', 'nightlight_total', 'gdp90_total', 'gdp90_corr', 'gdp95_total', 'gdp95_corr', 'gdp00_total', 'gdp00_corr', 'gdp05_corr', 'harvarea_sum', 'urban_gc_sum']

It can be seen from the result of this K-means classification that generally the features belonging to the same topic are clustered into the same group. For example, GDP in 2000 and GDP in 1995 are both belonging to the first group. Also, since the total GDP of a country displays strong correlation with the population of the country, they appear in the same group. Therefore, we are able to see how K-means classification give some satisfying results in this occasion. However, it still has some disadvantage: in fact, if we give different random initial centres, the resulting classes might appear different.

3.3.3 Using Simulated Annealing Algorithm to select features

Simulated annealing is a probabilistic technique for approximating the global optimum of a given function.[5] We want to use this algorithm to find the max-

[5] https://en.wikipedia.org/wiki/Simulated_annealing

imised solution of the goal function:

$$\sum_{i,j} \frac{|correlation(i,j)|}{|i-j|} \tag{6}$$

Here i, j represent two different features.
With the equation above we want to sweep the location of different features in the correlation matrix to move the high correlation terms near the diagonal. Expecting this procedure to lead to the formation of blocks along the correlation matrix diagonal, we then assume that each block corresponds to a group or class.
For the algorithm we follow those steps:
First, two random features are picked, and the algorithm checks if the goal function increases. If not, the exchange is rejected at the probability $1-e^{\frac{\delta}{T}}$, in which $\delta = score\ after\ change\ -\ score\ before\ change$ where T is a hyperparameter. Otherwise we accept the change, and perform the next sweep. The reason for accepting an exchange with small possibility even if it can reduce the score is to avoid being trapped in the local optimum. The idea of this method comes from simulation of paramagnetic-ferromagnetic phase transition in condensed matter physics, and that is why it is called "simulated annealing".

3.3.4 Result of Simulated Annealing Algorithm

The graphs below (Fig. 4, 5) show the total score variation during the evolution and the final result of the algorithm, i.e. the correlation matrix after switching the features. From the result we can see that the iteration can converge to a higher score compared to the original one. Normally, a greedy algorithm needs 128! searches for these 128 features in order to get the best score. However this algorithm can converge within 10^5 iterations although it cannot give the absolute maximum.

In the modified correlation matrix the most correlated features are grouped in blocks. To visualise the landscape of these blocks, the scores for individual features are computed and plotted (Fig. 6).

$$score_i = \sum_j \frac{|correlation(i,j)|}{|i-j|} \tag{7}$$

At this point we have the relevant features and we can use them to predict our target variable applying different methods. In this section we will briefly describe some of the methods we used and the main results.

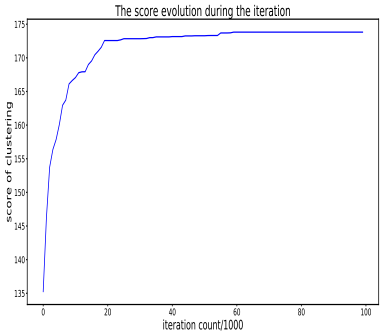

Figure 4: The score evolution in the iteration

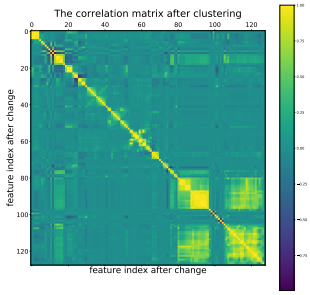

Figure 5: The correlation matrix after changing the feature position

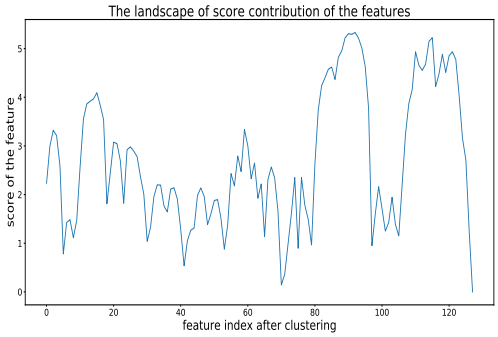

Figure 6: The score of features at the end of iteration

3.4 Models

3.4.1 Ridge Regression

Ridge Regression is a statistical supervised machine learning model that minimises the function (with respect to w):

$$||\hat{y}(w) - y||_2^2 + \alpha||w||_2^2 \tag{8}$$

where the norm used is the L^2 norm, α is an hyperparameter that has to be tuned in the training-validation steps, while w represents the set of parameters used to predict the value of y.
The role of the parameter α is therefore to give a weight to the coefficients and limit their sizes. In fact, if $\alpha = 0$ the Ridge model becomes equivalent to a least-squared solution, while if α has a larger size each coefficient will increase the loss function, therefore in order to reach a minimum the parameter will be

chosen to be as small as possible.
We used the Ridge model provided in the library sklearn. [6]
Due to the presence of the regularisation term, $\alpha||w||_2^2$, the Ridge regression can be used to evaluate the feature importance in the model, since the less relevant will shrink their value.

3.4.2 Results of Ridge model

3.4.3 Simple linear regression with Ridge

To apply this first simple method we use the features that are most correlated to the target feature in each classified set. These are:

> ['gdp95_corr', 'geo_typeid', 'status_senior', 'status_powerless', 'incidence_flag', 'min_coastal_km', 'egippop', 'diamsec_s_sum', 'upgraded10', 'upgraded_hist', 'downgraded_regaut_hist', 'downgraded_hist', 'upgraded_regaut_hist', 'mean_coastal_km']

We use these features as independent variables, to perform the simple linear regression of our target feature. For training and validation data, we also apply the cross check. The result of the model is shown in Fig. 7 and 8. We chose α in order to obtain the largest validation score.
The score obtained for the test data of the Ridge Linear Regression model is 0.486351.

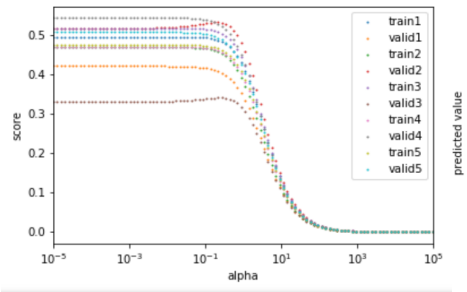

Figure 7: The score for the training and validation data changes as a function of varying alpha in the simple linear ridge model. The plot reports the results for the different sets used for the cross validation.

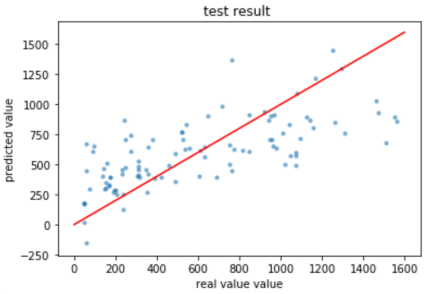

Figure 8: The result for the testing data of simple linear regression model. In the graph, the x-axis corresponds to the real value (y_i) and y-axis is the predicted value ($\hat{y}_i$). The red line is the ideal perfect fitting line, where y_i=$\hat{y}_i$.

[6] sklearn.linear_model.Ridge, documentation link: scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html

3.4.4 Exponential linear regression with Ridge

Another possibility is that our target might be exponential related with other features, such as GDP or land area. To check this hypothesis we use the exponential linear regression model, the results are plotted in Fig. 9 and 10.
The score for the exponential linear regression model is 0.508209.

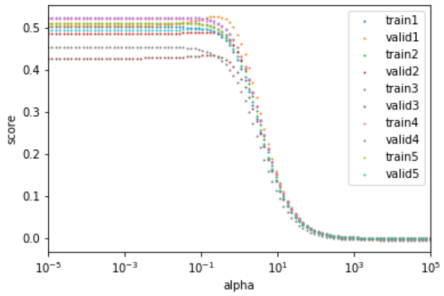

Figure 9: The score for the train and valid data changes as a function of α. We want the alpha which gives the largest score for the validation data. The plot reports the results for the different sets used for the cross validation.

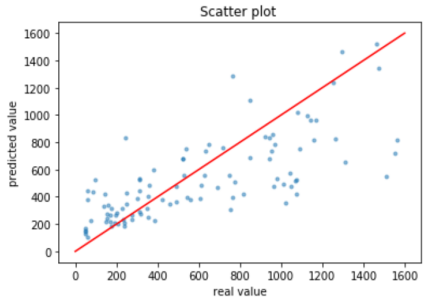

Figure 10: The result of the test data of exponential linear regression model.In the graph, the x-axis corresponds to the real value (y_i) and y-axis is the predicted value ($\hat{y}_i$). The red line is the ideal perfect fitting line, where y_i=$\hat{y}_i$.

Comparing these two linear models, we observe that the simple linear model has approximately the same residual for the whole imr range, while the exponential linear model has smaller residuals for small target value and a larger residuals for higher values of the target. Therefore, we can infer that the exponential model giver a better prediction for small imr while the simple linear model gives better results at the large value. The score of the test is not high, and this might be due to the fact that we have only 560 rows of data for this regression.

3.4.5 Decision Tree

This algorithm uses a tree-like decision method to learn from the data. In our case, it is a Regression Decision Tree model.
The decision tree model can be advantageous compared to linear regression in cases of discrete independent variables, strongly non-linear or non-smooth relations.
In the tree-based regression method, the dividing process is similar to the classification tree, but it can be used for continuous target variables. The iteration process is composed of recursive partitions of the training set depending on conditions related to independent variables, aiming to maximise the information that the division can give for the target variable. In this method the sum of

squared errors of the target in each leaf node is often used to decide the dividing quality. Specifically, the algorithm aims to minimise

$$\sum_{a} \sum_{i \in a} (y_i - \hat{y}_a)^2 \tag{9}$$

in which a represents the leaf nodes and $\hat{y}_a$ is the average of target variable y in leaf node a.
In each iteration the program uses a greedy search for the division of a leaf node that can minimise the sum of squared errors afterwards. The iteration ends when the stop criterion is satisfied. For example, the division of a node stops when the number of samples in it becomes lower than a user-defined value. Finally the method results in small blocks of variable spaces that are easier to fit into simple model, e.g. linear model or averaging the target values in each node.

3.4.6 Result of Regression Decision Tree

For the tree model we use the variables from different groups of features, which can represent different social aspects for ethnic groups. The independent variables are listed here:

['mean_coastal_km', 'peaceyears', 'ldiscrimpop', 'gdp95_corr',

'geo_typeid','landarea_sum', 'ed_rel1_size','meanelev']

Here we divide training and validating data into nine parts for cross validation. For each training process, the hyperparameters such as minimum instances in each leaf node, maximum depth of tree, or learning rate are tuned to improve the prediction result and to avoid over-training. Then the resulting fitted models are applied to the test set. The testing input variables are given to the nine models and the average of their predictions is compared with the true value of the testing target, which can give a direct insight into the prediction. The same cross validation technique will be applied to the next considered models.
Furthermore the tree methods can give the importance of each independent variables in the decision, which can also reflect how the infant mortality rate is related to these factors.

The result of regression tree method is plotted in Fig. 11 and 12. In the prediction for testing data the average score of the nine models is 0.696 ± 0.038.

3.4.7 Random Forest

The decision tree method can lead to over-training when the stopping criterion for minimum samples in each node is set to be too small, and in this case the growth of the tree is strongly affected by the statistical fluctuations of the

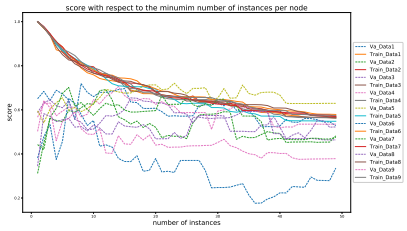

Figure 11: The score for training and validation data with the hyper-parameter minimum instances per node for regression tree model

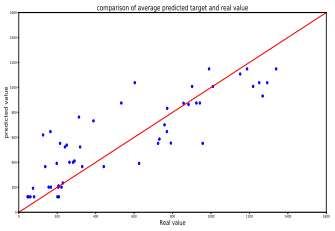

Figure 12: The average prediction of the regression tree models for test set vs. real values

training set, instead of following the expected general relations of independent variables and the target.

On the other hand, truncating the division of the tree can reduce the accuracy of the division. Moreover, a single decision tree can be sensitive to the initial condition. In another word, the division of one node can have large effect on its daughter nodes. The random forest method can be introduced to overcome these problems to some extent.
This algorithm is based on the previous one, but it is an improved version since it should lead to lower variance results. In this case, each tree is created from a sample coming from the training test and there is a randomness freedom in the choice of the splitting features. Many trees are thus created and averaged. Although each tree is normally less accurate than the case of one single decision tree, due to the final averaging the model usually leads to a result that has lower variance with respect to the Decision Tree model. Also, the averaging process can suppress the sensitivity to fluctuation of samples and initial condition of the trees.

3.4.8 Result of Random Forest Regression

The random forest method can slightly improve the result compared to pure regression tree, by averaging the result of a group of such trees. The result for this model is reported in Fig.13 and 14.
The score for random forest is 0.733 ± 0.014.

3.4.9 Gradient Boosted Decision Tree and Adaboost

This model is based on Decision Tree and Random Forest method, in general boosting methods can be introduced to further improve the original models.

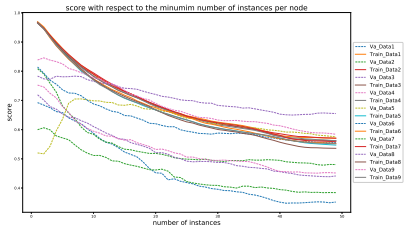

Figure 13: The score for training and validation data with the hyper-parameter minimum instances per node for random forest model

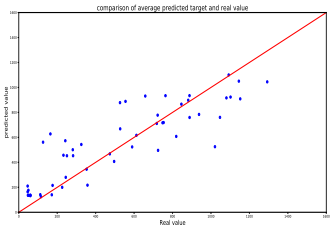

Figure 14: The average prediction of random forest models for test set vs. real values

The Gradient Boosting Decision Tree method modifies the Random Forest. In Random Forest, the final result gives the average of the predictions of the trees while in Gradient Boosting process, the averaging is replaced by a linear combination, and the coefficients are tuned by a gradient descent method. For one specific region of independent variables, the coefficients for the corresponding nodes in each tree are tuned separately. In another word, the coefficients for the leaf nodes in the same tree do not have to be the same. Each tree here is actually a function that can give predictions for input independent variables, and the gradient boosting is a linear regression with gradient descent method to fit the coefficients of these functions in order to minimise the error.

Another boosting method is Adaboost (Adaptive Boosting). It improves the result of random forest by assigning a weight distribution function depending on the independent variables for each weak learner.
At the beginning of iteration, the best learner is chosen with the minimum loss function and its weight distribution $D_t(i)$ is set to be identical.

$$D_t(i) = \frac{1}{m} \tag{10}$$

In the equation above m is the number of samples in training set, and i is the label of samples. t counts the iteration time, and in this case it is 1. Then a loss function is computed depending on the distribution function and errors of this tree. For estimation $f_t(x_i)$ and y_i and the error $l_t(i) = |f_t(x_i) - y_i|$, some options for the loss function are:
Linear:

$$L_t(i) = l_t(i)/Den_t \tag{11}$$

Square:

$$L_t(i) = (l_t(i)/Den_t)^2 \tag{12}$$

Exponential:

$$L_t(i) = 1 - exp(-l_t(i)/Den_t) \tag{13}$$

in which $Den_t = max_{i=1...m}(l_t(i))$. Then the program modifies the weight distribution for the next iteration.

$$D_{t+1}(i) = \frac{D_t(i)\beta_t^{1-L_t(i)}}{Z_t} \tag{14}$$

β_t is defined to be $\frac{\bar{L_t}}{1-\bar{L_t}}$ and $\bar{L_t}$ is the average of $L_t(i)$: $\bar{L_t} = \sum .L_t(i)D_t(i)$. Z_t is a normalisation constant. In the next step a second weak learner is chosen and assigned with this weight distribution. During the iteration the effect of each tree on the weight distribution is multiplied, which equals to adding the $ln\beta_t(1-L_t(i))$ on the exponential factor. In the end of the iteration the program gives an estimator that can give

$$f(x) = inf[y \in Y : \sum_{f_t x <= y} log(1/\beta_t) >= \frac{1}{2}\sum_t log(1/\beta_t)] \tag{15}$$

3.4.10 Result of Gradient Boosted Decision Tree

Based on the random forest method, the gradient boost can further improve the prediction. It can be seen in the scatter plot that the prediction of the models becomes closer to the the real data in the testing set, the results of this models are depicted in Fig. 15 adn 16.
The average score for the models on testing set is 0.913 ± 0.012.

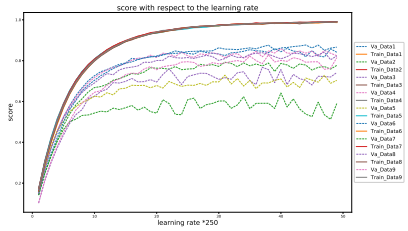

Figure 15: The score for training and validation data with the hyper-parameter learning rate for gradient boost decision tree

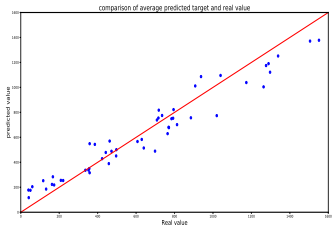

Figure 16: The average prediction of gradient boost decision tree models for test set vs. real values

3.4.11 Result of Adaptive Boost Decision Tree

The Adaptive Boost can also improve the prediction of the Random Forest model. Results of this model are reported in Fig. 17 and 18.
The average score of the Adaboosted decision tree model for the testing set is 0.855 ± 0.018.

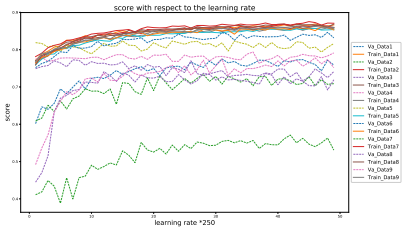

Figure 17: The score for training and validation data with the hyperparameter learning rate for Adaboost decision tree

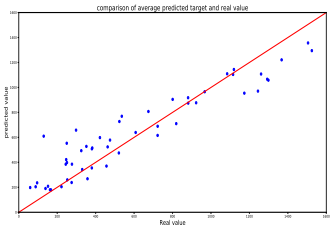

Figure 18: The average prediction of the Adaboost decision tree models for test set vs. real values

4 Results

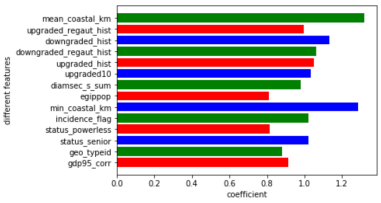

Figure 19: Coefficient of each features in the simple linaer model with Ridge regression.

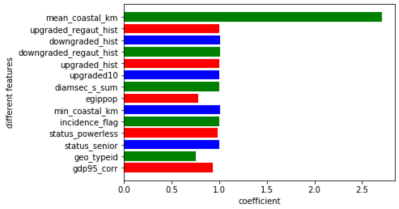

Figure 20: Coefficient of each features in the exponential linaer model with Ridge regression.

The results of regression models give some interesting insight on our sample. We can see from Fig. 19 and Fig. 20 that both simple linear regression and exponential linear regression with Ridge indicate the mean of the coastal length as the feature with the largest influence on the infant mortality rate. Moreover, the target variable has the strongest correlation with this quantity, 0.43. This result is not intuitive, and we will discuss it in the next paragraph. Other importantly related features, such as if the group has a high level of political relevance power or if it is downgraded, are more intuitive.

The tree models can give the importance of the independent variables when they are used to predict the target. The result of the four different tree-based models are shown below, in Fig. 21, 23, 22, 24.

From the ranking of the feature importance, we can see that the mean length of coast affects the target variable the most, and this is the same conclusion that

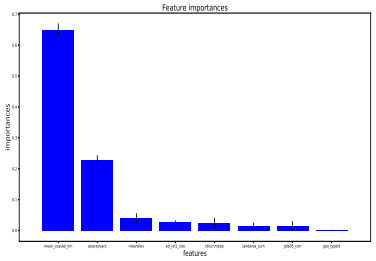

Figure 21: The importance of features plot for regression tree.

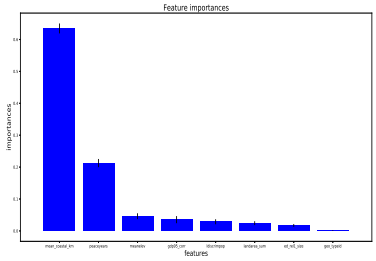

Figure 22: The importance of features plot for random forest.

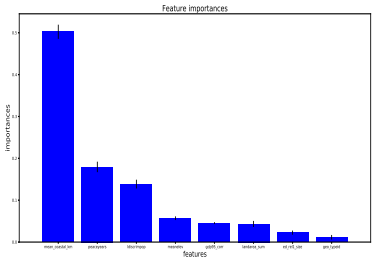

Figure 23: The importance of features plot for gradient boosted decision tree.

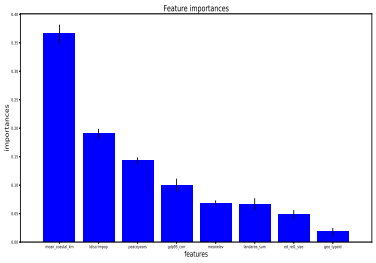

Figure 24: The importance of features plot for Adaboosted decision tree.

we derived from linear regression. The relation or potential causality between the two variables is not obvious at the first sight. It is likely that the coastal length is related to several factors such as climate, economy or transport conditions, which influence infant mortality rate when combined. Further research would be necessary to find out the actual cause of this relation and to confirm its validity.

To visualise the relation between two variables, we use the gradient boosted decision tree model after fit to make a quantitative analysis. The reason for using the models instead of original data is that the target is often affected by various variables at the same time, while with the models we can fix the other variables and analyse the influence of one single variable, in our case *mean_coastal_km*. The 9 models from cross validation are used to predict the target for given independent variables and then their average is used for the analysis. First we fix the other independent variables to be their average in the data set, and calculate the dependence of the target on *mean_coastal_km*. Then we also analyse the effect of *mean_coastal_km* when the influences of other in-

dependent variables are averaged out. For a specific value of $mean\_coastal\_km$, 1000 groups of other independent variables are randomly picked between their minimum and maximum respectively, then the predictions for this 1000 groups of values are averaged and the standard deviation is also computed and plotted in the graph. This process is equivalent to an integration over the other independent variables, which also gives a reliance of the target on $mean\_coastal\_km$. The non-linear dependence can be seen from graphs depicted in Fig. 25, 26.

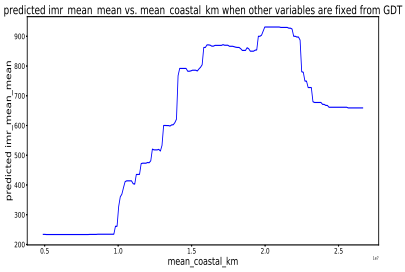

Figure 25: The dependence of infant mortality rate on mean coastal length when other features are fixed to be their average.

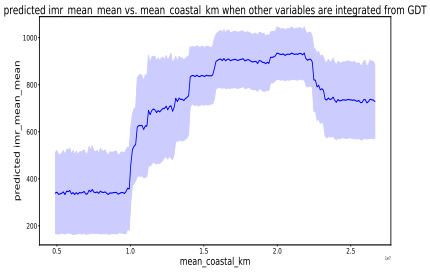

Figure 26: The dependence of infant mortality rate on mean coastal length when other features are integrated between their minimum and maximum

The result also shows that the variable *ldiscimpop* (i.e. fraction of discriminated population in the group) can strongly affect the target.

Previous research results such as [3] indicate that racial discrimination can cause the increase infant mortality rate in the discriminated group. Discrimination towards a group can often cause poverty and lack of education of the mothers, and this might affect the infant mortality rate. In addition, in further researches such as [4] it turns out that the racial effect still exists even if the income and education influences are excluded. More dramatically, it has also been shown in [5] that whether the mothers grow up in a discriminated environment is related to the infant health condition, when US-born and foreign born African mothers are compared, which indicates that the discrimination itself can have negative effect on the health of mothers and thus the health of their babies.
Other researches also point out that the gender discrimination towards women has impact on the infant mortality rate by affecting the healthcare, stress level and education of mothers[6]. The health of mothers is also more likely to be threatened by domestic violence in such groups with gender inequality[7].
The third factor affecting infant mortality rate is *peaceyears*, which counts the years from the last war. Wars can often cause the collapse of health care systems and lack of basic nutrition.[8] Moreover, women and children are more vulnerable under the violence and lack of water and food caused by war.[9][10].

Moreover, the relation between infant mortality rate and *peaceyears* also indicates that it takes time for infant mortality rate to decrease since the end of the wars, it might be because of the delay of economy and medical system reconstruction, or the effect of unstable political situation after the war.

5 Conclusions

In this work we analysed a set of data concerning ethnic groups and we focused our study on the prediction of the infant mortality rate, based on some other characteristics. One of the major challenge we had to face while dealing with this topic was to establish and measure in some way the importance of the characteristics and to deal with the lack of information due to the incompleteness of the data set.
For the feature selection we applied two different methods and both gave intuitively correct results, from which we deduced that grouping the features in clusters based on the correlation can be a good strategy to identify the most relevant features, and the centres of the clusters can have the function to summarise the information of its class.
In the second part of the work we applied different models, and we saw how the best performing model was the Gradient Boosted Decision Tree. From these models we inferred the importance of the features involved in the process of predicting the infant mortality rate.
The results we obtained might appear not completely intuitive at first sight, and in the previous section we tried to present a brief possible explanation for that, based on the literature related to our topic. However, it is certainly true that at this point it would be necessary to compare our modelled results with a socio-geo-political explanation or theory, that would require specific competence and knowledge of ethnic groups that goes beyond our abilities and our aim in this work. The possibility to compare our results with a different point of view, and in particular with the one of the pure social sciences, would be the best way of correcting and improving our model, that surely still have many limitations. If we had a comparison between our feature ranking and some theoretical or experience based behaviour study for those variable, we could modify our models accordingly and eventually this same procedure could be improved in order to be potentially applied in the prediction and study of other characteristics of the groups.

References

[1] Luc Girardin, Philipp Hunziker, Lars-Erik Cederman, Nils-Christian Bormann, and Manuel Vogt. 2015. GROWup - Geographical Research On War, Unified Platform. ETH Zurich. http://growup.ethz.ch/

[2] Tollefsen, Andreas Forø; Håvard Strand and Halvard Buhaug (2012) PRIO-GRID: A unified spatial data structure. Journal of Peace Research, 49(2): 363-374. doi: 10.1177/0022343311431287, website: http://grid.prio.org//codebook

[3] Alhusen, J. L., Bower, K. M., Epstein, E., and Sharps, P. (2016). Racial discrimination and adverse birth outcomes: an integrative review. Journal of midwifery and women's health, 61(6), 707-720.

[4] Collins Jr, J. W., and David, R. J. (1990). The differential effect of traditional risk factors on infant birthweight among blacks and whites in Chicago. American Journal of Public Health, 80(6), 679-681.

[5] Collins Jr, J. W., Wu, S. Y., and David, R. J. (2002). Differing intergenerational birth weights among the descendants of US-born and foreign-born Whites and African Americans in Illinois. American journal of epidemiology, 155(3), 210-216.

[6] Adhikari, R., and Sawangdee, Y. (2011). Influence of women's autonomy on infant mortality in Nepal. Reproductive health, 8(1), 7.

[7] Ahmed, S., Koenig, M. A., and Stephenson, R. (2006). Effects of domestic violence on perinatal and early-childhood mortality: evidence from north India. American journal of public health, 96(8), 1423-1428.

[8] Krug E (2002). World Report on Violence and Health. Geneva: Geneva WHO.

[9] Asling-Monemi, K.; Peña, R.; Ellsberg, M. C.; Persson, L. A. (2003). "Violence against women increases the risk of infant and child mortality: a case–referent study in Nicaragua". Bulletin of the World Health Organization. 81 (1): 10–16.

[10] Emenike E, Lawoko S, Dalal K (March 2008). "Intimate partner violence and reproductive health of women in Kenya". International Nursing Review. 55 (1): 97–102.