Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2021 doi:10.20944/preprints202110.0127.v1

Governance and socioeconomic factors contributing to antimicrobial resistance in European countries: a data panel and machine-learning analysis

Julian Riaño-Moreno[1,2], Jhoana P. Romero-Leiton [3*] and Kernel Prieto[4]

[1] *Faculty of Medicine, Cooperative University of Colombia, Villavicencio, Colombia*

2 El Bosque University, Bogotá D. C. Colombia, email: julian.camilo.riano@gmail.com

[3] *Faculty of Engineering, Cesmag University, Colombia*

[4] *Institute of Mathematics, National Autonomous University of Mexico, email: kernel@ciencias.unam.mx*

**Corresponding author. Email: jpatirom3@gmail.com*

Abstract

The aim of this work is to explain the behaviour of the multiresistance percentage of *Pseudomona aeruginosa* in some countries of Europe through a multivariate statistical analysis and machine learning validation, using data from the European Antimicrobial Resistance Surveillance System, the World Health Organization and the World Bank. First, we will use a descriptive analysis and a principal components analysis. Then, we use a k-means clustering to determine the countries and regions that are most affected by the antibiotic resistance. Second, we expand the database by adding some socioeconomic, governance and antibiotic-consumption variables. We then run a data panel regression analysis to determine some functions that relates the multiresistance percentage with

 doi:10.20944/preprints202110.0127.v1

those new variables. Finally, we use machine learning techniques to validate a pooling panel data case, using XGBoost and random forest algorithms. The results of the data panel analysis indicate that the most important variables for the multiresistance percentage are corruption control and the rule of law. Similar results are found with the machine learning validation analysis, where the human development index is an additional important variable for the multiresistance percentage.

Keywords: Descriptive analysis, principal components analysis, k-means clustering, data panel regression method, machine learning, XGBoost algorithms, random forest algorithms.

MSC2010: 92D30, 92D25, 49J15, 34D20, 62XX62, 62PXX.

---

1. Introduction

Antibiotics have completely revolutionized the world by prolonging human life. However, human behaviour has led to antibiotic abuse, or their inappropriate use in clinical settings, which has led to increasing antibiotic resistance.

Antibiotic resistance is an evergrowing concern globally in medicine and public health. Patients infected by antibiotic-resistant bacteria require extended hospital stays, costly and several treatments that result in an economic impacts on both the patients and the healthcare system [1]

Several pathogens have started to develop antibiotic resistance, particularly to first-line inexpensive broad-spectrum antibiotics, while the introduction of new drugs (e.g.

 doi:10.20944/preprints202110.0127.v1

fluoroquinolones) has been followed by the emergence and dissemination of resistant strains [2]. As resistance develops, disease related outbreaks have followed; for instance, acute respiratory tract infections cause 3.5 million deaths in children each year [3] and multidrug resistant tuberculosis (MDR-TB) caused 1.5 million deaths in 2018 (251,000 with HVI) [4].

The *Pseudomonas aeruginosas* [5] is a recognized micro-organism that is involved in antimicrobial resistance. It is a ubiquitous environmental bacterium that causes opportunistic human infections, such as urinary tract infections, respiratory system infections, dermatitis, soft tissue infections, bacteremia, bone and joint infections, gastrointestinal infections and a variety of systemic infections, particularly in patients with severe burns, and in cancer and AIDS patients who are immune-suppressed, [5], critical patients and in well-known bacteria involved in health-care associated infection [6].

The eradication of *P. aeruginosa* has become increasingly difficult due to its remarkable capacity to resist antibiotics, which includes biofilm-mediated resistance and the formation of multidrug-tolerant persister cells [7].

According to the European Antimicrobial Resistance Surveillance Network (EARS-Net), in 2017 30.8% of the *P. aeruginosa* isolates were resistant to at least one of the antimicrobial groups under regular surveillance. The highest European Union (EU) and European Economic Area (EEA) countries population-weighted mean resistance percentage in 2017 was reported for fluoroquinolones (20.3%), followed by piperacillin/tazobactam (18.3%), carbapenems (17.4%), ceftazidime (14.7%) and aminoglycosides (13.2%) [8].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2021 doi:10.20944/preprints202110.0127.v1

EARS-Net also claims that *P. aeruginosa* remains one of the major causes of healthcare-associated infection in Europe. Because of its ubiquitous nature and potential virulence, *P. aeruginosa* is a challenging pathogen to control in healthcare settings.

Antimicrobial resistance mechanisms are complex and a still unknown phenomenon, which includes well-described molecular phenomena (antibiotic-mediated selection, horizontal gene transfer, and others) [9] fostered by different social and behavioural determinants that are still unknown.

Several studies have found that the general and imprudent consumption of antimicrobials are the main causes for antimicrobial resistance. However, other factors have been suggested, such as socioeconomic factors and corruption [10]

Some researches claim that higher antimicrobial resistance rates can be found in low-income and middle-income countries, in which per-person consumption of antibiotics is much lower than in high-income countries [11,12,13]. In fact, the quality of governance and public spending on health, poverty, education, and community infrastructure are known to affect health outcomes [13].

We will include socioeconomic and governance variables such as total gross domestic product, gross domestic product for health, control of corruption (among others), and a variable that represents the consumption of antibiotics.

We will use a descriptive analysis to determine the most important characteristics of the data. Then, a Principal Components Analysis (PCA) is used to establish the most influential antibiotics in the data. Then, we run a k-means clustering analysis by country to determine which countries are most affected by antibiotic resistance. We then use the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2021 doi:10.20944/preprints202110.0127.v1

data panel regression method to determine the functions that relate to the multiresistance percentages with some socioeconomic, governance and antibiotic-consumption variables. Finally, we validate the previous results using Machine Learning (ML) techniques for the pooling panel data case. We use several kinds of methods as filters to keep the most important variables of a polynomial up to six degrees. We use a threshold value filter for the covariance between the target variables. We then use the XGBoost and random forest Algorithms.

2. Materials and methods

In this section we describe the methods used in this study. We have uploaded all the codes and source data used in this paper to the following https://github.com/jpatirom3/Governance_socioeconomic_resistance Github link for a detailed review.

2.1 Study area

Europe is located entirely in the Northern Hemisphere. and mostly in the Eastern Hemisphere. It is the second smallest continent of the world after Oceania. It has an area of 1,0530,751 km^2, which represents just 2% of the Earth's surface. Europe contains the borders of many countries. Europe contains around 50 countries, 27 of which are part of the European Union (EU) and some of the others are members of the European Economic Area (EEA). The EU/EEA is an economic and political union of 30 countries. It operates an internal (or single) market, which allows for the free movement of goods,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2021 doi:10.20944/preprints202110.0127.v1

capital, services and people between member states. A common geographical distribution splits the EU/EEA countries into four regions: Northern Europe, Southern Europe, Eastern Europe and Western Europe. Figure 1 shows the geographical distribution of the countries of the European regions that are used in this study.

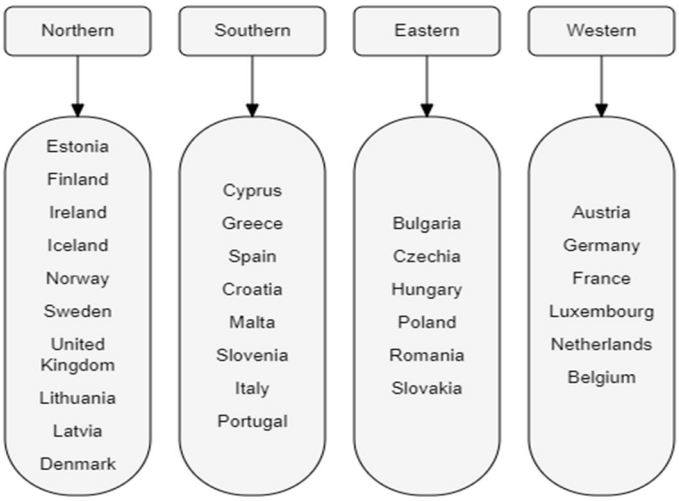

Figure 1: Geographical distribution of the countries of the EU/EEA regions used in this study.

2.2 Data collection

The data were collected from the European Antimicrobial Resistance Surveillance System (EARSS), the World Health Organization (WHO) and the World Bank. Their databases are available through the European Centre for Disease Prevention and Control (ECDC), the World Health Organization (WHO) and the Worldwide Governance Indicators (WGI) project.

For the purposes of this study we filtered the EARSS database for *P. aeruginosa* and the percentage of resistance to the following antibiotics: piperacillin tazobactam, carbapenems, fluoroquinolones, ceftazidime and aminoglycosides, by country and year, and from 2005 to 2018.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2021 doi:10.20944/preprints202110.0127.v1

We used the EARSS database corresponding to the resistance percentages and multiresistance percentages to *P. aeruginosa* to the five antibiotics mentioned previously. The original database contained information collected from 2000 to 2018 of eight (8) bacteria, namely: *Acinetobacter spp*; *Enterococcus faecalis; Enterococcus faecium*; *Escherichia coli*; *Klebsiella pneumoniae*; *Pseudomonas aeruginosa*; *Staphylococcus aureus*; *Streptococcus pneumoniae*) corresponding to the 30 EU/EEA countries. This first database had 67,542 observations and nine variables (i.e., HealthTopic, Population, Indicator, Unit, Time, RegionCode, RegionName, NumValue and TxtValue).

After an exhaustive data curation and reorganization, 3,478 observations and 72 variables were obtained. These 72 variables corresponded to the type of indicator for antibiotic (e.g., column name or variable *ceftazidime-r-percentage*, which corresponded to the resistance percentage to ceftazidime). These variables were used for each of the reported bacteria.

We then eliminated the variables that were not informative, such as those that referred to the number of isolates in each of case. We removed non-informative data (NAs) and we maintained only the variables corresponding to the resistance percentage for each of the antibiotics reported for each bacteria. Subsequently, we only restrict the data for the bacteria: *Pseudomona Aeruginosa*. Additionally, we implemented a *web scrapping* strategy, to incorporate the socioeconomic and antibiotic-consumption variables and also to obtain the regions of the European continent and matched it by country. Thus, 10 new variables (i.e., region, GDP_total, GOV_effect, GDP_health,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2021 doi:10.20944/preprints202110.0127.v1

CTRL_corrup, Rule_law, Per_cap_US, Out_pocket_exp, HDI and DDD_sys_comun) were then added to data. The final variables considered in this study are specified in Table 1.

Variable name	Definition
Year	Years from 2005 to 2018
Country	Country name
Region	Region name (eastern, northern southern, western)
Antibiotic	Resistance and multiresistance (R_multi) percentages to five antibiotics Aminoglycosides, Fluoroquinolones Carbapenems, Piperacilina_taz and Ceftazidime
GDP_total	Gross domestic product
GOV_effect	Government effectiveness
GDP_health	Gross domestic product for health
CTRL_corrup	Control of corruption estimated
Rule_law	Rule of law
Per_cap_US &	Current health expenditure per capita in US
Out_pocket_exp	Out-of-pocket expense
HDI	Human development index
DDD_sys_commun	Daily doses of antibiotic per 1,000 inhabitants per day

Table 1: Study variables: description.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2021 doi:10.20944/preprints202110.0127.v1

2.3 Multivariate Analysis

2.3.1 Correlation Analysis

We did a correlation analysis between the resistance percentages of *P. aeruginosa* to the antimicrobial groups under regular surveillance, including the resistance percentage to at least three antibiotics simultaneously (multiresistance percentage). Using the **R** software, we first determined the confidence intervals (CI) for the resistance percentages to each antibiotic and we then computed the correlation matrix.

2.3.2 Principal Component Analysis

A PCA was only done on the standardized resistance percentage to the five antibiotics (we did not consider the multiresistance percentage). The new variables were ordered by the percentage of original variance that they described. A reasonable number of components should be retained to avoid an under-performing forecast. The Cattell's Scree Test [14] was chosen as a criterion of relevance.

PCA was also done to discriminate the percentage of resistance observations by region. We used a biplot-type graph (see Figure 2).

The selected variables that appeared in the dimensions for each antibiotic were identified and only the first component (eigenvalue of 4.29 and it explains 85.7% of data variance) was considered.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2021 doi:10.20944/preprints202110.0127.v1

2.3.3 K-means clustering

Through a k-means clustering, we analysed the correlation between the EU/EEA countries with respect to the resistance percentage to *P. aeuruginosa*. K-means is a partition technique that allows to the data to be grouped into clusters, so that the objects within a cluster are similar but objects in the other groups are different. A centroid-based partitioning method was applied, using the centroid of a cluster to represent the partition. During the application of the technique, one k centroid was defined for each cluster, and the objective function to be minimized was the square sum of the error [15], as follows

$$\varepsilon = \sum_{i=1}^{k} \sum_{p \epsilon C_i} dist(p, C_i), \quad (1)$$

where ε is the square sum of the error for all objects in the dataset; p is the point in space representing an object; and C_i is the centroid of each cluster. The k-means clustering technique was applied to group the percentage of resistance data to identify how the countries are grouped in relation with the resistance percentages. The selection of the appropriate number of clusters is one of the most influential factors on the results of k-means clustering. The Mojena criterion was used to determine the optimal number of clusters [16]. Once the best number of clusters was determined, the k-means technique was applied to group the data.

2.4 Panel data analysis

Panel (or longitudinal) data provides us with observations on cross-section units (e.g. individuals, firms, industries, countries, regions), $i = 1,2,\ldots,N$ over repeated time

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2021 doi:10.20944/preprints202110.0127.v1

periods, $t = 1,2,\ldots,T$ The cross-section units will be referred to as units or groups. In this study, each of the EU/EEA countries was considered as a unit. Meanwhile, the difficulty to obtain the information through the time of the individuals was found. Due to the amount of missing data and to maintain the homogeneity of the data, a polynomial interpolation of the missing data was done. However, it was not possible to make up for the lack of data from the EARSS database (only with the resistance percentage information), which forced the construction of an unbalanced panel data (i.e., when there are missing elements that result in an incomplete data series for an individual or individuals are absent in some years for some variable) with $i = 30$ and $t \in [8,14]$, being Slovakia the country with the fewest time periods ($t = 8$) followed by Belgica ($t = 10$) and most of them (20) with time periods of 14. According to Hsiao [17], although many statistical proposals are built from the consideration of balanced panels, most of the empirical studies and the data that can be used only enables unbalanced panels-such as the one presented in this study.

We used the Fixed Effects method (FE). Thus, it is necessary to apply a transformation of the data to eliminate the unobserved heterogeneity, which allows the fixed effects estimator to take the form of an Ordinary Least Squares (OLS) estimator. Thus, we proposed FE models for multiresistance percentage with the following structure:

$$R\_multi = d_1\text{Gdp\_total} + d_2\text{Gov\_effect} + d_3\text{Gdp\_health} + d_4\text{Ctrl\_corrup} +$$

$$d_5\text{Rule\_law} + d_6\text{Rural\_pop} + d_7\text{DD\_sys\_commun} + d_8\text{Per\_cap\_us} +$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2021 doi:10.20944/preprints202110.0127.v1

$$d_9\text{Out\_pocket\_exp} + + d_{10}\text{HD} \quad (2)$$

where d_i and c_i, $i = 1,\ldots,10$ are the model coefficients.

2.5 Pooling panel data analysis using machine learning

We performed the Hausman test to determine if the Fixed (FE) or Random Effects (RE) is most suitable for our panel data. The Hausman test is a test of endogeneity. The null hypothesis is that the preferred model is the random effects versus the alternative the fixed effects.

In the ML framework, we used models such as those used by [18,19]. We used the *Scikit-learn* package in *Python*. We also used the Ordinary-Least-Square method. We noted that it is possible to also use the *Statsmodels* package of *Python*. We modeled the relationship between the input variables $x_1, x_2,\ldots,x_n$, called features in the ML framework, and the output variable, y, called the target variable, as a non-linear relationship with the following polynomial of degree m, called the complexity on ML framework, on the variables $x_1, x_2,\ldots,x_n$

$$y = d_0 + d_1x_1 + \cdots + d_nx_n + d_{n+1}x_1x_2 + \cdots + d_{n+m_1}x_1^m + d_{n+m_1+1}x_2^m \ldots d_{n+M}x_n^m. \quad (3)$$

Here, the variables $x_1, x_2,\ldots,x_n$ were the $n = 9$ variables given in (2). Although equation (3) is non-linear on the features x_i, the estimation of the parameters d_i is still a linear regression because the equation (3) is linear on the parameters d_i. First, we used

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2021 doi:10.20944/preprints202110.0127.v1

a polynomial of degree $m = 3,4,5$ of the polynomial (3) and built the covariance matrix of these polynomial variables $x_1, \ldots, x_i^{l1} x_j^{l2} x_k^{l3}, \ldots, x_n^m$, such that $l1 + l2 + l3 = m$, with respect the target variable, y, variables and have selected the polynomial variables, x_i, which covariance value with respect to the target variable, $Cov(x_i, y)$, have a greater value than a threshold value, $\delta = \{0.3, 0.5, 0.6\}$, R_multi. Next, we split our data: 80% is used as training data and 20% is used as testing data. We presented the results of the R^2 and $R^2_{adjusted}$ on the test data for polynomial degrees $m = 3,4,5$ and the threshold values $\delta = \{0.3, 0.5, 0.6\}$ selected.

We also used another type of variables filters. We first used the Low Variance Filter, which consists in eliminating the variables which variance are less than a threshold value, η. Apart of the Linear Regression (LR) Algorithm, we have used the k-Nearest Neighbors (kNN) and the Decision Tree (DT) algorithms. The hyperparameters for the KNN are $k = 5$ neighbors and $n = 5$ trees.

Furthermore, we applied the k Best Variable Selection (kBVS) and the Recursive Feature Elimination (RFE) methods as other filter techniques of the polynomial features. The k Best Variable Selection method selects the k best variables based on the Fisher Test. The Recursive Feature Elimination method recursively removes the weakest features or the least sensitive features over the target variable until the desired manually selected variables are reached. We noted that the RFE method cannot be applied for the kNN Algorithm.

Finally, we used the XGBoost and the Random Forest Algorithms as two more options for the Pooling Method with the combination of the Shapley Additive exPlanations

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2021 doi:10.20944/preprints202110.0127.v1

(*SHAP*) Package [20]. The latter package was used to make some plots to better explain the models. We used the following hyperparameters for the XGBoost: $\eta = 0.3$, $\text{max\_dept} = 3$, $\text{subsample} = 0.5$, $\text{iterations} = 10000$.

3. Results

3.1 Multivariate Analysis

Table 2 shows the confidence intervals with 95% as confidence levels, the mean vector ($\bar{X}$), the standard deviation vector (σ) and the variation coefficients vector (***CV***) of the resistance percentage for each antibiotic. Table 3 shows the correlation matrix.

	Amino.	Carba.	Cefta.	Fluoro.	Piper/taz	R_multi
$\bar{X}$	16.16	18.01	14.32	20.56	17.22	14.35
σ	14 .2	13.43	11.85	13.19	12.19	12.71
CV	87.89	74.59	82.76	64.18	70.77	88.58
CI(95%)	[15, 17.33]	[16.91, 19.12]	[13.34, 15.29]	[19.47, 21.64]	[16.22, 18.22]	[13.31, 15.40]

Table 2: Mean vector, standard deviation vector, variation coefficients vector and confidence interval with 95% of confidence level for each antibiotic. The error of estimation was $e = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ The values are given on percentages.

	Amino.	Carba.	Cefta.	Fluoro.	Piper/taz	R_multi
Amino.	1	0.79	0.79	0.89	0.85	0.93
Carba.	0.79	1	0.84	0.81	0.77	0.88
Cefta.	0.79	0.84	1	0.8	0.84	0.89

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2021 doi:10.20944/preprints202110.0127.v1

Fluoro.	0.89	0.81	0.8	1	0.83	0.92
Piper/taz	0.85	0.77	0.84	0.83	1	0.9
R_multi	0.93	0.88	0.89	0.92	0.9	1

Table 3: Correlation matrix

From Tables 2 and 3 we observed that in mean, the highest resistance percentage is for fluoroquinolones (20.56%) followed by carbapenems (18.01%) and piperaciclin/tazobactam (17.22%). We could see also that there is a high correlation between the resistance percentages of all pairs of antibiotics (the correlation coefficients are greater than 0.77), being the highest correlation between the multiresistance percentage and all other antibiotics, which is an obvious result, and the lowest correlation between carpapenems and piperacillin/tazobactam (0.77). This indicates that when a bacteria is resistant to an antibiotic, then it is also resistant to others. The results obtained in the correlation analysis were of vital importance in the subsequent analyses.

Table 4 and Figure 2 summarize the results of PCA, such as eigenvalues for each dimension, the percentage of variance explained (PVE), the cumulative percentage of variance explained (CPVE) and the contribution by country of each dimension.

	Eigenvalue	PVE	CPVE
Dimension 1	4.29	85.7	85.7
Dimension 2	0.274	5.48	91.2
Dimension 3	0.221	4.41	95.6

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2021 doi:10.20944/preprints202110.0127.v1

Dimension 4	0.12	2.41	98
Dimension 5	0.0988	1.98	100

Table 4: Summary of the PCA.

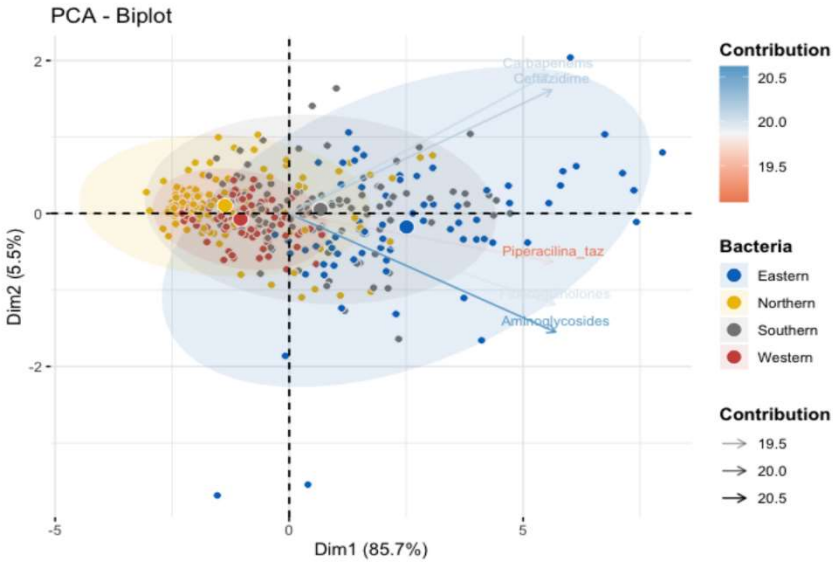

Figure 2: Contribution biplot of antibiotics by region of Europe.

We observed that the Eastern European countries contribute the most in Dimension 1. There are also some countries in the Southern and Northern regions of Europe. These results suggested that the least contribution to the problem of bacterial resistance is made by Western countries. In particular, the Eastern countries contribute to the antibiotic resistance of groups such as aminoglycosides and fluoroquinolones, which are last-line antibiotics for the treatment of infections caused by *P. aeuroginosa*. To corroborate these results, we use a k-means cluster analysis.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2021 doi:10.20944/preprints202110.0127.v1

We got three clusters, which are shown on Figure 3, which corroborated the results obtained through the PCA.

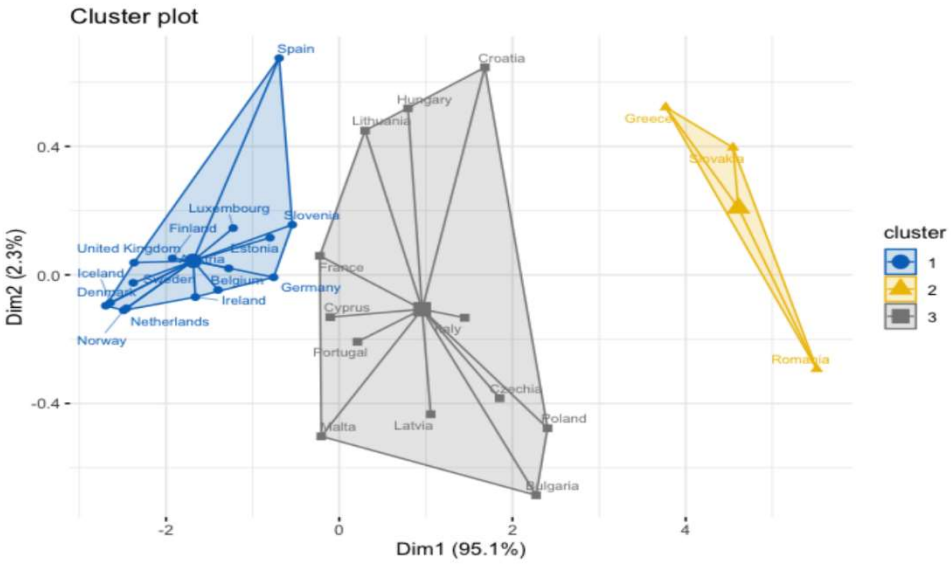

Figure 3: Clusters using k-means method.

3.2 Panel data analysis

Tables 5 and 6 show the results of panel date analysis after removing those coefficients that are not significant (p-value greater than 0.05).

These results were extracted using **EViews** software.

Variable	Coefficient (d_i)	Std. Error	T-statistic	P-value
Gdp health	-0,021294735	0,006690584	-3,182791556	0,001588912
Ctrl corrup	-0,073232514	0,026884398	-2,723978198	0,006772701
Rule law	0,118845636	0,03168187	3,751219069	0,000205806
DDD sys commun	0,004008661	0,001592144	2,517775888	0,012254289

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2021 doi:10.20944/preprints202110.0127.v1

Per cap US	$2{,}86 \times 10^{-5}$	$9{,}44\times 10^{-6}$	3,025597026	0,002664484
Out pocket exp	0,005861159	0,001933574	3,031256929	0,002616302
HDI	-1,426357188	0,581775562	-2,451731012	0,014703742
Intercept (C)	1,217893722	0,499095622	2,440201174	0,01517321

Table 5: Coefficients of the initial panel data for the FE multiresistance model.

From the results given in Table 5, we obtain the following structure for the models:

$$\mathrm{R\_multi} = \bar{d}_1 \mathrm{Gdp\_health} + \bar{d}_2 \mathrm{Ctrl\_corrup} + \bar{d}_3 \mathrm{Rule\_law} + \bar{d}_4 \mathrm{DDD\_sys\_commun} +$$

$$\bar{d}_5 \mathrm{Per}_{\mathrm{cap_{us}}} + \bar{d}_6 \mathrm{Out}_{\mathrm{pocket_{exp}}} + \bar{d}_7 \mathrm{HDI} + \mathrm{C}.$$

In Tables 6, the R^2 and $R^2_{adjusted}$ for the model is shown. Table 7 shows the results of the validation tests of the models and Tables 9 and 10 show the cross time series effects for years and countries.

R^2	0,821638359
$R^2_{adjusted}$	0,796738871

Table 6: R^2 and $R^2_{adjusted}$ for the final FE multiresistance model.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2021 doi:10.20944/preprints202110.0127.v1

Test	Statistic	P-value
Durbin-Watson	1,998868923	8.58×10^{-10}
Breusch-Pagan	273,4608126 (DF=91)	3.83×10^{-20}
Jarque-Bera	955,3028	0.0000
Bartlett	3.489150 (Df=3)	0.3221719

Table 7: FE multiresistance model validation tests.

Year	Effect (Multiresistance)
2005	0,023793591
2006	-0,019273908
2007	-0,036315213
2008	-0,015436554
2009	-0,006274706
2010	-0,000474912
2011	-0,002099464
2012	-0,000364172
2013	-0,014729078
2014	0,003484427
2015	0,014952204
2016	0,017928067
2017	0,017056765

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2021 doi:10.20944/preprints202110.0127.v1

2018	0,1765423365

Table 8: FE panel model time series effects.

Country	Effect (Multiresistance)
Cyprus	-0,20126558
Luxembourg	-0,152954443
Malta	-0,117017084
Norway	-0,107294776
Latvia	-0,103793424
Iceland	-0,083119698
Estonia	-0,062798702
Austria	-0,052746151
Lithuania	-0,04973579
Ireland	-0,047151656
Denmark	-0,04695085
Finland	-0,046832957
Sweden	-0,045701303
United Kingdom	-0,039522724
Portugal	-0,026484986
Bulgaria	-0,019421904
Spain	-0,018007836
Belgium	-0,013962841

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2021 doi:10.20944/preprints202110.0127.v1

Hungary	-0,009903188
Netherlands	0,01802192
France	0,048518441
Germany	0,089771891
Poland	0,096052476
Italy	0,098525797
Slovenia	0,108944932
Greece	0,1269621
Czechia	0,140159238
Croatia	0,164864454
Romania	0,222494055
Slovakia	0,228095313

Table 9: FE panel model country effects.

3.3 Pooling panel data analysis using machine learning

From the Hausman test, the p −value was less than 0.05, thus the FE model was statistically better choice than the RE model. The Hausman test results are given in Table 10. Moreover, the F-test between to use pooled OLS or fixed effects said that there is no significant heterogeneity across time and countries (i.e., the pooled OLS is the better choice). The result of this F-test is presented in Table 11, which showed that we must accept the null hypothesis. We noted that similar results are obtained using a Hausman test between the pooled OLS and the fixed effects model. Since the F-test's result showed

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2021 doi:10.20944/preprints202110.0127.v1

that we must opt for a pooled OLS model, we opted to use ML without considering heterogeneity across time and countries.

Figure 4 shows the results of ML analysis. We noted that we did not apply the polynomial features technique given in (3) at using the XGBoost and the Random Forest Algorithms.

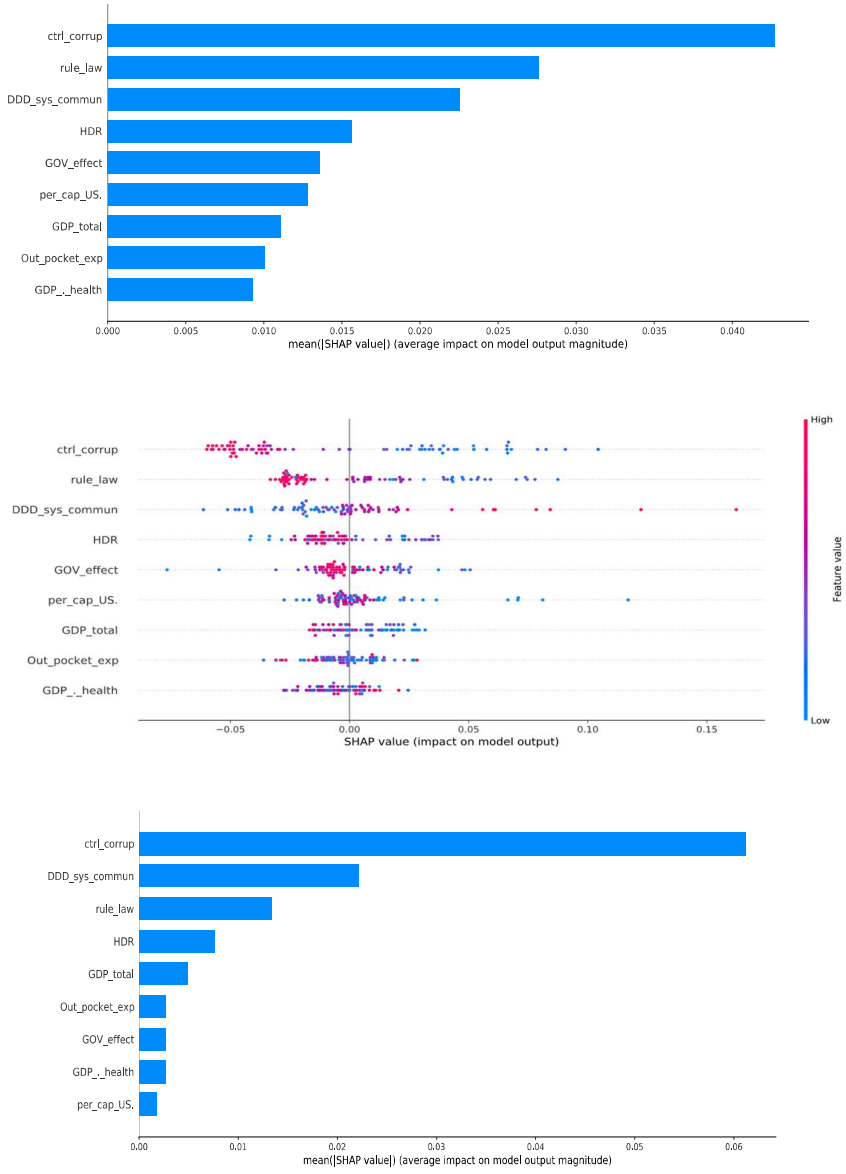

Figure 4: Machine Learning results. Top row: feature importance of the training data on the target variable R_multi using the SHAP and XGBoost Algorithm, respectively. Middle row: feature importance of the training data on the target variable R_multi using the SHAP and the XGBoost Algorithm, respectively.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2021 doi:10.20944/preprints202110.0127.v1

Bottom row: feature importance of the training data on the target variable R_multi using the SHAP and the Random Forest Algorithm.

Tables 12 and 13 show the performance comparison of polynomial features and the threshold value δ with respect to the covariance value $Cov(x_i, y)$ on the target variable R_multi. From Tables 12 and 13 we can observe that the highest value of R^2 on the test data is obtained with the combination $m = 5$, $\delta = 0.6$ for the output variable: R_multi. We chose the complexity $m = 3,4,5$ based on the number of features, $n = 5$, and the data size equal to 401. As a rule of thumb, the complexity of our model, in this case the polynomial degree of (3), must be selected as follows: the number of data points should be no less than some multiple between 3 and 10 of the number of parameters to estimate, i.e., the number of terms of our polynomial (4) given by $\binom{n+m-1}{m-1}$, which for $m = 2,3,5$ are $\binom{10}{8} = 45$, $\binom{11}{8} = 165$, respectively. Therefore, after applying the threshold number δ, we should keep in mind empirically to select at most 134 terms.

Next, we present the results obtained with the Low Variance Filter. We eliminate the variables whose variance is less than a threshold value, η. Apart from the Linear Regression (LR) algorithm, we have used the k-Nearest Neighbors (kNN) and the Decision Tree (DT) algorithms. The hyperparameters for the KNN are $k = 5$ neighbors and $n = 5$ trees.

Furthermore, we applied the k Best Variable Selection (kBVS) and the Recursive Feature Elimination (RFE) methods as filter techniques of the polynomial features. The k Best Variable Selection method selects the k best variables based on the Fisher Test. The

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2021 doi:10.20944/preprints202110.0127.v1

Recursive Feature Elimination method removes recursively the weakest features or the least sensitive features over the target variable until the desired manually selected variables is reached. Table 14 shows the performance comparison of polynomial features, type of Algorithm and the threshold value η on the target variable R_multi. From Table 14, we can observe that the highest R^2 value on the validation set was obtained with the combination of $m = 3$ degree of the polynomial, Algorithm set to LR and $\eta = 0.3$ (selected features equal to 85) on the R_multi variable.

Table 14 show Performance of the k-Best Variable Selection and the Recursive Feature Elimination methods on the target variable R_multi. We can observe from Table 14 that the highest R^2 value on the validation set was obtained with the Recursive Feature Elimination although the time execution is about 1.5hrs for a five degree polynomial.

ξ^2	DF	p-value
96.50	9	2.2×10^{-16}

Table 10: Hausman test between Fixed and Random Effects.

F	DF1	DF2	p-value
1.099	29	362	0.3342

Table 11: F-test between Pooled OLS and Fixed Effects.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2021 doi:10.20944/preprints202110.0127.v1

m	δ	Selected vars	R^2 (test)	RMSE (test)	R^2 (train)	$R^2_{adjusted}$(train)
3	0.5	76	0.602011	0.076700	0.843	0.793
3	0.6	34	-0.41801192	0.1447774	0.722	0.688
4	0.5	77	0.32147335	0.10014847	0.850	0.804
4	0.4	110	0.59188053	0.07767021	0.894	0.838
4	0.6	34	-0.41801192	0.14477741	0.722	0.688
3	0.5	76	0.602011	0.076700	0.843	0.793
3	0.6	34	-0.41801192	0.1447774	0.722	0.688
5	0.6	45	0.655947668	0.07131372	0.789	0.754
5	0.55	122	0.49342470	0.086533223	0.894	0.828
2	0.3	13	0.6161932032	0.0753211962	0.725	0.713
2	0.2	24	0.6400755133	0.0729401371	0.753	0.733
2	0.1	41	0.64501943033	0.072437452652	0.780	0.748

Table 12: Performance comparison of polynomial features and the threshold value δ with respect to the covariance value $Cov(x_i, y)$ on the target variable $y = \text{R\_multi}$.

m	Algorithm	η	Selected vars	R^2 (test)	RMSE (test)
3	LR	0.01	135	0.75079266782	0.06069337475
3	kNN	0.01	135	0.71696136129	0.06468204494
3	DT	0.01	135	0.59647645671	0.07723165076
3	LR	0.03	85	0.76871932762	0.05846965734
3	kNN	0.03	85	0.72046320501	0.06428066623
3	DT	0.03	85	0.61279712114	0.07565369902

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2021 doi:10.20944/preprints202110.0127.v1

3	LR	0.05	42	0.60366993964	0.07654016341
3	kNN	0.05	42	0.73875824488	0.06214156333
3	DT	0.05	42	0.38112062778	0.09564535472
2	LR	0.01	49	0.63975956573	0.07297214411
2	kNN	0.01	49	0.71305481234	0.06512689170
2	DT	0.01	49	0.43579278492	0.09132300555
2	LR	0.03	36	0.62109266821	0.07483889782
2	kNN	0.03	36	0.72170833139	0.06413734540
2	DT	0.03	36	0.40578655115	0.09371996482
5	LR	0.05	77	0.57646543152	0.07912346719
5	kNN	0.05	77	0.73228675897	0.06290654082
5	DT	0.05	77	0.40679530773	0.09364037996

Table 13: Performance comparison of polynomial features, type of algorithm and the threshold value η on the target variable $\mathrm{y} = \mathrm{R\_multi}$.

m	Algorithm	Filter Method	Selected vars	R^2 (test)	RMSE (test)
3	LR	kBVS	60	0.65393025273	0.07152250163
3	kNN	kBVS	60	0.74928967132	0.06087612379
3	DT	kBVS	60	0.54737632879	0.08179551852
3	LR	kBVS	70	0.67226953928	0.06960160952
3	kNN	kBVS	70	0.75930523349	0.05964776868

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2021 doi:10.20944/preprints202110.0127.v1

3	DT	kBVS	70	0.66984972895	0.06985809002
3	LR	kBVS	80	0.69905119445	0.06669713674
3	kNN	kBVS	80	0.76583084091	0.05883364106
3	DT	kBVS	80	0.65731394839	0.07117198724
5	LR	kBVS	60	0.63975956573	0.06963926882
5	kNN	kBVS	60	0.71305481234	0.06574826883
5	DT	kBVS	60	0.67191479342	0.07777055852
3	LR	RFE	55	0.71741552514	0.07757176375
3	DT	RFE	7	0.66243686127	0.08478265748
5	LR	RFE	36	0.71595373164	0.07777214212
5	DT	RFE	64	0.71650937527	0.07769603700

Table 14: Performance of the k-Best Variable Selection and the Recursive Feature Elimination methods on the target variable $y = R\_multi$.

4. **Discussion**

This study presented the complexity of the antibiotic resistance and multiresistance phenomena of *P. Aeruginosa* for the EU/EEA countries. We proposed an approach that uses a multivariate statistical analysis, a data panel strategy and machine learning techniques. Addressing this problem cannot be limited to good control practices centred on the institutions of health alone, nor private relationships between health personnel within institutions or antibiotic self-formulation. We proposed a much broader scenario where the willingness of governments to distribute their resources in basic goods such as

 doi:10.20944/preprints202110.0127.v1

health, governance practices and human development become basic needs to be solved to improve control practices for multiresistant antibiotics.

Meanwhile, the control of corruption is essential to guarantee the adequate distribution of the resources destined to cover the basic health needs of the population, so that, in a country where the control of corruption is poor, there is less probability to guarantee adequate spending on health to respond to problems such as antibiotic resistance and there is also less confidence on the part of the population in their institutions. The latter is related to the variable follow-up of the laws (Rule_law); if the population does not trust on their institutions, then the population will not follow the established norms, which can be reflected in the disobedience to follow antibiotic regimens or other medications, which can increase the problem of antibiotic resistance. Additionally, the models proposed in this study showed the complexity of the antibiotic resistance problem by including variables of consumption, health spending and governance. This indicates that the antibiotic resistance in *P. Aeruginosa* does not only lie in an undisclosed or inappropriate use of antibiotics in the community, but there are also some basic structural conditions in the countries that contribute to this problem. Health spending, which is measured here by the variables GDP_health and Per_cap_US from the public dimension, is correlated with the quantity and quality of prevention, surveillance and public health programs in the countries. Hence, countries where the percentage of GDP allocated to health is greater, will have greater guarantees for the health for the community, as could be reflected in this study when GDP_health was inversely related with the multiresistance.

 doi:10.20944/preprints202110.0127.v1

Meanwhile, out-of-pocket expense is considered the main source of spending in low and middle income countries. This is also related to the lack of contribution from governments for public spending for health, therefore, people individually must spend money for their health. This can be correlated with antibiotic resistance, in two ways: first, the countries with the highest out-of-pocket expenses have a higher direct consumption of non-formulated antibiotics to private entities such as drugstores or pharmacies, due to distrust of the health system and because they do not find guarantees for the care of the population. Second, because they occur more frequently in low-income countries, the unavailability of resources and drug cost overruns can lead people to stop using antibiotics without following the guidelines for proper use.

As we mentioned previously, this is the first study to include HDI as a study variable for antibiotic resistance. In the multiresistance model obtained, HDI is inversely related to antibiotic resistance; that is, the higher human development in a country, the lower the antibiotic resistance. This is important given the characteristics of the indicator. This indicator includes three variables that must be taken into account: life expectancy, schooling, and GDP total. From the initial model where GDP was included, it was considered that it was not a significant variable for antibiotic resistance, and therefore it was not included in the final model. For this reason, it can be considered that the components of life expectancy and schooling may be affecting antibiotic resistance.

The analysis of coefficients in the time series for multiresistance model showed a particular pattern of positive influence on antibiotic resistance since 2014. In comparison, the previous years show a lesser influence (except for 2005). This may be due to multiple

 doi:10.20944/preprints202110.0127.v1

factors that may be related to the recent awareness of the problem of antibiotic resistance and greater reporting of information, in addition to geopolitical and economic problems that have mainly affected the countries of South-East of Europe, which can be related to changes in governance systems, trust in governments and spending on basic goods.

As a recommendation for future works, more studies are required on the effect of these variables, including other bacteria, to study the problem more broadly and to be able to generate alternatives to face this antimicrobial resistance pandemic of the twenty-first century.

5. **Conclusions**

In this work, we first used a multivariate statistical analysis to understand the multiresistance percentage behaviour of *Pseudomona aeruginosa* to five antibiotics commonly used for its treatment using data from the EARSS, the WHO and the World Bank. We used a descriptive analysis to determine the most important characteristics of the data. The results of these analysis show that there is a positive high correlation between the resistance percentages, which indicates that there is a high linear dependence between the variables. This indicated that when there is an increase in the resistance percentage to an antibiotic, it is very likely that there is an increase in the resistance percentage to other antibiotics, and vice-versa. Based on these results, a PCA was used to establish the most influential antibiotic on the data. The Cattell's Scree Test was applied to determine the number of principal components needed to analyse the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2021 doi:10.20944/preprints202110.0127.v1

problematic. The first component showed 20.30% of aminoglycosides contribution, 20.30% of fluoroquinolones contribution, 20% of piperaciclin/tazobactam contribution, 19.90% of ceftazidime contribution and 18% of carbapenems contribution. Thus, we could conclude that the resistance percentage to aminoglycosides and fluoroquinolones are the variables that most contribute to antibiotic resistance problem, while resistance percentage to carbapenems has the minimal contribution to this problematic in the EU/EEA countries. Then, a k-means clustering by region was used to cluster the data. This technique was performed to group the data by countries. By applying the Mojena criterion, it was found that the data was best represented by three clusters, where the minor contribution to the problem is made by the Western countries. In particular, Eastern European countries contribute to the antibiotic resistance of groups such as Aminoglycosides and Fluoroquinolones.

Second, we expanded the database by adding some socioeconomic and antibiotic-consumption variables. In this part, we used data panel regression method to determine functions that relate the multiresistance percentage with those new variables. From those results, we could conclude that the governance variables predict the same behaviour that is already defined in the literature. However, to the best of the authors' knowledge this is the first study to include the Human Development Index (HDI) variable and its inverse relationship with the antibiotic resistance. For the multiresistance panel data model, the most significant coefficients (smaller p-value), in its respective order of significance, corresponded to two governance variables (Rule_law and Ctrl_corrupt), three health expenditure variables (GDP_health, Per_cap_US and Out_pocket_exp), and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2021 doi:10.20944/preprints202110.0127.v1

the variable that corresponds to the indicator of systemic antibiotic consumption in the community (DDD_sys_commun) and the HDI. We can also conclude that there is an inverse relationship between the multiresistance percentage and the variables Ctrl_corrup, Gdp_health and HDI. The multiresistance model presented a $R^2 = 0{,}821638359$ (and a $R^2_{adjusted} = 0{,}796738871$), which indicates that the variables obtained represent approximately 82% of the antibiotic multiresistance of *P. Aeruginosa*. From the validation tests of the model (Dubin Watson, Breusch Pagan, Jarque-Bera), we can conclude that there is no residual autocorrelation (the Breusch Pagan cross-sectional dependency test of Lagrange multipliers with $p - value < 0.05$ strongly rejects the hypothesis of no correlation).

Finally, the Jarque-Bera test for the residuals showed that they do not behave under the assumption of normality. However, due to the number of observations and under a standard regression model and subject to certain regularity conditions, the residuals will behave asymptotically normal.

Tables 8 and 9 show the effects of the coefficients discriminated by time and country. These results were consistent with the results obtained in the cluster analysis of Section 2.3.3, where it could be established that the countries that contribute the most to antibiotic resistance are those found mainly in the South-Eastern region of Europe, particularly Romania, Slovakia and Croatia, and the countries that contribute least to antibiotic resistance are the North-Western countries. It is striking that Cyprus, being a South-Eastern country, is the one that contributes the least to antibiotic resistance, this may be because being an island it is possible that there are no neighborhood effects and

its foreign relations are more common with countries such as Turkey and North Africa than with continental European countries.

We can observe from Table 14 that the highest R^2 value on the validation set was obtained with the Recursive Feature Elimination although the time execution is about 1.5hrs for a five degree polynomial. Because the data size is small, only 401 observations, this method is still valuable. We found that Ctrl_corrup and Rule_law were the most relevant features and Ctrl_corrup and the DDD_sys_common for the multiresistance variable, using the XGBoost Algorithm. We found Ctrl_corrup and Rule_law the most relevant features and the Ctrl_corrup and the DDD_sys_common for the multiresistance variable, using the XGBoost Algorithm.

Acknowledgments

J. Romero thanks to Fundación Ceiba, Colombia. K. Prieto wishes to acknowledge to CONACYT through its program *Cátedras CONACYT*.

References

[1] P. Dadgostar, Antimicrobial resistance: implications and costs, Infection and drug resistance 12 (2019) 3903.

[2] C. W. Hoge, J. M. Gambel, A. Srijan, C. Pitarangsi, P. Echeverria, Trends in antibiotic resistance among diarrheal pathogens isolated in Thailand over 15years, Clinical infectiousdiseases26 (2) (1998) 341345.
URL https://doi.org/10.1086/516303

[3] S. Green, J. Cheesbrough, Salmonella bacteraemia among young children at narural hospital in western zaire, Annals of tropical paediatrics 13 (1) (1993) 45–53.
URL https://doi.org/10.1080/02724936.1993.11747624

[4] M. Pena, B. García, F. Baquero-Artigao, D. Pérez, R. Pérez, A. Echevarría, J. Amador, D. Durán, A. Julian, Actualización del tratamiento de la tuberculosis en niños, in: Anales de pediatría, Vol. 88, Elsevier, 2018, pp.52–e1.
URL https://doi.org/10.1016/j.anpedi.2017.05.013

[5] W. Wu, Y. Jin, F. Bai, S. Jin, Pseudomonas aeruginosa, in: Molecular medical microbiology, Elsevier, 2015, pp. 753–767.
URL https://doi.org/10.1016/B978-0-12-397169-2.00041-X

[6] J. A. Driscoll, S. L. Brody, M. H. Kollef, The epidemiology, pathogenesis and treatment of pseudomonas aeruginosa infections, Drugs 67(3) (2007)351–368.
URL https://doi.org/10.2165/00003495-200767030-00003

[7] Z. Pang, R. Raudonis, B. R. Glick, T.-J. Lin, Z. Cheng, Antibiotic resistance in pseudomonas aeruginosa: mechanisms and alternative therapeutic strategies, Biotechnology advances 37 (1) (2019) 177–192.
URL https://doi.org/10.1016/j.biotechadv.2018.11.013

[8] ECDC, Surveillance of antimicrobial resistance in Europe: annual report of the European antimicrobial resistance surveillance network (ears-net) 2017, S. ECDC; Editor (2018).

[9] J. M. Munita, C. A. Arias, Mechanisms of antibiotic resistance, Virulence mechanisms of bacterial pathogens (2016) 481–511. URL 10.1128/microbiolspec.VMBF-0016-2015

[10] P. Collignon, P. C. Athukorala, S. Senanayake, F. Khan, Antimicrobial resistance: the major contribution of poor governance and corruption to this growing problem, PLoSOne 10 (3) (2015) e0116746.
URL https://doi.org/10.1371/journal.pone.0116746

[11] R. Laxminarayan, D. Sridhar, M. Blaser, M. Wang, M. Woolhouse, Achieving global targets for antimicrobial resistance, Science353 (6302) (2016) 874–875. doi: 10.1126/science.aaf9286.

[12] E. Y. Klein, T. P. Van Boeckel, E. M. Martinez, S. Pant, S. Gandra, S .A. Levin, H. Goossens, R. Laxminarayan, Global increase and geographic convergence in antibiotic consumption between 2000 and 2015, Proceedings of the National Academy of Sciences 115 (15) (2018) E3463–E3470. URL https://doi.org/10.1073/pnas.1717295115

[13] P. Collignon, J. J. Beggs, T. R. Walsh, S. Gandra, R. Laxminarayan, Anthropological and socioeconomic factors contributing to global antimicrobial resistance: a univariate and multivariable analysis, The Lancet Planetary Health 2 (9) (2018) 398–405. URL https://doi.org/10.1016/S2542-5196(18)30186-4

[14] R. B. Cattell, The scree test for the number of factors, Multivariate behavioral research 1 (2) (1966) 245–276. URL https://doi.org/10.1207/s15327906mbr0102_10

[15] F. Franceschi, M. Cobo, M. Figueredo, Discovering relationships and forecasting pm10 and pm2.5 concentrations in Bogotá, Colombia, using artificial neural networks, principal component analysis, and k–means clustering, Atmospheric Pollution Research 9 (5) (2018) 912–922. URL https://doi.org/10.1016/j.apr.2018.02.006

[16] A. Mikulec, A. Kupis-Fija lkowska, Anempirical analysis of the effectiveness of wishart and mojena criteria in cluster analysis, Statistics in Transition new series 3 (13) (2012) 569–580. URL https://www.ceeol.com/search/article-detail?id=444454

[17] C. Hsiao, Analysis of panel data, no. 54, Cambridge university press, 2014.

[18] A. Müller, S. Guido, Introduction to Machine Learning with Python, O'Reilly, 2016.

[19] G. Aurélin, Hands-On Machine Learning with Scikit-Learn, Keras & and Tensor Flow, 2nd Edition, O'Reilly, 2019.

[20] S. M. Lundberg, S.-I. Lee, A Unified Approach to Interpreting Model Predictions, Vol. 30, Curran Associates, Inc., 2017.