TPM Vol. 32, No. S5, 2025
ISSN: 1972-6325
https://www.tpmap.org/

Open Access

PREDICTING AND EXPLAINING CORRUPTION IN THE MENA REGION: A MACHINE LEARNING APPROACH

NADIA FARJALLAH
UNIVERSITY OF SOUSSE, EMAIL nadiafarjallah25@gmail.com

AFTAB HUSSAIN TABASSAM*
DEPARTMENT OF BUSINESS ADMINISTRATION, UNIVERSITY OF POONCH RAWALAKOT, PAKISTAN, EMAIL : aftabtabasam@upr.edu.pk

ABDUL LATIF
DEPARTMENT OF BUSINESS ADMINISTRATION, UNIVERSITY OF POONCH RAWALAKOT, PAKISTAN, EMAIL : abdullatif@upr.edu.pk

NADIA MAHFOUZ
MANAGEMENT SCIENCES, UNIVERSITY OF POONCH RAWALAKOT, PAKISTAN, EMAIL : nadiamahfooz047@gmail.com

HIFSA AKHLAQ
PHD SCHOLAR, RIPHAH INTERNATIONAL UNIVERSITY, ISLAMABAD, PAKISTAN, EMAIL : Hifsaakhlaq@gmail.com

ALTAF HUSSAIN
MS SCHOLAR, UNIVERSITY OF POONCH RAWALAKOT, EMAIL : altafhussai01@gmail.com

UMAR ZAIB
MS SCHOLAR, UNIVERSITY OF POONCH RAWALAKOT, AZAD KASHMIR, PAKISTAN: EMAIL: uk115171@gmail.com

Abstract

Corruption is still pervasive and is considered one of the greatest challenges facing modern societies. Many academic studies have attempted to identify and explain the causes and potential consequences of corruption, primarily through theoretical lenses using correlation and regression-based statistical analyses. The present study approaches the phenomenon from the predictive analytics perspective by employing contemporary machine learning techniques to discover the most important corruption perception predictors based on enriched/enhanced nonlinear models with a high level of predictive accuracy. Specifically, within the regression modeling setting employed herein, the Random Forest (an ensemble-type machine learning algorithm) was found to be the most accurate prediction model, followed by Gradient Boosting and CART. Practically, the increased predictive power of machine learning algorithms coupled with a multi-source database revealed the most applicable corruption-related information, contributing to the body of knowledge and generating actionable information for administrators, academics, citizens, and politicians. The variable importance results indicated that trading across borders, time (men in days), getting electricity, and cost (% of warehouse value) are the most influential factors in defining the corruption level of significance.

1. INTRODUCTION

Since the early 1980s, corruption has been at the heart of international policy and development debate. This corruption is defined by most economists as the abuse of public office for private gain. This corruption has affected all countries, especially developing countries. This corruption can influence all countries, rich or poor, democratic or non-democratic, etc. Corrupt behavior in politics limits economic growth, embezzles public funds, and promotes socio-economic inequality in modern democracies. Over the past 27 years, Ribeiro et al. (2018) have analyzed well-documented political corruption scandals in Brazil, focusing on the dynamic structure of networks in which two individuals are connected if they are involved in the same scandal.

Generally, corruption is a serious crime that weakens the state. According to the international transparency organization, "corruption comes from the behavior of public sector agents, whether politicians or civil servants, who illicitly enrich themselves or their relatives through the abuse of public powers entrusted to them.” Thus, numerous research studies by the World Bank have shown that the causes of corruption can be classified into four categories: economic, social, political, and institutional (Banque Mondiale, 2002[1]). Furthermore, the causes of corruption have been extensively explored through the use of regression-based statistical analysis, but the results remain unclear or at least insufficient to support conclusions.

Many researchers agree that corruption has an enormous deleterious effect on economies. For example, Nuijten and Anders (2017) take a stance that differs in three fundamental ways from the current perspectives in academic and public debates about corruption. First, they do not treat the global anti-corruption industry and dominant social-scientist approaches as frameworks of analysis but rather as subjects of anthropological study t. Second, they do not conceive corruption as an individual act but as a phenomenon that is institutionalized and embedded in the wider matrix of power relations in society. The contextualization of individual acts reveals the systemic and structural dimensions of corruption. Third, and most importantly, we distance ourselves from the commonly held view that corruption is simply the law’s negation, a vice afflicting the body politic. Corruption is still pervasive and is considered one of the greatest challenges facing modern societies. Many academic studies have attempted to identify and explain the causes and potential consequences of corruption, primarily through theoretical lenses using correlation and regression-based statistical analyses. This paper examines the phenomenon from the predictive analytics perspective by employing contemporary machine learning techniques to discover the most important corruption perception predictors based on enriched enhanced nonlinear models with a high level of predictive accuracy. Specifically, within the regression modeling setting employed herein, the Random Forest (an ensemble-type machine learning algorithm) was found to be the most accurate prediction model, followed by Gradient Boosting and CART.

The main objective of this analytical study is to reveal the likely factors and their relative importance as predictors of corruption in the MENA region. To obtain reliable results, we used modern and popular machine learning algorithms. The use of artificial intelligence and machine learning techniques to detect and understand government fraud and corruption has gained popularity in the literature in recent years (Stockemer, 2018; Sun & Medaglia, 2019; Tang et al., 2019). According to de Sun and Medaglia (2019), there is growing interest in studies involving artificial intelligence in the public sector. This study attacks corruption with modern machine learning techniques, potentially reinforcing government strategies toward a cleaner society. The main objective of this analytical study is to reveal the likely factors and their relative importance as predictors of corruption in the MENA region. To obtain reliable results, we used modern and popular machine learning algorithms. The remainder of this paper is organized as follows. Section 2 provides a literature review of corruption. Section 3 summarizes the data acquisition and preprocessing. Section 4 presents the results and discussion. The last section conclusions.

2. REVIEW OF THE LITERATURE ON THE THEORY OF CORRUPTION

The literature has established that corruption not only erodes state legitimacy but also incurs substantial economic and social costs to societies. Studies show that corruption negatively affects both public revenues (Aghion et al., 2016; Besley & Persson, 2014) and public expenditures (Mauro, 1998), thus limiting the state’s ability to carry out its functions. It is also found to lead to lower quality of public investment (Iliopulos & Arnone, 2007) and less private investment (Al-Sadig, 2010; Godinez & Liu, 2015). What is more, corruption can contribute to poor social and environmental outcomes. It causes inefficiencies in public service provision (Reinikka & Svensson, 2006) and can deprive a country of its human capital by fostering emigration to places that are perceived as more meritocratic (Cooray & Schneider, 2016). Corruption can undermine the enforcement of environmental regulations, leading to increased pollution (Pellegrini, 2011) and the overextraction of natural resources (OECD, 2012). Furthermore, data disaggregated by gender show that women tend to suffer more from the negative consequences of corruption than men (Transparency International, 2010).

Over the last quarter century, donors have directed large amounts of resources toward anti-corruption efforts around the world. Reflecting on the achievements of these efforts, many corruption scholars find it difficult to trace major positive results from the anticorruption programming that the World Bank and other international development organizations have launched since the mid-1990s (Rothstein, 2018; Hough, 2017; Mungiu-Pippidi, 2015a; Heeks & Mathisen, 2012). Overall, the empirical picture corroborates this bleak view of the achievements of international anti-corruption efforts. The most widely used global indices that include measurements of

[1] Banque mondiale, 2002.Rapport sur le développement dans le monde 2002 : des institutions pour les marchés.Banque mondiale, Washington, DC, Éditions Eska, Paris.

corruption, such as the Corruption Perception Index (CPI), the Worldwide Governance Indicators (WGI), and the International Country Risk Guide (ICRG), all show that, on a global scale, corruption remains about as prevalent today as it was when the global anticorruption agenda started twenty-five years ago.

Institutional corruption occurs when an institution or its officials receive a benefit that is directly useful to performing an institutional purpose, and systematically provides a service to the benefactor under conditions that tend to undermine procedures that support the primary purposes of the institution (Thompson 2013). This theorist who has taken this turn call attention to the Institutional corruption does not receive the attention it deserves partly because it is so closely (and often unavoidably) related to conduct that is part of the job of a responsible official, the perpetrators are often seen as (and are) respectable officials just trying to do their job, and the legal system and public opinion are more comfortable with condemning wrongdoing that has a corrupt motive. Miller (2011) contends that institutional corruption necessarily involves corrupt individuals who are conscious of their deleterious behavior. There seems to be some confusion in the extant literature regarding boundary limits between the concepts of institutional and individual corruption. According by Thompson (2018), normative theorists of corruption have developed an institutional conception that is distinct from both the individualist approaches focused on quid pro quo exchanges and other institutional approaches found in the literature on developing societies. These theorists emphasize the close connection between corruption patterns and the legitimate functions of institutions. Also, institutional corruption does not require that its perpetrators have corrupt motives, and it is not limited to political institutions.

Few papers examine the relation between economic freedom and corruption. The empirical results of these studies are consistently the same: the more freedom, the lower the level of corruption, implying that economic freedom is a deterrent to corruption (Chafuen and Guzmán, 2000, Paldam, 2002). In this sense, Graeff and Mehlkop (2003) imply that there is a strong relationship between economic freedom and corruption. This relationship depends on a country's level of development. They identify a stable pattern of aspects of economic freedom influencing corruption that differs depending on whether countries are rich or poor.

Moreover, a large body of academic studies has attempted to identify and explain the potential causes and consequences of corruption, at varying levels of granularity, mostly through theoretical lenses by using correlations and regression-based statistical analyses. According by Marcio and Dursun (2020), the phenomenon from the predictive analytics perspective by employing contemporary machine learning techniques to discover the most important corruption perception predictors based on enriched nonlinear models with a high level of predictive accuracy.

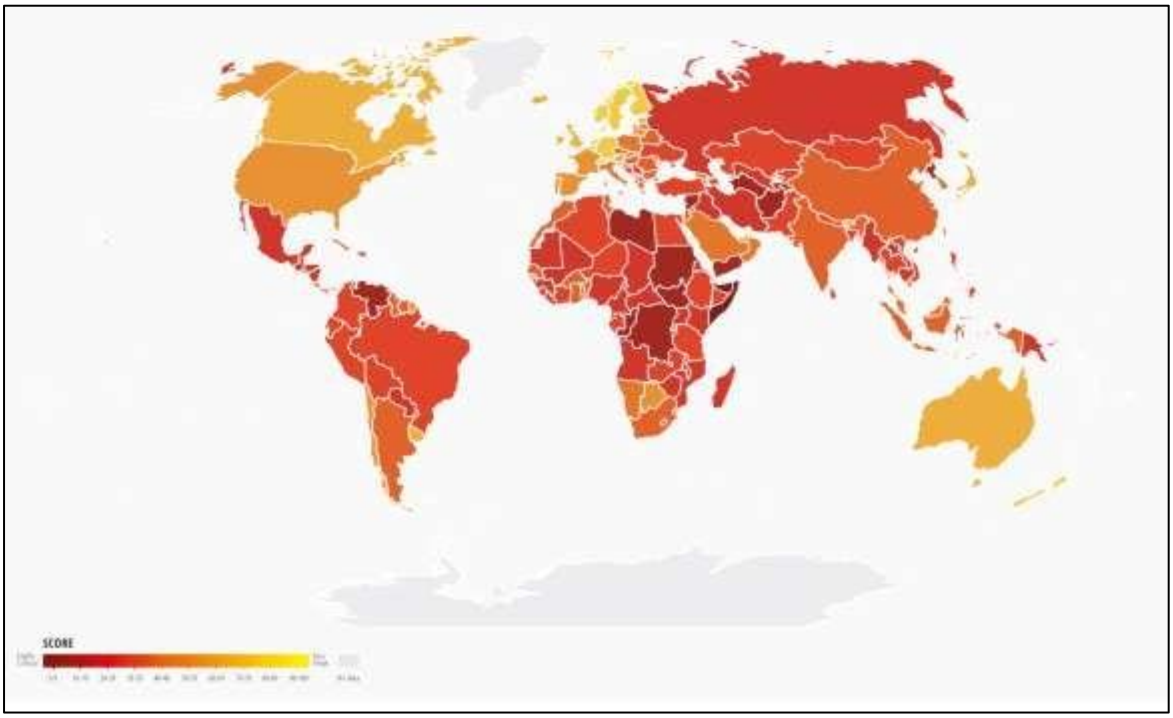
3. Data acquisition and data preprocessing

The data was acquired from several sources, including Ease of Doing Business Indexes [2] ,Transparency International[3],the Human Development Reports of the United Nations Development program[4] and the World Bank[5] for the year 2019 and 2020 from 17 countries in the MENA region (annex1). Transparency International is a global movement that works in over 100 countries to end corruption. They work to expose the systems and networks that allow corruption to flourish, demanding greater transparency and integrity in all areas of public life. The Corruption Perception Index (CPI) sorts 180 countries according to their perceived corruption levels. The index captures the assessments of domain experts on corrupt behavioural information, originating a scale from 0 to 100 where economies close to 0 are perceived as highly corrupt while economies close to 100 are perceived as less corrupt. The Human Development Report aims to support country-related analyses by collecting and exploring data from many countries. The Education Index represents the average mean years of schooling of adults and the expected years of schooling of children. The index uses a scale related to the corresponding maxima[6].

The Doing-Business project contains measures of trade regulations and the efficiency with which these regulations are enforced and provides data from 190 countries. It has been used in several academic studies in several research areas, such as politics, economics, and law (Roe and Siegel, 2009; Dixit, 2009). The information provided by the Ease of Doing Business Project is mainly related to starting a business, paying taxes, trading across borders, obtaining credit, and registering properties. The Ease of Doing Business ranking system assesses economies by comparing the distance-to-frontier score to benchmark countries for regulatory best practices. Essentially, it is based on systematic comparisons between countries with a baseline that is drawn from the country with the best and most efficient economic practice for a certain indicator. On the 2019 Doing Business questionnaire, survey

[2] www.doingbusiness.org
[3] www.transparency.org
[4] www.undp.org
[5] www.worldbank.org
[6] hdr.undp.org/en/indicators/103706.

responses were collected from over 13,000 local experts, including public officials, lawyers, and business specialists. A sample item of a binary variable is the Quality Control Before Construction – whether licensed or technical experts approve building plans. For this type of variable, where the only possible scores are 0 and 1, the output was converted to a string-type variable before computing the model results. A sample item that is treated as a continuous variable is the Time Required to Complete Each Procedure – Registering Property. For this type of variable, the score captures the median time that field professionals take to complete all necessary procedures to register properties. Scores are computed in calendar days. The distance to frontier score takes into account the simple average of scores on indicators and measures how far or how close an economy is to the most efficient practice score. For more details on the Ease of Doing Business methodology, please visit the official website.[7]
The World Bank is a financial institution that provides loans and grants. With 189 member states, staff from over 170 countries, and over 130 branches worldwide, the World Bank Group is an unparalleled partnership of five institutions working together to find lasting solutions to reduce poverty and promote prosperity in developing countries. Government Effectiveness captures perceptions of the quality of public services, civil service, degree of independence from political pressures, policy formulation and implementation, and credibility of the government's commitment to such policies. Regulatory Quality captures perceptions of the government’s ability to formulate and implement sound policies and regulations that permit and promote private sector development. The rule of law captures perceptions of the extent to which agents have confidence in and abide by the rules of society, particularly the quality of contract enforcement, property rights, the police, and the courts, as well as the likelihood of crime and violence. Voice and Accountability capture perceptions of the extent to which a country's citizens can participate in selecting their government, as well as freedom of expression, freedom of association, and free media. The scores for each variable varied between -2.5 and 2.5. Table 1 presents the names and descriptive statistics (including the Shapiro-Wilk p-value for the normality test) of all variables included in this analytical study. In general, authors have shown (Delen et al., 2012; Sharda et al., 2017) that machine learning algorithms can extract deeply hidden knowledge from large datasets involving several types of input variables that are not necessarily normally distributed. Unlike traditional stochastic data models, the machine learning community assumes that data are generated in complex ways that are not necessarily normally distributed or linearly correlated. The only assumption for algorithmic cultivation is that data generated by natural processes follow an unknown multivariate distribution (Breiman, 2001). Our research contributes to the corruption literature by using state-of-the-art machine learning techniques to identify the most important predictors of perceived corruption in economies. Fig 1 uses average CPI scores from 2019 to create a color scale (from red to blue) to illustrate and differentiate the high and low levels of corruption across countries.



[7] www.doingbusiness.org/methodology

Fig.1. Corruption Perception Index across countries. Dark yellow countries (higher CPI scores) are less corrupt economies, while dark red countries (low CPI scores) are more corrupt economies. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

3. Machine Learning Algorithmes

The purpose of machine learning is to analyze the information available from a large number of statistical data to learn how to carry out a study without being specifically programmed beforehand. She has achieved great success in the past few years. These flexible machine learning techniques have the potential to capture complex, nonlinear, interactive selection patterns. However, to the best of our knowledge, their performance in analyzing missing context data has never been evaluated. The forecasts are mediocre and associated uncertainties models can increase construction costs[8]. These costs can be reduced by predicting the stability number closer to actual conditions[9]. Therefore, a meaningful and robust model can be both cost-effective and time-saving. Therefore, advanced machine learning algorithms have been studied to overcome the limitations mentioned above.

3.1. Random Forest

The random forest is an efficient and widely used tree-based ensemble learning method. Random forest algorithms depend in concept on bagging[10]. It uses sampling with replacement and develops multiple decision trees from the training data set. The decision, obtained from the maximum number of trees, is considered as the final result of the corresponding random forest model (RFM)[11]. This allowed us to reduce the prediction deviation with respect to an individual tree. Therefore, the model becomes more generalized. First, a bootstrapped dataset is created by choosing random inputs from the main dataset, and a decision tree is built based on this dataset. Other trees are developed in the same way. Specifically, each tree depends on a random value of the number of predictors that is chosen as the split input at each iteration. Consequently, the best distribution was found when building a tree. This randomness decreases the variance in the forest estimates. Generally, an individual decision tree tends to overfit and thus has a high variance in the prediction. Random forest produces reduced variance with its wide variety of trees and shows high accuracy because it combines single trees and takes an average of their predictions. Out-of-bag observations, which were not retained in the sampling process, are used in random forest to estimate the error rate [12]. The random forest model performs well in predicting both numerical and categorical variables, and it can estimate the missing values in the dataset. Moreover, it can handle complex interactions and noisy data. A schematic of the random forest algorithm is shown in Fig. 2.

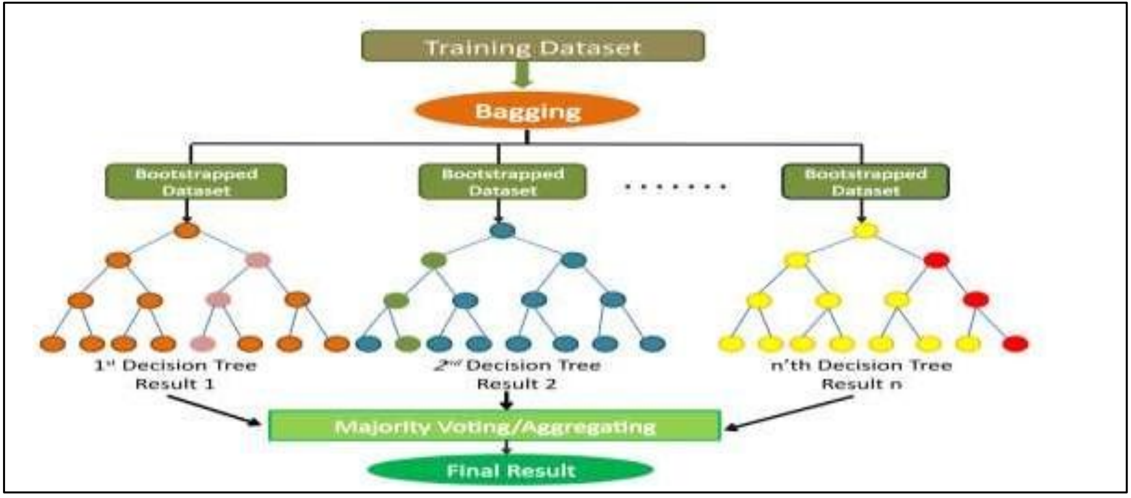


Fig.2. schematics of random forest algorithms

3.2. Gradient Boosting

Generally, gradient boosting is a tree boosting procedure composed of arbitrary differentiable loss functions. Il a été reconnu comme une technique d'apprentissage automatique efficace pour les problèmes de classification et de régression. Le modèle de régression est utilisé pour prédire la valeur continue[13]. Gradient boosting regressors

[8] D.H. Kim, W.S. Park, Ocean Eng. 32 (2005) 1332–1349.

[9] T. Erdik, Expert Syst. Applic. 36 (2009) 4162–4170.

[10] G. Louppe, arXiv:1407.7502v3 (2015).

[11] L. Breiman, Mach. Learn. 24 (2) (1996) 123–140

[12] L. Breiman, Mach. Learn. 24 (2) (1996) 123–140.

[13] J.P. Bieman, J.M. Wilms, H. Boogaard, Water (Basel) 12 (6) (2020). 2073-4441(1703)

use several decision trees of fixed size to build an additive model. This is the main difference between the gradient boosting regressor and conventional AdaBoost model. Usually, decision stumps with one node and two leaves are used in the AdaBoost algorithm. The model fitting procedure began with a leaf as the mean value of the target variable. Then, a tree is added based on the obtained residuals, and the contribution of the tree is scaled in the next step with a learning rate until the final estimate is obtained. The other trees are added by considering the new residuals based on the error obtained from the previous trees. In this way, the decision trees are adjusted to estimate the negative gradient of the samples, and for each subsequent estimator, the gradients are updated in each operator. The schematic idea of gradient-boosting regression algorithms is illustrated in Fig. 3

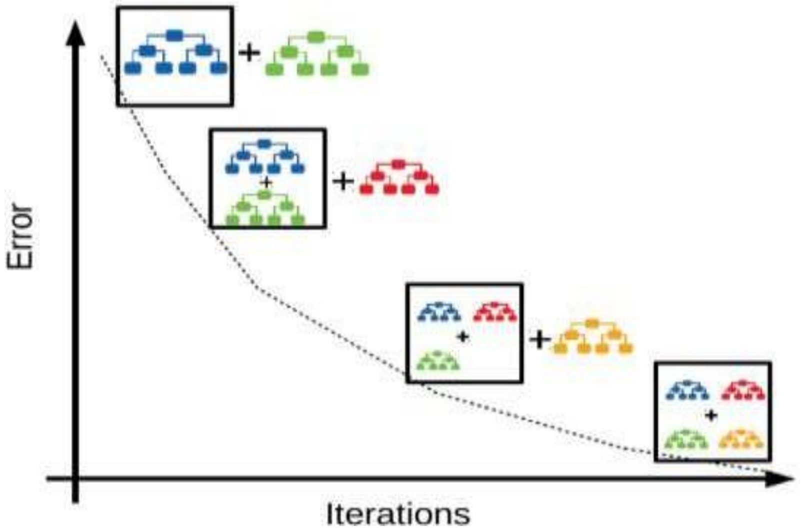


Fig. 3. A schematic of the gradient boosting regression algorithms

3.3. CART (Classification And Régression Trees)

This model was the first to use regression trees for machine learning. It is very simple and not very efficient; however, it serves as the basis for several more elaborate machine learning models, such as bagging and random forest models.

* Maximum tree: If the decision tree is a regression tree, the variable to be explained is a continuous variable. We want to determine the value of a quantity of interest. On note:

-Y : The answer variable.

- p : The number of covariates.

-X_j, 1≤ j ≤ : the covariates.

-π_0 : The amount of interest to be predicted.

$\pi_0 = [Y/X \quad = x]$

The expectation is the solution of the minimization of the quadratic error, and the quantity of interest chosen is the solution of the following equation:

$$\pi(x) = \arg \underbrace{min}_{\pi(x)} E[\emptyset(Y,\pi(x))/X=x]$$

With: $\emptyset(Y,\pi(x)) = (Y-\pi(x))^2$

To build the tree, each node is segmented into two child nodes, minimizing the variance of the two new nodes. At each constructed node, the new estimator of E[Y] becomes the empirical expectation of all the observations of the node.

3.4. Testing and evaluating – cross - validation

For these methods, we used k-fold cross-validation to randomly divide the data into k mutually exclusive subsets for “training” and “testing” sets. The folds (k-1) of the data are used to build the model, and the remaining fold is used to test it. Delen et al. (2012) proved that a single random assignment can potentially lead to heterogeneous subsets of data, which, in turn, would produce biased results. Therefore, we used five rounds (k = 10) of cross-validation on the entire dataset. In each round of the 5-fold cross-validation, the model was trained in all but one of the folds and tested on the excluded fold, which was the test subset for that specific round. Finally, the average of the results of the five rounds was compiled for the final analysis. Olson and Delen (2008) reported that the use of stratified cross-validation tends to decrease bias compared to regular cross-validation. According to Delen et al. (2012), the overall accuracy is measured using the average of each k accuracy measurement.

$CV = \frac{1}{}\sum_{=1}\ {}_{i}$ (5)

3.5. Model Performance Comparison

To evaluate and compare the prediction performance of the four machine learning algorithms, we chose to use the mean squared error (MSE), which is defined as the mean for each individual of the deviation test basis squared between the prediction of the model to be tested and the true output value. For each model, we evaluated the following parameters:

$MSE_j = \frac{1}{n_{test}}\sum_{i=1}^{n}(y - \hat{y}_{i,j})$ (6)

Where:

- m~~: the~~: number of models to be tested.
- n~~: the~~: number of individuals in the initial base.

-n_{tes} : the number of individuals in the test database.

- y : the actual output variable of the individual.

-$\hat{y}_{i,}$: the output variable of individual i predicted by model j.

We are interested in the model which has MSE_j: min $(MSE_j)_{1\le j\le m}$. The model that minimizes the MSE appears to be the best predictor of the dependent variable.

4. RESULTS AND DISCUSSIONS

The objective is to build artificial intelligence models to predict corruption in the public sector. To achieve this, we used three machine learning approaches: **Random Forest (RF)**, **Gradient Boosting (GB),** and **CART.** The training and test data were split, with 70% of the data used to train the model and 30% used to examine the model efficiency and accuracy. To compare the performance of the models, we used the mean square error (MSE). In general, the smaller the error, the more accurate the model. Table 2 shows that the learning method **Random Forest** improves the accuracy of the prediction method, achieving an overall accuracy of **99.66%. Gradient Boosting (GB)** was the second best model with an overall accuracy of 93.79% and **CART** with 92.30%. The Random Forest method remains a "black box" method since we cannot visualize the decision tree, allowing us to obtain the final prediction. This type of model provides excellent predictive performance, is versatile, and detects interactions without having to specify a parametric form (Hamza and Larocque, 2005). Graphs (4) explain the mechanism of the Random Forest algorithm to avoid overfitting. Generally, the root mean square error of the training and test data decreases. To assess the most influential factors on corruption in the MENA region, we used the random forest-based predator importance technique.

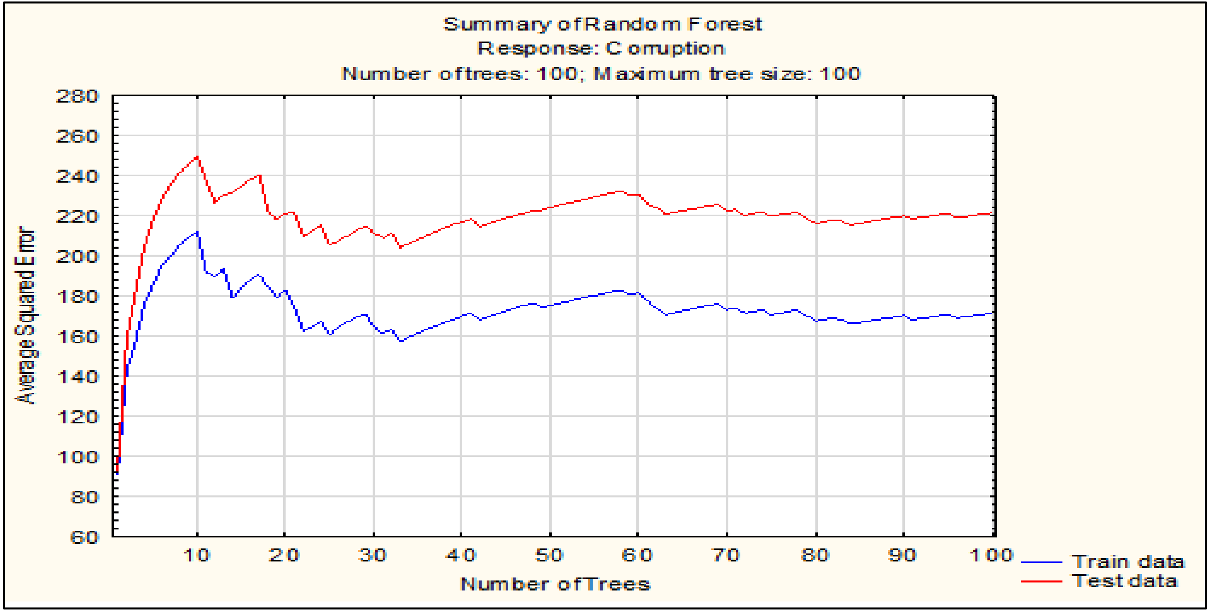


Fig 4. Summary of the random forest response.

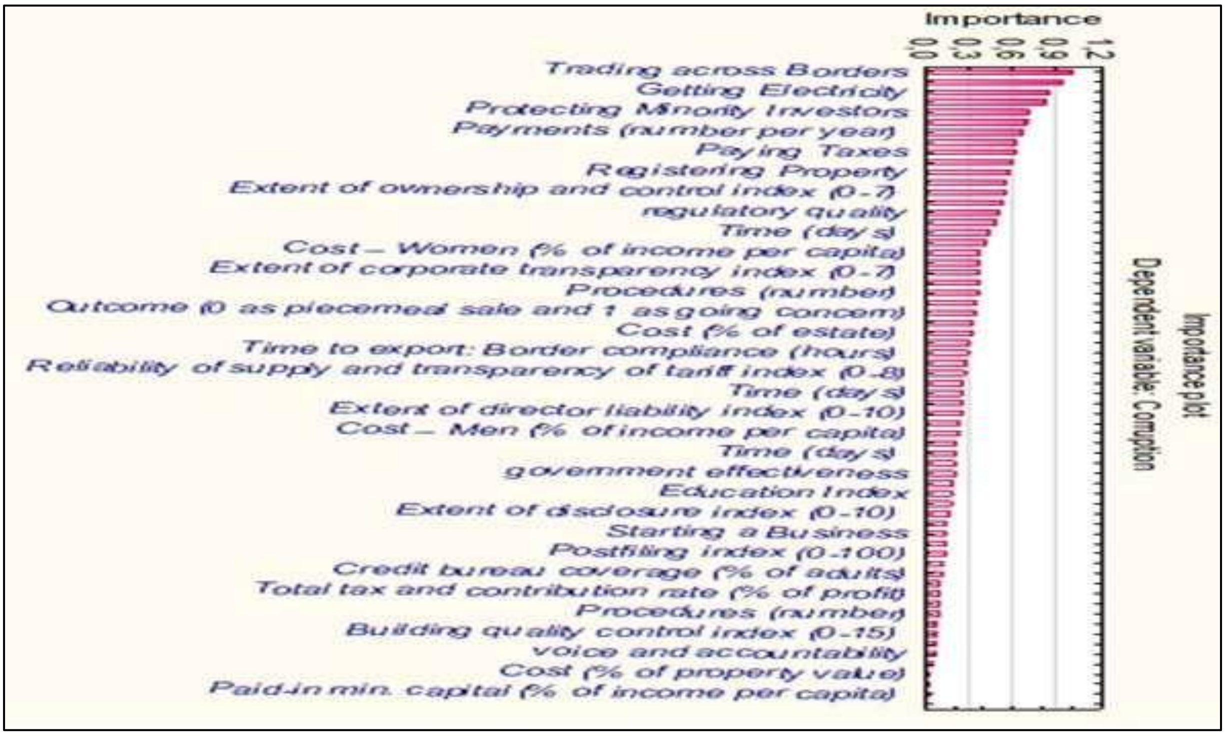


Fig 5. Importance of variables

Understanding the nature of data is one of the most important objectives in the field of machine learning (ML). An approach to estimating the relative influences of input characteristics on corruption is presented in this study by introducing feature analysis using the proposed ensemble learning models. Erdik[14] mentioned that the success of these forecasting models can be increased by understanding more intensively the stability parameters. Feature selection is a procedure that identifies a subset of the original input data set. This feature selection property is very effective in building a good prediction model and understanding the physical knowledge of the dataset. There are many benefits to a good feature importance analysis. To the best of the author's knowledge, no strong and compact feature analysis is considered in any of the existing literatures related to the corruption. Feature selection can reduce the dimensionality of the prediction problem with appropriate logic, which will speed up the ML algorithms. It reduces storage requirements and improves the accuracy of the prediction models. The underlying relationships between the corruption variables in the corresponding dataset can be better understood by identifying the most relevant variables. Therefore, feature analysis provides deep insight into the dataset and improves the understanding of the data. In addition, the easiest way to meet the storage requirements and achieve the required speed is to retain only those variables in the dataset that are more important than the response variables. Incorporating variable importances into a dedicated iterative feature selection procedure can yield more accurate predictions at the cost of a small computational effort. Trading across Borders (100%), Time - Men (94%), Getting Electricity (85%), and Cost (% of warehouse value, 84%) are the most influential factors in defining the corruption level of significance.

CONCLUSIONS

In this study, we searched for several potential predictors of the Corruption Perceptions Index in 17 MENA countries. We chose to use variables provided by the World Bank, Transparency International, and the Heritage Foundation for 2019 and 2020. After several experiments, we chose the methods Gradient Boosting, Classification and Regression Trees (CART) and Random Forest. The cross-validation results showed that the Random Forest was the most accurate classification method for predicting CPI. Gradient Boosting and CART were, respectively, the second and the third best models, showing satisfactory prediction performances. Our results prove that trading across borders is the most important predictor of corruption. On the other hand, machine learning models, specifically speaking of classification-type algorithms such as Random Forest, have the capability of revealing important predictors regardless of significant linear correlations and complex relationships between the input variables.

[14] T. Erdik, Expert Syst. Applic. 36 (2009) 4162–4170.

Most machine learning algorithms are considered predictive instruments with limited descriptive capabilities. Although the predictive accuracy of machine learning models has been consistently higher than that of traditional regression models, their process of operation has been referred to as a "black box" by some researchers because a machine learning model is trained byby assuming that the data are generated in a complex way. It is necessarily correlated to produce accurate predictions of a certain outcome. In other words, in machine learning, the relationship between the input and output variables can only be inferred using heuristic methods of experimentation, with an emphasis on predictive accuracy.

Customs and trade regulations appear to be greater obstacles for businesses in the MENA region than in other countries. Businesses need more time to clear customs to import or export than they do in other countries. The MENA region depends on high levels of imports compared to low levels of exports. To ensure sustainable growth in the region's private sector, the report calls on the MENA region to lower regulatory barriers for businesses, promote competition, and reduce disincentives resulting from political influence and informal business practices. The region also needs reforms to facilitate innovation, the adoption of digital technologies, and investments in human capital, while being in line with the global agenda to limit climate change, enhance sustainability, and protect the natural environment. Therefore, the lack of explicit theoretical explanations regarding the relationships between variables (strength and direction of influence) can be considered a limitation of predictive analytics studies, including the one proposed in this study.

Table1. List of variables and their Descriptive Statistics

Variables	Observations	Mean	Min	Max	Std Dev	Median	W	Prob« W
Corruption	34	41,0000	15,0000	71,000	16,2033	42,0000	0.96197	0.27737
Education Index	34	0,7074	0,3500	0,919	0,1260	0,7210	0.96974	0.45424
Starting a Business	34	84,0971	67,0000	94,800	8,5421	85,4000	0.91986	0.1601
Procedure – Men (number)	34	6,1176	2,0000	12,000	2,8473	6,0000	0.93675	<0.04937
Time – Men (days)	34	18,3529	3,5000	72,000	16,8853	12,0000	0.72928	<0.000
Cost – Men (% of income per capita)	34	14,1882	1,0000	42,300	14,0317	6,3000	0.82734	<0.0000
Procedure – Women (number)	34	6,7647	3,0000	12,000	2,8183	7,0000	0.93553	<0.04544
Time – Women (days)	34	19,0000	4,5000	73,000	16,9393	13,0000	0.72231	<0.0000
Cost – Women (% of income per capita)	34	14,1882	1,0000	42,300	14,0317	6,3000	0.82734	<0.0000
Paid-in min. capital (% of income per capita)	34	5,2412	0,0000	41,500	12,0021	0,0000	0.49870	<0.000
Dealing with Construction Permits	34	64,4294	0,0000	89,800	25,4309	71,5500	0.69765	<0.000
Procedures (number)	34	14,7941	9,0000	22,000	4,0585	14,0000	0.91924	<0.01537
Time (days)	34	120,5588	47,5000	276,000	58,0771	114,0000	0.89920	<0.00437

Cost (% of warehouse value)	34	3,5441	0,1000	12,100	3,1448	3,3000	0.94144	0.06807
Building quality control index (0-15)	34	11,9706	4,5000	15,000	2,7577	12,5000	0.92711	0.0258
Getting Electricity	34	72,3500	0,0000	100,000	21,0710	76,5500	0.88526	<0.00191
Procedures (number)	34	4,4706	2,0000	6,000	1,1074	5,0000	0.93024	<0.003179
Time (days)	34	57,5294	7,0000	118,000	29,3611	53,0000	0.90064	<0.0477
Cost (% of income per capita)	34	308,1765	0,0000	1308,800	388,3218	128,0000	0.88974	0.00248
Reliability of supply and transparency of tariff index (0-8)	34	5,0000	0,0000	8,000	2,4863	6,0000	0.89125	<0.00272
Registering Property	34	64,9647	0,0000	96,200	20,5760	66,5000	0.95631	0.18944
Procedures (number)	34	5,4412	1,0000	10,000	2,6651	6,0000	0.85941	<0.00046
Time (days)	34	25,7353	1,0000	76,000	20,7201	19,5000	0.73077	<0.0000
Cost (% of property value)	34	4,1706	0,0000	9,000	3,0564	6,0000	0.89994	<0.00457
Quality of the land administration index (0-30)	34	16,1471	7,0000	26,000	5,3563	17,0000	0.84766	<0.00030
Getting Credit	34	41,1559	0,0000	95,000	24,2137	45,0000	0.86447	<0.01796
Strength of legal rights index (0-12)	34	2,8824	0,0000	11,000	2,8044	2,0000	0.84925	<0.01044
Depth of credit information index (0-8)	34	5,7647	0,0000	8,000	3,3126	8,0000	0.64990	<0.00014
Credit registry coverage (% of adults)	34	13,5235	0,0000	60,300	17,0603	5,0000	0.83672	<0.0000
Credit bureau coverage (% of adults)	34	25,1176	0,0000	100,000	29,1078	22,9000	0.64339	<0.0001
Protecting Minority Investors	34	51,8235	18,0000	86,000	20,1050	54,0000	0.93463	<0.0005
Extent of disclosure index (0-10)	34	6,4706	2,0000	10,000	2,2861	7,0000	0.78761	<0.02580

Extent of director liability index (0-10)	34	4,7059	1,0000	10,000	3,0104	4,0000	0.81267	<0.00191
Ease of shareholder suits index (0-10)	34	4,4118	1,0000	9,000	1,9403	4,0000	0.94144	<0.03179
Extent of shareholder rights index (0-6)	34	3,2941	0,0000	6,000	1,9311	4,0000	0.92711	0.10382
Extent of ownership and control index (0-7)	34	3,8235	0,0000	7,000	2,4429	4,0000	0.88526	0.06560
Extent of corporate transparency index (0-7)	34	3,7647	0,0000	7,000	2,4502	4,0000	0.93024	0.06951
Paying Taxes	34	75,0559	40,0000	100,000	16,2081	74,1000	0.90064	0.29476
Payments (number per year)	34	14,3529	3,0000	44,000	10,8149	12,0000	0.88974	<0.02337
Time (hours per year)	34	203,9647	10,4000	889,000	201,1199	155,0000	0.89125	<0.00038
Total tax and contribution rate (% of profit)	34	30,8765	11,3000	66,100	16,2346	27,4000	0.95631	<0.0046
Postfiling index (0-100)	33	54,7455	19,0000	98,600	28,6773	49,8000	0.84766	<0.0003
Trading across Borders	34	61,0647	0,0000	85,600	22,9799	68,8500	0.85951	<0.00046
Time to export: Border compliance (hours)	33	54,3333	6,0000	101,000	30,2713	53,0000	0.93129	<0.03810
Cost to export: Border compliance (USD)	33	358,2424	47,0000	1118,000	267,5850	319,0000	0.87663	<0.00139
Time to export: Documentary compliance (hours)	33	65,4545	3,0000	504,000	119,6378	24,0000	0.51337	<0.0000
Cost to export: Documentary compliance (USD)	33	227,7576	50,0000	1800,000	413,3370	100,0000	0.40450	<0.0000

Time to import: Border compliance (hours)	33	97,5152	39,0000	240,000	61,3505	72,0000	0.80825	<0.0005
Cost to import: Border compliance (USD)	33	486,7273	206,0000	790,000	174,2753	553,0000	0.92361	<0.02310
Time to import: Documentary compliance (hours)	33	73,0909	7,0000	265,000	64,0568	60,0000	0.78742	<0.0002
Cost to import: Documentary compliance (USD)	33	254,3636	60,0000	1000,000	228,3422	144,0000	0.72430	<0.000
Enforcing Contracts	34	56,5912	40,0000	75,900	8,1126	57,7500	0.95425	0.16454
Time (days)	34	605,7059	5,0000	1010,000	210,9447	598,0000	0.83153	<0.00011
Cost (% of claim value)	34	37,2176	14,7000	248,000	53,7263	26,2000	0.33789	<0.0000
Quality of judicial processes index (0-18)	34	7,1765	1,5000	14,000	3,4198	6,5000	0.93427	<0.04171
Resolving Insolvency	34	35,8206	0,0000	72,800	20,1564	39,2500	0.90724	<0.00716
Recovery rate (cents on the dollar)	34	37,7412	15,0000	62,600	12,8168	34,1500	0.93678	<0.04949
Time (years)	34	4,3029	1,0000	27,700	6,2199	2,9000	0.4449	<0.0000
Cost (% of estate)	34	12,9529	3,2000	23,000	6,3711	10,0000	0.88443	<0.00182
Outcome (0 as piecemeal sale and 1 as going concern)	34	1,2353	0,0000	20,000	4,7677	0,0000	0.27292	<0.0000
Strength of insolvency framework index (0-16)	34	8,6029	4,0000	12,500	2,7074	8,2500	0.91104	<0.00910
rule of law	34	-0,2342	-2,0000	1,048	0,9428	0,0812	0.91586	<0.01236
regulatory quality	34	-0,3249	-2,3471	1,281	1,0159	-0,0841	0.95745	0.20474
government effectiveness	34	-0,2647	-2,3000	1,377	0,9857	-0,1112	0.95696	0.19795
voice and accountability	34	-0,9384	-1,8000	0,700	0,6383	-1,1135	0.87713	<0.00120

Note. The Prob « W value listed in the last column is the p-value. The Shapiro-Wilk p-value tests the null hypothesis that the data are normally distributed.

Table 2: Performance analysis of ensemble learning-based corruption prediction models.

Model	**MSE**
Random Forest	**0.003413**
Gradient Boosting	0.062134
CART	0.070208

REFERENCES

1. Aghion, P., Akcigit, U., Cagé, J. and Kerr, W. (2016). Taxation, Corruption, and Growth. NBER Working Paper 21928, Cambridge, Massachusetts: National Bureau of Economic Research.
2. Al-Sadig,A.,(2010). Corruption and Private Domestic Investment: evidence from developing countries, International Journal of Economic Policy in Emerging Economies 3(1):47-60, DOI:10.1504/IJEPEE.2010.032794
3. Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32. https://doi.org/10. 1023/A: 1010933404324.
4. Breiman, L., Friedman, J., Olshen, R. Stone, C. [1984] : Classification And Regression Trees.
5. Besley, T. and Persson, T. (2014). Why Do Developing Countries Tax So Little? *Journal of Economic Perspectives*, 28(4), pp. 99–120.
6. Cooray and Schneider, 2016, Does Corruption Promote Emigration? An Empirical Examination, Journal of Population Economics 29(1), DOI:10.1007/s00148-015-0563-y.
7. Chafuen, A., Guzman, E. (2000). Economic Freedom and Corruption. In: O'Driscoll, G.P., Holmes, K.R., Kirkpatrick, M. (Eds.), Index of Economic Freedom. The Heritage Foundation, Washington D.C., 51-63. Retrieved from https://mpra.ub.uni-muenchen.de/18731/.
8. Delen, D., Cogdell, D., & Kasap, N. (2012). A comparative analysis of data mining methods in predicting NCAA bowl outcomes. International Journal of Forecasting, 28, 543–552. https://doi.org/10.1016/j.ijforecast.2011.05.002.
9. Dixit, A. (2009). Governance institutions and economic activity. American Economic Review, 99, 5–24. https://doi.org/10.1257/aer.99.1.5.
10. Friedman, T. Hastie, R. Tibshirani, Ann Stat 28 (2) (2000) 337–407.
11. Graeff, P., & Mehlkop, G. (2003). The impact of economic freedom on corruption: Different patterns for rich and poor countries. European Journal of Political Economy, 19, 605–620. https://doi.org/10.1016/S0176-2680(03)00015-6.
12. Godinez, Jose R. & Liu, Ling, 2015. "Corruption distance and FDI flows into Latin America," International Business Review, Elsevier, vol. 24(1), pages 33-42.
13. Hamza, M.,Larocque,D. 2005. “An Empirical Comparison of Ensemble Methods Based on Classification Trees.” Journal of Statistical Computation and Simulation 75 (8): 629–43. https://doi.org/10.1080/00949650410001729472.
14. Hough, Dan (2017) Analysing corruption. Agenda, Newcastle upon Tyne. ISBN 9781911116547
15. Miller, S. (2011). Corruption. In E. N. Zalta (Ed.). Stanford encyclopedia of philosophy http:// plato.stanford.edu/archives/spr2011/entries/corruption/.
16. Mungiu-Pippidi, A. (2015a). Corruption: Good governance powers innovation. Nature News, 518(7539), 295.

17. Marcio, S, M, L; Dursun,D.(2020) Predicting and explaining corruption across countries: A machine learning approach Government Information Quarterly Volume 37, Issue 1, January 2020, 101407
18. Nuijten, M., & Anders, G. (2017). Corruption and the secret of law: An introduction. Corruption and the secret of law (pp. 1–24). New York: Routledge.
19. Ribeiro, H. V., Alves, L. G., Martins, A. F., Lenzi, E. K., & Perc, M. (2018). The dynamical structure of political corruption networks. Journal of Complex Networks, 6, 989–1003. https://doi.org/10.1093/comnet/cny002.

20. Reinikka,R., Svensson,J.(2006).Using Micro-Surveys to Measure and Explain Corruption, Elsevier, Volume 34, Issue 2, February 2006, Pages 359-370.
21. Richard Heeks, Harald W. Mathisen, (2012), Understanding success and failure of anti-corruption initiatives, Economics.
22. Roe, M. J., & Siegel, J. I. (2009). Finance and Politics: A Review Essay Based on Kenneth Dam’s analysis of Legal Traditions in the Law-Growth Nexus. Journal of Economic Literature, 47, 781-800. https://doi.org/10.1257/jel.47.3.781.
23. Sharda, R., Delen, D., & Turban, E. (2017). Business intelligence, analytics, and data science: A managerial perspective (4th ed.). London: Pearson.
24. Shahhosseini, M.,H. Guiping, S.V. Archontoulis, Front. Plant Sci. 11 (2020) 1120.
25. Stockemer, D. (2018). The internet: An important tool to strengthening electoral integrity.Government Information Quarterly, 35(1), 43–49.
26. Sun, T. Q., & Medaglia, R. (2019). Mapping the challenges of artificial intelligence in the public sector: Evidence from public healthcare. Government Information Quarterly, 36(2), 368–383.
27. Tang, Z., Chen, L., Zhou, Z., Warkentin, M., & Gillenson, M. L. (2019). The effects of social media use on control of corruption and moderating role of cultural tightness-looseness. Government Information Quarterly (in press).
28. Thompson, D. (2018). Theories of institutional corruption. Annual Review of Political Science, 21, 495–513. https://doi.org/10.1146/annurev-polisci-120117-110316.
29. Thompson DF. 2013. Two concepts of corruption. Safra Res. LabWork. Pap., Harvard Univ., Cambridge, MA.Soc. Sci. Res. Netw. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2304419
30. Transparency International. 2010. Regulating the revolving door. Transparency Int. Work. Pap.No.06/2010.https://www.transparency.org/whatwedo/publication/working_paper_06_2010_regulating_the revolving door.
31. Pellegrini, L. (2011) Chapter 4: The Effect of Corruption on Growth and Its Transmission Channels. In: Pellegrini, L., Ed., Corruption, Development and the Environment, Springer, Berlin, 53-74.
32. Paldam, M. (2002). The Big Pattern of Corruption: Economics, Culture and the Seesaw Dynamics. European Journal of Political Economy, 18(2), 215-240.
33. Rothstein,B.,(2018), Fighting Systemic Corruption: The Indirect Strategy, Vol. 147, No. 3, Anticorruption: How to Beat Back Political & Corporate Graft, pp. 35-49 (15 pages), https://www.jstor.org/stable/48563079
34. Iliopulos, E., & Arnone, M. (2007). The cost of corruption The cost of corruption. Milan, Italy: V&P
35. OECD. (2012). Illegal trade in environmentally sensitive goods Illegal trade in environmentally sensitive goods. OECD Publishing. Paris, France.

Annex 1
Algeria
Yemen
Bahreïn
Egypt
Iran
Iraq
Israel
Jordan
Kuwait
Libye
Maroc
Oman
Qatar
Saudia
Tunis
Enirate
Lebanon