Government Information Quarterly 37 (2020) 101407

Contents lists available at ScienceDirect

Government Information Quarterly

journal homepage: www.elsevier.com/locate/govinf

Predicting and explaining corruption across countries: A machine learning approach

Marcio Salles Melo Lima[a], Dursun Delen[b,*]

[a] Research and Development, Metalsider, Brasil, and Spears School of Business, Oklahoma State University, USA
[b] Department of Business Analytics, Department of Management Science and Information Systems, Spears School of Business, Oklahoma State University, USA

ARTICLE INFO

Keywords:
Corruption perception
Machine learning
Predictive modeling
Random forest
Society policies and regulations
Government integrity
Social development

ABSTRACT

In the era of Big Data, Analytics, and Data Science, corruption is still ubiquitous and is perceived as one of the major challenges of modern societies. 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. 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 multiclass classification modeling setting that is employed herein, the Random Forest (an ensemble-type machine learning algorithm) is found to be the most accurate prediction/classification model, followed by Support Vector Machines and Artificial Neural Networks. From the practical standpoint, the enhanced predictive power of machine learning algorithms coupled with a multi-source database revealed the most relevant corruption-related information, contributing to the related body of knowledge, generating actionable insights for administrator, scholars, citizens, and politicians. The variable importance results indicated that government integrity, property rights, judicial effectiveness, and education index are the most influential factors in defining the corruption level of significance.

1. Introduction

It is not humanly possible to foresee what exactly the next few years will bring when it comes to informed, data-driven decision-making, especially when one realizes that the majority of the extant digitized data in the world was generated over the last couple of years. It is nearly inconceivable for most of us to wrap our minds around the huge potential of knowledge discoveries available from this data using modern-day machine learning and data mining techniques. Not only computational sciences and applied engineering benefits from these new computational endeavors; recently, social sciences have been the primary subject of such studies. As it turns out, these discovery focused endeavors are even more fascinating and valuable when they are applied to the soft/behavioral/social science problems.

Human behavior-related societal issues such as corrupt behavior have been explored by scholars for more than 100 years. Pearson (1901) used the Gauss least squares method to introduce principal component analysis at the beginning of the twentieth century; and the recent explosion of "Big Data" has substantially advanced social science studies (Wolfe, 2013). Responding to the challenge of making sense of the enormous amount of data generated by humanity during the last decade, powerful machine learning techniques were developed and used to bring light to important knowledge fronts in a number of fields. As corruption is enduring and permeates human societies over time and space, its pervasiveness and significance around the world make it a worthy challenge for modern and powerful machine learning techniques.

In 2016, the World Bank estimated that the cost of corruption was over $2.6 trillion, accounting for 5% of the global domestic product (GDP). Transparency International, which claims to give voice to the victims of corruption, points out that two out of five business executives pay bribes at some point when dealing with public institutions in developing countries (Ribeiro, Alves, Martins, Lenzi, & Perc, 2018). Many researchers agree that there exists an enormous deleterious effect of corruption on economies. For example, the International Monetary Fund (IMF) points out that bribery alone accounts for U.S.$1.5 to $2

* Corresponding author at: Department of Business Analytics, Journal of Business Analytics, Spears and Patterson Endowed Chairs in Business Analytics, Director of Research—Center for Health Systems Innovation, Department of Management Science and Information Systems, Spears School of Business, Oklahoma State University, 700 North Greenwood Avenue, T-NCB 302, Tulsa, OK 74106, USA.
E-mail addresses: marcio@metalsider.com.br (M.S.M. Lima), dursun.delen@okstate.edu (D. Delen).

https://doi.org/10.1016/j.giq.2019.101407
Received 30 May 2019; Received in revised form 9 September 2019; Accepted 10 September 2019
Available online 11 October 2019
0740-624X/

trillion, which constitutes just one part of the many possible forms of corrupt practices. The definition of corruption goes beyond transgressing government rules when conducting public affairs, according to Nuijten and Anders (2017). Although related institutions such as the World Bank and Transparency International define corruption as the practice of abuse of public offices for private gain, using moral arguments to base the concept of corruption seems to make sense in a time of bribery, theft, and misappropriation of public and private assets (Li, Xu, & Zou, 2000). It is evident that corruption worries policy makers and international organizations. However, to some extent these societal entities seem to move slowly when addressing the factors that lead to corrupt practices. The aim of the present research is to contribute to the corruption research by employing powerful machine learning algorithms to reveal the important predictors for corruption perception across 132 nations.

Subtraction of public assets can fundamentally affect crucial resources for the population. Therefore, the drainage of public funds through corruption tends to increase revolutionary feelings among economic actors such that tax collections are perceived as an official type of robbers (Power & Taylor, 2011). This, in turn, may lead to recursive tax evasion or even more variants of corruption, all of which can sabotage a nations' economic wellbeing. Spain may be said to be a legitimate example of this vicious cycle. In 2008, the unemployment rate in Spain started to collapse. It raised from a one-digit figure to 27% in the short time frame of four years. An alarming public deficit of 11% and a drop of around €168 billion in Spanish households' income led Spanish treasury bonds to skyrocket by 7% in 2012. As efforts to reduce government expenditures and stabilize public deficits were enforced by Spanish society, several political corruption cases were revealed. The premium on the Spanish treasury bond rate in the summer of 2012 was credited to high corruption levels (López-Iturriaga & Sanz, 2017).

As corruption distorts public trust, impairs economic growth, and interferes with effective political decision making (Grimes, 2013), it also jeopardizes private investment intentions in diverse fields. The negative social consequences of corruption have been widely reported across the globe. Recently, Brazil has been revealed to have among the worst corruption scandals of modern history. The federal operation “Java Jato” revealed around $2 billion in bribes distributed to government administrators and several private companies. According to the Brazilian Prosecution Office, this recent Brazilian corruption scandal condemned 132 individuals to prison as of August 2018.[1] Although the deleterious consequences of corruption have been addressed in a number of studies, it remains a frequent subject for scholars. Similarly, even though corruption causes have been substantially explored with the use of regression-based statistical analysis, the results remain unclear or at least insufficient to emphasize conclusions.

According to Hillard, Purpura, and Wilkerson (2008), improvements in information technology tend to generate larger databases, which in turn require more efficient methods for solving classification problems. An interesting work by Clark, Morris, and Lomax (2018), applied machine learning algorithms to estimate “Leave Vote” percentage on the United Kingdom's 2016 referendum (i.e., Brexit) that addressed political intentions to exit the European Union. Their results suggested that machine learning methods are capable of predicting the outcome using the e-petition data from the Westminter Parliament. To our knowledge and to this date, no academic study has explored factors that are likely to predict corruption perceptions across countries through the lens of modern data mining techniques. The major goal of this analytic study is to reveal probable factors and their relative importance as predictors of corruption. The dataset used for this purpose is drawn from 132 countries across the globe. In order to obtain reliable results, we employ modern and the most prevailing machine learning algorithms. Although recent research has applied artificial neural networks to predict corruption in Spain, our approach relies on methodological innovations. That is, we use multiple prediction models and a heuristic method to assess variable importance that is based on the ratio of the number of splits by candidates to splits from Random Forest attribute statistical output. Another important contribution from this investigative work lays in employing machine learning techniques to reveal potential corruption predictors at the country level as opposed to the regional level.

Use of artificial intelligence and machine learning techniques in detecting and understanding governmental fraud and corruption has been gaining popularity in literature in the recent years (Stockemer, 2018; Sun & Medaglia, 2019; Tang, Chen, Zhou, Warkentin, & Gillenson, 2019). According to de Sousa, de Melo, Bermejo, Farias, and Gomes (2019), there is a growing interest for studies involving artificial intelligence in the public sector. As the economic affairs became one of the most explored government subjects, AI has contributed to untangle risk and fraud related issues by working with massive data sets generated inside government institutions. In line with this increasing trend, and to complement the extant literature, the present study tackles corruption with modern machine learning techniques, potentially bolstering government-related strategies towards a cleaner society.

This rest of the paper is organized as follows. Section 2 provides a literature review that conceptualizes corruption, prevailing corruption measures, and a brief overview of the traditional and algorithmic data modeling approaches. Section 3 summarizes the research methodology that includes a succinct description of the data, data preprocessing, machine learning models, and methods employed in the study. Section 4 presents the modeling results, explains the predictive accuracy and predictor order of importance of all model types. The last section, Section 5, summarizes the study, its findings, and provides related insights and implication of the findings.

2. Literature review

2.1. Theory of corruption

As Thompson (2005, p. 4) nicely stated it, “the pollution of the public by the private” underlines the long-established concept of corruption. According to Thompson (2005), corruption can be tackled with two different lines. Individual corruption follows a quid pro quo fashion, in which institutions or public officials collect benefits that do not serve public interests, providing advantages outside professional relationships. Institutional corruption happens when public officials or institutions collect some type of advantage that is convenient to the institutional objective, and in turn consistently support the giver in ways that hamper the procedural structure of institutions (Thompson, 2005). In this sense, Philp and Dávid-Barrett (2015) point out that Thompsons' concept of institutional corruption enlarges the study of corruption by contrasting motives, intentions, and the violation of appearance standards as potential causes of wrongdoing. From a different perspective, 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.

In line with our purpose in this study, it is worth mentioning that Lipset's (1960) elaboration on this topic suggested that societies with higher levels of education are likely to have less corrupt behavior. Regardless of the type of corruption, Thompson (2018, p. 5) suggests that the differentiation between individual and institutional corruption should “encourage students of corruption to conduct normative and empirical analysis of the patterns of access to institutions, conflicts of interest, the organization of staff, and the relations between legislators and bureaucrats.” Lessig (2011) points out that some scholars attempt to use social science theories to explain corruption. He builds on the

[1] http://www.mpf.mp.br/para-o-cidadao/caso-lava-jato/atuacao-na-1a-instancia/parana/resultado

idea that many corruption theorists are more interested in philosophical elaborations to the neglect of empirical research. Following these calls and in a scholarly attempt to shed new light on the field, we contribute to knowledge by using modern data mining techniques to establish an independent view of the potential corruption perception causes across several countries.

2.2. Corruption predictors

Corruption is constant and ubiquitous across societies. It surpasses the notion of violating the limits of government rules when public and private entities take actions that disregard basic ethical principles (Nuijten & Anders, 2017). The existing literature on corruption points out that decentralization of public decisions benefits individuals while neglecting the general public interests. According to Prud'homme (1994) and Tanzi (1995), who worked on exploring the role of decentralization on macroeconomics, decentralization may actually disorganize audits and ethics control by regulatory federal entities so that corruption may take place. A number of scholars from this body of research find that political decentralization is positively related to corruption because it fosters abundant and perhaps exaggerated interactions between officials and private investors (Blanshard & Schleifer, 2000; Shah, 1998; Shleifer & Vishny, 1993). Although scholars appear to agree that decentralization is a strong predictor of corruption, it is still unclear to what extent other macroeconomic indicators may influence corruption perception levels, including private property rights, judicial effectiveness, difficulties in starting new businesses, obtaining construction permits, and financial freedom. For example, Graeff and Mehlkop (2003) imply that there is a strong effect of economic freedom on corruption perception indexes and suggest that capital restrictions are likely to be one of its major causes. Although these researchers use a regression-based study, they provide evidence that regulations can impact Corruption Perception Indexes (CPI) by influencing the transaction costs associated with corruption practices. They find that political instability weakens institutions responsible for ensuring adequate compliance with good economic practices, which in turn, increases corruption (Damania, Fredriksson, & Mani, 2004). Again, using a regression-based methodology, Apergis and Cooray (2017) find a statistically significant effect of legal systems and property rights on corruption. Their conclusion emphasizes the need for further investigation of the issue with perhaps more solid insights. In this sense, the current state-of-the-art machine learning algorithms emerge as novel and possibly more robust tools for revealing the actual predictors of CPI.

The potential causes of corruption have been attempted to be addressed by a number of regression-based studies from a wide variety of configurations and perspectives. For example, authors like Persson, Tabellini, and Trebbi (2003) and Chang and Golden (2007) suggest that corruption may be correlated with electoral system features such as size and type of political system. Addressing corruption at the regional level in China, Dong and Torgler (2013) find that more access to media is likely to reduce corruption. Also, at the state level, Lipset (1960) studies the potential influences of educational levels on corruption. Glaeser and Saks (2006) find a consistent relationship between more educated American states and less corrupt states.

One common issue of regression-based studies addressing potential corruption predictors is the direction of causal relationships between variables. For example, Depken and La Fountain (2006) show that in a sort of vicious cycle, taxpayers from more corrupt nations with lower bond ratings are likely to bear increased interest rates when borrowing money to grow their businesses. How evident is it that corruption leads to lower bond rates and therefore increased interest rates, but not the other way around? This apparent ambiguous causal relationship between increased interest rates and corruption is not rare in regression-based studies. Following the same rationale, flawed conclusions may occur if researchers are not careful to explain the direction of causal relationships between corruption and decentralization (Fisman & Gatti, 2002), government size (Goel & Nelson, 1998), business-venturing (Mitchell & Campbell, 2009), or income inequality (Apergis, Dincer, & Payne, 2010). On the other hand, mainly because of the broad use of the internet, the enormous amount of data generated, fosters the development of powerful data mining techniques such as machine learning algorithms. In this sense, employing machine learning models to help untangle dubious causal relationships between corruption and corruption predictors seems to be a shrewd strategy for the field. Due to its ultimate focus on predictive accuracy results that are free from traditional stochastic model limitations (e.g., multicollinearity, normality). machine learning algorithms are consistently used, and have shown to be superior predictive tools when compared to traditional regression-based models (Delen, Cogdell, & Kasap, 2012; Seifert, 2004; Sharda, Delen, & Turban, 2017). To the best of our knowledge, this is the first scholarly study to address potential corruption predictors across countries with the use of powerful machine learning techniques.

2.3. Corruption measures

One of the most widely accepted definitions to date claims that corruption establishes practices of illegal payments to public agents to obtain benefits that may not be deserved, or the abuse of public offices for private gains (Klitgaard, 1988; Rose-Ackerman, 1996; Shleifer & Vishny, 1993). By its very nature, corruption leaves scant identifiable traces, which creates a serious puzzle for scholars. The principal approach to measuring corruption has been based on perception indexes built upon expert or public opinion (Mauro, 1995). Measuring corruption by assessing how societies perceive it became recognized as a valid method when employing the Transparency International Corruption Perception Index (Lambsdorff, 2003) and the governance indicators of the World Bank (Kaufmann, Kraay, & Mastruzzi, 2009).

Recent developments in testing have improved the validity of subjective measures of corruption based on perception indexes. Mocan (2008) and Razafindrakoto and Roubaud (2010) suggest that expert opinions, such as those used by Transparency International, to assess corruption levels in countries may fail to approximate potential results to the ones obtained by using objective measures. Along these lines, Olken (2007), Donchev and Ujhelyi (2008), and Razafindrakoto and Roubaud (2010) question corruption perception as a valid measure by elaborating on potential threats to validity associated with it. However, a number of published articles apply CPI as a reliable measure. Examples include Damania et al. (2004), who use CPI to explore the influence of legal regulation compliance on the resilience of corruption practices, and Graeff and Mehlkop (2003), who find that the direction and the magnitude of the effect of certain economic freedom aspects on corruption depend on whether the country is rich or poor.

Although objective measures of corruption can be inferred through institutional data such as procurement practices or budget procedures that may create opportunities for corruption, this type of data is not widely available. There have been efforts towards revealing alternative measures of corruption (Seligson, 2006). The ratio of the number of public official convictions to the total population is an example. Glaeser and Saks (2006) employ the aforementioned corruption measure to posit that higher quality education and superior per capita income are positively associated with lower levels of corruption. However, considering that the concept of corruption goes beyond misbehavior of public officials (Nuijten & Anders, 2017); using the ratio of convictions of public officials to the total population as a measure of corruption may not be the most appropriate measure to reveal country-level corruption predictors. Recent research attempting to predict corruption in Spain with artificial neural networks found taxation of real state and economic growth among the most relevant corruption predictors (López-Iturriaga & Sanz, 2017). Alas, in addition to using only one type of algorithm to explore the data, the corruption measure that was

adopted may not be generalized to the country level. These results lack generalizability given that depending on a country's judicial effectiveness, the number of convictions may not reflect the actual number of corruption cases, which would in turn, mask the corruption measurement analysis.

According to Kaufmann, Kraay, and Mastruzzi (2006), the World Bank enumerates three different manners for measuring corruption: collecting corruption perceptions of important stakeholders within the society, a detailed examination of specific public projects, and a critical assessment of a country's features that are potentially related to corruption. The CPI is the most largely applied corruption measure and fits into the first measurement method. Given that the CPI measure aggregates multiple perception scores into one measure, it presents the advantage of providing approximate calculations of the targeted output variable with partial data. According to Transparency International, the CPI measure is composed of scores on surveys from 13 different sources. This information compiling process helps the claim that this measure is the most reliable and therefore, more appropriate to capture a country's corruption level. CPI surveys focus on assessing bribery, fraud involving public resources, and manipulation of public power for the benefit of individuals. As the CPI comprises a broader concept of corruption, the other objective measures lack generalizability by focusing on a particular empirical analysis and targeting a specific type of corruption (Kaufmann et al., 2006).

Finally, although it is possible that people perceive greater corruption than they experience, it is difficult to establish whether the corruption experienced is the true representation of the underlying reality (Hussmann, 2011, p. 36). Moreover, studies that use CPI-type measures seem to be more easily replicable by scholars over time and across countries, which in turn produces more meaningful and believable insights.

2.4. The predictive power of modern data mining techniques

The machine learning community has found convincing evidence of the superior predictive power of modern-day algorithms over traditional regression-based statistical models. For instance, the recent study by Delen, Tomak, Topuz, and Eryarsoy (2017) provide invaluable insights to help mitigate injury severity risks incurred in car crashes. Artificial Neural Networks (ANN), Support Vector Machine, and Decision Tree models achieved, respectively, 81.31%, 90.41%, and 86.61% predictive accuracy while the traditional logistic regression model scored 76.97% accuracy. When tackling the survivability of breast cancer patients, logistic regression was also found to yield inferior predictive accuracy when compared to Artificial Neural Networks and Decision Trees (Delen, Walker, & Kadam, 2005). The relatively recent development of efficient open-source data modeling tools, as well as the impressive amount of data that is being generated by modern society, bolsters the development of machine learning research (Sharda et al., 2017). On the other hand, the majority of studies exploring factors that may influence corruption aim at evaluating significant and non-significant effects between variables. However, in social science as well in a number of other disciplines, data interpretation may be highly susceptible to flawed conclusions when relying solely on techniques based on null-hypothesis significant testing (NHST). In this sense, Loftus (1996) argues that a chaotic phenomenon occurs in a number of social studies building on the observation that similar results (e.g., $p = 0.049$, $p = 0.050$, and $p = 0.051$) can lead to entirely opposing scholarly inferences. Giving that the knowledge-building process must rely on accurate findings, it is crucial to improve the prediction accuracy of sociological arenas such as political sciences. As the extant literature shows, social scientists have been addressing corruption for a long time with the use of traditional stochastic data modeling. An exception is the work of López-Iturriaga and Sanz (2017), who attempt to use neural networks to reveal potential predictors related to public corruption in Spanish provinces. However, in addition to concerns about the use of only one type of algorithm, that particular study addresses potential corruption predictors at the regional level within a specific country, Spain. Our study tackles corruption at the country level, employing reliable data from well-known sources such as the World Bank, Transparency International, and the Heritage Foundation.

In light of the extant literature, and to the best of our experiential knowledge, this study presents the first scholarly effort to assess macroeconomic and global indexes as potential predictors for corruption at the country level using machine learning algorithms. By exploring a multimodal machine learning style, the present work determines the most important CPI predictors at the country level, adding to the body of knowledge to be used and further explored by political scientists, machine learning scholars as well as to government administrators.

3. Method

To build the underlying structure of the proposed methodology, we first briefly describe the sources of the data sets and the techniques that we have used to obtain, integrate, and process the dataset for further analyses. We, then, succinctly explain the specific models we developed and the measures of predictive accuracy we employed. We also describe the details of the heuristic method (called *actual splitting ratio*) we used to assess the relative importance of the input variables. To effectively and efficiently conduct our analysis, we employed a combination of data mining best practices. To execute the proposed methodology, we used a number of software tools, including Microsoft Excel, Tableau, SAS JMP, and KNIME.

3.1. Data acquisition and data preprocessing

Our initial dataset was acquired from multiple sources, including Ease of Doing Business Indexes,[2] the Heritage Foundation,[3] Transparency International,[4] and the Human Development Reports of the United Nations Development program[5] for the years 2017 and 2018. The Doing-Business Project, which contains measures of business regulations and how effectively these regulations are actually executed, provides data from 190 countries. It has been used by a number of academic studies in several research fields such as politics, economics, and law (Dixit, 2009; Roe & Siegel, 2009). The information provided by the Ease of Doing Business Project are mainly related to starting a business, paying taxes, trading across borders, getting credit, and registering property. The Ease of Doing Business ranking system assesses economies by comparing the distance to frontier score to benchmark countries regarding 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 2018 Doing Business questionnaire, survey answers are 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 is converted for a string-type variable before computing 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

[2] www.doingbusiness. org
[3] www.heritage.org
[4] www.transparency.org
[5] hdr.undp.org

details on Ease of Doing Business methodology, please visit their official website.[6]

The Heritage Foundation has delivered effective research metrics on key policy issues for over 20 years. The index provides information on government economic practices across 186 countries and relies on data from internationally recognized data sources (Miller, Kim, & Holmes, 2015). Examples of economic matters addressed by the Heritage Foundation are Rule of Law, Government Size, Regulatory Efficiency, and Market Openness. Our research uses the 12 components of the overall Economic Freedom index as potential predictors for increased levels of corruption perception across countries: Property Rights, Judicial Effectiveness, Government Integrity, Tax Burden, Government Spending, Fiscal Health, Business Freedom, Labor Freedom, Monetary Freedom, Trade Freedom, Investment Freedom, and Financial Freedom. Each of the 12 components' scores are calculated from subfactor numerical data that were previously preprocessed (e.g., averaged, normalized) for comparative purposes. More details about the methodology applied to develop the Economic Freedom index measures can be found on the organization's website.[7]

Transparency International is an entity that defines itself as being a global movement giving voice to the victims of worldwide corruption. The Corruption Perception Index (CPI) sorts 180 countries according to their perceived corruption levels. The index captures the assessments of domain experts on corrupt behavioral 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. Examples of related topics embraced by the CPI include bribery, diversion of public funds, use of public office for private gain, and nepotism. For details about the methodology to compute the CPI, the reader is referred to description file on the organization's website.[8]

The Human Development Report, a part of the United Nations Development Program, is intended to support several nation-related analyses by gathering and exploring data from several countries across the globe. One dimension, the Education Index, represents an average of mean years of schooling of adults and expected years of schooling of children. The index uses a scale related to the corresponding maxima.[9]

Delen, Sharda, and Bessonov (2006) argue that even when the imputation technique is highly sophisticated, imputed values may lead to biased results because they are not real. On the other end, Schafer and Olsen (1998) point out that in some cases, statistical analysis can be hampered by missing data. After processing and joining the data obtained from the three different above-mentioned sources, we had a final dataset containing 30,420 data points with only 24 missing values. Variables that contained missing values were Paying Taxes - Time to comply with corporate income tax audit (hours), Paying Taxes - Time to complete a corporate income tax audit (weeks), Paying Taxes - DTF-Postfiling index (0 − 100), overall score, tax burden, trade freedom, investment freedom, financial freedom and education index. Given the fairly small amount of missing data, representing $< 0.07\%$ of the entire dataset, we decided to apply multivariate normal imputation (MVNI) to replace the missing values. Given that MVNI has been suggested as one of the most effective imputation methods for continuous, binary, and ordinary variable types that do not necessarily follow a normal distribution (Lee & Carlin, 2010), we employed MVNI to generate the 24 missing data points. Specifically, we applied the MVNI shrinkage approach to the dataset using its implementation in SAS JMP 13 (the detailed explanation of this method can be found in Schäfer and Strimmer (2005)). Table 1 lists the names and descriptive statistics (including the Shapiro-Wilk p-value for normality test) of all variables included in this analytics study.

The final database comprised of information on 117 variables across 132 countries from five distinct world regions (Americas, Asia Pacific, Europe and Central Asia, Middle East and North Africa and Sub-Saharan Africa). It is worth noting that for the year 2017 data for the following four countries: Equatorial Guinea, Seychelles, Swaziland and Vanuatu, were not available and therefore were not used in the present study. As shown in Table 1, the majority of the variables are not normally distributed at the 5% significance level. Previous research on corruption has explored nonparametric methods such as the Mann–Whitney U test to compare mean differences between predictors that are not normally distributed (López-Iturriaga & Sanz, 2017). However, as evidence consistently shows, machine learning algorithms are capable of extracting knowledge that is hidden deep in large datasets involving several types of input variables that are not necessarily normally distributed (Delen et al., 2012; Sharda et al., 2017). Free from inherent constrains of traditional stochastic data models, the machine learning community assumes that in nature, data is generated in complex ways that are not necessarily normally distributed nor linearly correlated. According to Breiman (2001), the one assumption for the algorithm culture is that data generated by natural processes follow an unknown multivariate distribution. In this sense and in a classification model setting, our research contributes to the corruption literature by employing the most prevailing machine learning techniques to unearth the most important corruption perception predictors across economies. Fig. 1 uses average CPI scores from the years 2017 and 2018 to create a color scale (from red to blue) to illustrate and differentiate high and low levels of corruption across countries.

3.2. Model building

Machine learning is a subset of Artificial Intelligence, which is often applied when computational devices try to mimic the human cognitive functions related to the processes of learning and problem solving to achieving “optimal” results. Within the machine learning community, the development of methods that enable and improve the process of learning by computers is often tested by model comparison methods. In other words, the goal of the process is to identify the machine learning model and its optimal set of parameters that achieve the highest unbiased predictive accuracy for a given problem and related dataset. In parallel to traditional stochastic data modeling, the algorithmic culture has been exploring and improving the predictive accuracy of machine learning models for decades (Huang, Shi, & Suykens, 2014). Our proposed methodology explores the predictability of the phenomenon by applying and comparing three popular machine learning techniques—Random Forest, Artificial Neural Networks (ANN), and Support Vector Machine (SVM)—to a feature-rich dataset.

Random Forest is a bagging-type of ensemble model composed of a collection of small trees. It has been known for its superior performance and robust outcomes. Random Forest works well with large datasets and handles imperfect data effectively while maintaining a high level of accuracy. Compared to ANN and SVM, Random Forest provides a better understanding of the mechanics behind the predictive model (Duro, Franklin, & Dubé, 2012). Although the generalization error that occurs in a Random Forest depends on the predictive strength of the decision trees in the models, ensemble models are suggested to provide robust models with a high predictive accuracy when compared to individual classifiers (Svetnik et al., 2003). According to Breiman (2001), at each node within the forest, input variables are selected at random such that the nodes are split according to internal impurity criteria while growing trees of the forest. Each tree in the forest provides a classification outcome that is recorded as a vote. Votes from all trees are pooled, and the class with the maximum votes is finally computed. Successful applications of the Random Forest algorithm are reported in several fields such as e-commerce, finance, sports, and medicine (Sharda et al., 2017).

[6] www.doingbusiness.org/methodology
[7] https://www.heritage.org/index/pdf/2018/book/methodology.pdf
[8] https://files.transparency.org/content/download/2183/13748/file/CPI_2017_TechnicalMethodologyNote_EN.pdf
[9] hdr.undp.org/en/indicators/103706

Table 1
List of variables and their descriptive statistics.

	Min	Max	Mean	Std Dev	Median	25% Quart	75% Quart.	W	Prob < W
Starting a Business - DTF - Starting a Business	33.11	99.96	83.43	11.93	86.03	79.99	91.38	0.84	< 0.0001
Starting a Business - Procedures - Men (#)	1.00	16.00	6.81	2.95	7.00	5.00	8.00	0.95	< 0.0001
Starting a Business - Time - Men (Days)	0.50	99.00	18.61	19.37	11.50	6.00	24.00	0.73	< 0.0001
Starting a Business - Cost - Men (% of Income per Capita)	0.00	219.30	21.30	34.27	8.05	1.80	26.25	0.62	< 0.0001
Starting a Business - Procedures - Women Number)	1.00	16.00	6.94	3.00	7.00	5.00	9.00	0.95	< 0.0001
Starting a Business - Time - Women (Days)	0.50	99.00	18.74	19.38	11.50	6.50	24.00	0.73	< 0.0001
Starting a Business - Cost - Women (% of Income per Capita)	0.00	219.30	21.30	34.27	8.05	1.80	26.25	0.62	< 0.0001
Starting a Business - Minimum capital (% of Income per Capita)	0.00	556.60	12.31	52.34	0.00	0.00	6.68	0.23	< 0.0001
Dealing with Construction Permits - DTF - Dealing with Construction Permits (DB16-18 Methodology)	22.54	87.04	65.26	12.33	67.27	58.18	73.41	0.95	< 0.0001
Dealing with Construction Permits - Procedures (#)	7.00	32.50	14.94	4.37	14.70	12.00	17.00	0.95	< 0.0001
Dealing with Construction Permits - Time (Days)	50.50	652.00	167.88	88.44	152.50	110.50	206.75	0.82	< 0.0001
Dealing with Construction Permits – Cost (% of Warehouse Value)	0.10	89.80	6.30	10.70	2.50	1.10	6.88	0.54	< 0.0001
Dealing with Construction Permits - Building Quality Control Index (0-15) (DB16-18 Methodology)	1.00	15.00	9.94	2.99	11.00	8.00	12.00	0.94	< 0.0001
Dealing with Construction Permits - Quality of Building Regulations Index (0-2) (DB16-18 Methodology)	0.00	2.00	1.61	0.59	2.00	1.00	2.00	0.68	< 0.0001
Dealing with Construction Permits - Quality Control Before Construction Index (0-1) (DB16-18 Methodology)	0.00	1.00	0.86	0.35	1.00	1.00	1.00	0.41	< 0.0001
Dealing with Construction Permits - Quality Control During Construction Index (0-3) (DB16-18 Methodology)	0.00	3.00	1.68	0.85	2.00	1.00	2.00	0.74	< 0.0001
Dealing with Construction Permits - Quality Control After Construction Index (0-3) (DB16-18 Methodology)	0.00	3.00	2.64	0.66	3.00	2.00	3.00	0.57	< 0.0001
Dealing with Construction Permits - Liability and Insurance Regimes Index (0-2) (DB16-18 Methodology)	0.00	2.00	0.82	0.75	1.00	0.00	1.00	0.82	< 0.0001
Dealing with Construction Permits – Professional Certifications Index (0-4) (DB16-18 Methodology)	0.00	4.00	2.33	1.49	2.00	1.00	4.00	0.85	< 0.0001
Getting Electricity - DTF - Getting Electricity (DB16-18 Methodology)	19.91	99.92	67.25	18.55	70.99	54.87	82.31	0.95	< 0.0001
Getting Electricity - Procedures (number)	2.00	9.80	5.16	1.42	5.00	4.00	6.00	0.92	< 0.0001
Getting Electricity - Time (Days)	10.00	482.00	95.59	64.98	82.00	61.00	113.75	0.71	< 0.0001
Getting Electricity - Cost (% of Income per Capita)	0.00	16,917.50	1381.39	2720.47	349.60	69.95	1223.50	0.54	< 0.0001
Getting Electricity - Reliability of Supply and Transparency of Tariff Index (0-8) (DB16-18 Methodology)	0.00	8.00	4.17	3.20	5.00	0.00	7.00	0.80	< 0.0001
Getting Electricity - Total Duration and Frequency of Outages per Customer a Year (0-3) (DB16-18 Methodology)	0.00	3.00	1.32	1.27	1.00	0.00	3.00	0.78	< 0.0001
Getting Electricity - Mechanisms for Monitoring Outages (0-1) (DB16-18 Methodology)	0.00	1.00	0.73	0.44	1.00	0.00	1.00	0.55	< 0.0001
Getting Electricity - Mechanisms for Restoring Service (0-1) (DB16-18 Methodology)	0.00	1.00	0.72	0.45	1.00	0.00	1.00	0.56	< 0.0001
Getting Electricity - Regulatory Monitoring (0-1) (DB16-18 Methodology)	0.00	1.00	0.74	0.44	1.00	0.00	1.00	0.56	< 0.0001
Getting Electricity - Financial Deterrents Aimed at Limiting Outages (0-1) (DB16-18 Methodology)	0.00	1.00	0.48	0.50	0.00	0.00	1.00	0.64	< 0.0001
Getting Electricity - Communication of Tariffs and Tariff Changes (0-1) (DB16-18 Methodology)	0.00	1.00	0.86	0.35	1.00	1.00	1.00	0.41	< 0.0001
Registering Property - DTF - Registering Property (DB17-18 Methodology)	27.50	94.47	63.63	16.01	63.22	51.38	76.44	0.98	0.0025
Registering Property - Procedures (#)	1.00	13.60	5.64	2.25	6.00	4.00	7.00	0.97	< 0.0001
Registering Property - Time (Days)	1.00	312.00	44.61	52.18	34.00	15.20	55.00	0.64	< 0.0001
Registering Property - Cost (% of Property Value)	0.00	19.00	5.43	4.01	4.90	2.50	7.60	0.94	< 0.0001
Registering Property - Quality of Land Administration Index (0-30) (DB16 Methodology)	2.50	29.00	14.72	7.40	14.50	8.00	21.50	0.95	< 0.0001
Registering Property - Quality of Land Administration Index (0-30) (DB17-18 Methodology)	2.50	29.00	14.65	7.46	14.50	7.63	21.50	0.95	< 0.0001
Registering Property - Reliability of Infrastructure Index (0-8) (DB16-18 Methodology)	0.00	8.00	4.17	2.90	4.60	1.00	7.00	0.88	< 0.0001
Registering Property - Transparency of Information Index (0-6) (DB16-18 Methodology)	0.00	6.00	2.74	1.41	3.00	1.50	3.65	0.97	< 0.0001
Registering Property - Geographic Coverage Index (0-8) (DB16-18 Methodology)	0.00	8.00	2.78	3.12	2.00	0.00	4.00	0.77	< 0.0001
Registering Property - Land Dispute Resolution Index (0-8) (DB16-18 Methodology)	1.00	8.00	5.03	1.49	5.00	4.00	6.00	0.98	0.0007
Registering Property - Equal Access to Property Rights Index (-2-0) (DB17-18 Methodology)	−1.00	0.00	−0.07	0.25	0.00	0.00	0.00	0.27	< 0.0001
Getting Credit - DTF - Getting Credit (DB15-18 Methodology)	0.00	100.00	52.06	22.27	50.00	35.00	70.00	0.98	0.0047
Getting Credit - Strength of Legal Rights Index (0-12) (DB15-18 Methodology)	0.00	12.00	5.35	3.00	6.00	2.25	7.00	0.96	< 0.0001
Getting Credit - Depth of Credit Information Index (0–8) (DB15-18 Methodology)	0.00	8.00	5.06	2.94	6.00	3.25	7.00	0.77	< 0.0001
Getting Credit - Credit Registry Coverage (% of Adults)	0.00	100.00	14.79	25.66	0.15	0.00	22.68	0.64	< 0.0001
Getting Credit - Credit Bureau Coverage (% of Adults)	0.00	100.00	31.94	37.11	14.55	0.00	62.70	0.79	< 0.0001
Getting Credit - Getting Credit Total Score (DB15-18 Methodology)	0.00	20.00	10.41	4.45	10.00	7.00	14.00	0.98	0.0047
Protecting Minority Investors - DTF – Protecting Minority Investors (DB15-18 Methodology)	10.00	85.00	55.18	13.66	56.67	46.67	64.92	0.98	0.0022
Protecting Minority Investors - Extent of Disclosure Index (0-10)	0.00	10.00	6.11	2.38	7.00	4.00	8.00	0.95	< 0.0001
Protecting Minority Investors - Extent of Director Liability Index (0-10)	0.00	10.00	4.75	2.56	5.00	2.25	6.75	0.95	< 0.0001
Protecting Minority Investors - Ease of Shareholder Suits Index (0-10) (DB15-18 Methodology)	0.00	9.00	6.23	1.93	6.00	5.00	8.00	0.94	< 0.0001
Protecting Minority Investors - Extent of Shareholder Rights Index (0-10) (DB15-18 Methodology)	0.00	10.00	5.77	1.91	6.00	4.00	7.00	0.96	< 0.0001
Protecting Minority Investors - Extent of Ownership and Control Index (0-10) (DB15-18 Methodology)	0.00	9.00	4.39	2.03	4.00	3.00	6.00	0.96	< 0.0001
Protecting Minority Investors - Extent of Corporate Transparency Index (0-10) (DB15-18 Methodology)	0.00	10.00	5.85	2.24	6.00	4.00	8.00	0.94	< 0.0001

(*continued on next page*)

Table 1 (*continued*)

	Min	Max	Mean	Std Dev	Median	25% Quart	75% Quart.	W	Prob < W
Protecting Minority Investors - Extent of Conflict of Interest Regulation Index (0-10)(DB15-18 Methodology)	1.70	9.30	5.70	1.49	5.70	4.70	6.70	0.98	0.004
Protecting Minority Investors - Extent of Shareholder Governance Index (0-10) (DB15-18 Methodology)	0.30	9.00	5.34	1.59	5.30	4.00	6.70	0.97	0.0001
Protecting Minority Investors - Strength of Minority Investor Protection Index (0-10) (DB15-18 Methodology)	1.00	8.50	5.52	1.37	5.70	4.70	6.50	0.98	0.0025
Paying Taxes - DTF - Paying Taxes (DB17-18 Methodology)	17.92	99.44	70.02	16.61	72.13	59.90	83.81	0.95	< 0.0001
Paying Taxes - Payments (#/ Year)	3.00	60.00	22.27	15.19	18.00	9.00	33.00	0.90	< 0.0001
Paying Taxes - Time (Hours / Year)	12.00	2038.00	237.00	211.36	197.25	136.75	270.00	0.59	< 0.0001
Paying Taxes - Total Tax Rate (% of Profit)	8.00	216.50	40.53	22.41	38.10	28.98	48.50	0.72	< 0.0001
Paying Taxes - Profit Tax (% of Profit)	0.00	53.00	15.48	9.30	16.20	9.30	21.68	0.97	0.0001
Paying Taxes - Profit Tax (% of Profit)_1	0.00	53.00	15.48	9.30	16.20	9.30	21.68	0.97	0.0001
Paying Taxes - Labor Tax and Contributions (% of Profit)	0.00	54.70	17.38	11.11	15.90	10.23	24.80	0.96	< 0.0001
Paying Taxes - Labor Tax and Contributions (% of Profit)_1	0.00	54.70	17.38	11.11	15.90	10.23	24.80	0.96	< 0.0001
Paying Taxes - Other Taxes (% of Profit)	0.00	184.50	7.67	20.08	2.00	0.83	4.30	0.36	< 0.0001
Paying Taxes - Other Taxes (% of Profit)_1	0.00	184.50	7.67	20.08	2.00	0.83	4.30	0.36	< 0.0001
Paying Taxes - Time to Comply with Corporate Income Tax Audit (Hours) (DB17-18 Methodology)	0.00	207.50	15.27	25.57	6.00	3.00	17.50	0.53	< 0.0001
Paying Taxes - Time to Complete a Corporate IncomeTax Audit (Weeks) (DB17-18 Methodology)	0.00	87.10	11.24	18.19	0.00	0.00	18.30	0.68	< 0.0001
Paying Taxes - DTF-Post-Filing Index (0-100) (DB17-18 Methodology)	0.00	100.00	59.72	27.47	59.15	47.53	83.78	0.95	< 0.0001
Trading across Borders - DTF - Trading across Borders (DB16-18 Methodology)	15.99	100.00	71.75	21.32	71.50	59.63	91.45	0.94	< 0.0001
Trading across Borders - Time to Export: Documentary Compliance (Hours) (DB16-18 Methodology)	1.00	504.00	49.34	66.11	26.00	2.55	72.00	0.70	< 0.0001
Trading across Borders - Time to Import: Documentary Compliance (Hours) (DB16-18 Methodology)	1.00	324.00	61.19	68.02	37.50	2.00	110.25	0.84	< 0.0001
Trading across Borders - Time to Export: Border Compliance (Hours) (DB16-18 Methodology)	0.00	312.00	51.84	47.26	42.00	16.00	75.00	0.88	< 0.0001
Trading across Borders - Time to Import: Border Compliance (Hours) (DB16-18 Methodology)	0.00	402.00	71.16	76.52	58.00	4.25	98.00	0.83	< 0.0001
Trading across Borders - Cost to Export: Documentary Compliance (USD) (DB16-18 Methodology)	0.00	1800.00	125.78	176.82	90.00	40.00	160.00	0.50	< 0.0001
Trading across Borders - Cost to Import: Documentary Compliance (USD) (DB16-18 Methodology)	0.00	1025.00	161.58	190.48	100.00	40.00	203.75	0.77	< 0.0001
Trading across Borders - Cost to Export: Border Compliance (USD) (DB16-18 Methodology)	0.00	1633.00	359.96	299.11	307.00	144.75	496.75	0.90	< 0.0001
Trading across Borders - Cost to Import: Border Compliance (USD) (DB16-18 Methodology)	0.00	1585.00	419.17	358.82	367.00	125.00	644.00	0.92	< 0.0001
Enforcing Contracts - DTF - Enforcing Contract (DB17-18 Methodology)	25.94	83.61	56.48	12.89	57.91	48.04	66.95	0.98	0.0006
Enforcing Contracts - Time (Days)	164.00	1785.00	638.33	320.56	535.00	447.93	718.25	0.84	< 0.0001
Enforcing Contracts - Filing and Service (Days)	6.00	165.00	37.65	22.96	30.00	21.00	48.75	0.84	< 0.0001
Enforcing Contracts - Trial and Judgment (Days)	90.00	1420.00	419.17	258.00	365.00	280.00	460.25	0.83	< 0.0001
Enforcing Contracts - Enforcement of Judgment (Days)	30.00	600.00	181.51	114.70	150.00	90.00	220.00	0.88	< 0.0001
Enforcing Contracts - Cost (% of Claim)	9.00	110.30	32.43	18.58	27.10	21.60	36.40	0.80	< 0.0001
Enforcing Contracts - Attorney Fees (% of Claim)	5.00	95.80	20.24	13.68	17.15	12.00	23.65	0.73	< 0.0001
Enforcing Contracts - Court Fees (% of Claim)	0.10	40.20	6.52	5.45	5.20	3.50	8.00	0.72	< 0.0001
Enforcing Contracts - Enforcement Fees (% of Claim)	0.00	38.30	5.67	6.11	3.00	1.10	8.58	0.81	< 0.0001
Enforcing Contracts - Quality of the Judicial Processes Index (0-18) (DB16 Methodology)	2.50	15.50	8.54	2.98	8.00	6.00	10.50	0.96	< 0.0001
Enforcing Contracts - Quality of the Judicial Administration Index (0-18) (DB17-18 Methodology)	1.50	15.50	8.47	3.07	8.00	6.00	10.50	0.97	< 0.0001
Enforcing Contracts - Court Structure and Proceedings (0-5) (DB17-18 Methodology)	0.00	5.00	3.36	1.08	3.00	3.00	4.50	0.92	< 0.0001
Enforcing Contracts - Case Management (0-6) (DB16-18 Methodology)	0.00	5.50	1.86	1.48	1.50	1.00	3.00	0.92	< 0.0001
Enforcing Contracts - Court Automation (0-4) (DB16-18 Methodology)	0.00	4.00	1.04	1.15	0.50	0.00	2.00	0.83	< 0.0001
Enforcing Contracts - Alternative Dispute Resolution (0-3) (DB16-18 Methodology)	0.00	3.00	2.21	0.47	2.50	2.00	2.50	0.86	< 0.0001
Resolving Insolvency - DTF - Resolving Insolvency (DB15-18 Methodology)	0.00	93.89	48.53	22.78	45.38	34.20	67.48	0.97	< 0.0001
Resolving Insolvency - Recovery Rate (Cents on the Dollar)	0.00	93.10	39.55	25.79	35.55	21.30	52.00	0.94	< 0.0001
Resolving Insolvency - Strength of Insolvency Framework Index (0-16) (DB15-18 Methodology)	0.00	15.00	8.72	3.87	9.00	6.00	11.50	0.95	< 0.0001
Resolving Insolvency - Commencement of Proceedings Index (0-3) (DB15-18 Methodology)	0.00	3.00	2.43	0.48	2.50	2.00	3.00	0.79	< 0.0001
Resolving Insolvency - Management of Debtor's Assets Index (0-6) (DB15-18 Methodology)	0.00	6.00	4.19	1.66	4.50	3.00	5.50	0.88	< 0.0001
Resolving Insolvency - Reorganization Proceedings Index (0-3) (DB15-18 Methodology)	0.00	3.00	1.01	1.02	0.50	0.00	2.00	0.83	< 0.0001
Resolving Insolvency - Creditor Participation Index (0–4) (DB15-18 Methodology)	0.00	4.00	1.49	1.00	1.00	1.00	2.00	0.88	< 0.0001
Overall Score	42.00	88.80	62.10	9.78	61.10	54.43	69.10	0.98	0.0003
Property Rights	12.30	98.40	54.73	19.88	53.70	38.20	69.08	0.98	0.0009
Government Integrity	12.20	95.70	44.54	19.76	38.80	30.25	53.60	0.91	< 0.0001
Judicial Effectiveness	10.30	93.80	49.03	20.35	48.50	32.70	62.70	0.97	< 0.0001
Tax Burden	37.20	99.90	76.04	12.78	77.05	68.03	85.20	0.98	0.0006
Government Spending	0.00	96.30	62.04	22.28	66.65	48.05	78.90	0.95	< 0.0001
Fiscal Health	0.00	100.00	67.34	29.76	78.55	51.58	91.78	0.87	< 0.0001
Business Freedom	27.20	95.10	65.35	14.22	66.45	53.95	75.78	0.99	0.0188

(*continued on next page*)

Table 1 (*continued*)

	Min	Max	Mean	Std Dev	Median	25% Quart	75% Quart.	W	Prob < W
Labor Freedom	29.70	92.60	60.50	13.26	59.70	50.43	70.38	0.99	0.1319
Monetary Freedom	47.40	90.10	77.57	7.83	79.10	73.23	83.50	0.93	< 0.0001
Trade Freedom	47.80	90.00	77.00	10.33	79.40	69.73	86.90	0.91	< 0.0001
Investment Freedom	0.00	95.00	59.29	21.75	60.00	45.00	75.00	0.95	< 0.0001
Financial Freedom	0.00	90.00	50.74	18.67	50.00	40.00	60.00	0.97	< 0.0001
Education Index	0.21	0.94	0.66	0.18	0.69	0.50	0.81	0.96	< 0.0001

The Prob < W value listed on the last column is the p–value. The Shapiro-Wilk p–value tests the null hypothesis that the data are normally distributed.

ANN is a proficient computing system also called Artificial Connectionist Systems or Parallel Distributed Processing Systems. ANN derives from a series of machine learning techniques that mimic the human biological neural network functioning process. Often applied to solve complex nonlinear relationships between input and output variables, ANN essentially uses a series of neurons to learn how certain patterns of pre-existing factors tend to produce outcomes. The nodes or neurons transmit signals through connection links according to the learned patterns of information. Similar to the human brain, the connections have different weights that inhibit or excite signals across the neural network. That is, every neuron has an internal state generated by an active signal so that output signals are produced as a combination of activation rules and input signals (Cortes, Gonzalvo, Kuznetsov, Mohri, & Yang, 2016). We apply a common neural network mechanism known as multilayer perceptron (MLP) with a back-propagation type that is one of the most popular ANN architectures (Wu, Jennings, Terpenny, Gao, & Kumara, 2017).

SVM was first introduced by Boser, Guyon, & Vapnik, 1992. As another classification and regression prediction system that maximizes prediction accuracy, it is commonly used in handwriting recognition functions and facial analysis (Varma, Rao, Raju, & Varma, 2016). SVM is based on linear combinations of factors and revolves around the analyses of a theoretical hyperplane learned from training data by using optimization procedures that maximize prediction. During the learning process of a theoretical hyperplane generated from the data, all the training instances, which are called support vectors, fall within a certain distance margin and influence the format and position of the hyperplane (Zhang et al., 2015). In SVM, nonlinear kernel functions are applied to convert complex nonlinear relationships into high dimensional feature spaces. Parallel hyperplanes are generated to maximize the separation of targeted output classifications in the training data. Essentially, the larger the distance between the two parallel hyperplanes, the better the predictive accuracy of the model (Delen, Oztekin, & Kong, 2010). One advantage of SVM over ANN is that it is less prone to overfitting because it is based on structural risk minimization (separation of hyperplanes) while ANN uses empirical risk minimization. SVM has been successfully used in diverse research problems such as exploring electroencephalographic efficiency on predicting working memory (Johannesen, Bi, Jiang, Kenney, & Chen, 2016) and prognostic analysis of thoracic transplantations (Delen et al., 2010).

3.3. Testing and evaluating - cross-validation

In this study, we used *k*-fold cross-validation to randomly split the data into mutually exclusive *k* number of subsets for "training" and "testing" sets. In this method, the *k*-1 folds of the data are used to build the model and the remaining fold is used to test the model. Echoing Delen et al. (2012) concerns that one single random split can potentially lead to heterogeneous subsets of data that would, in turn, produce biased results. Hence, herein, we employed ten rounds ($k = 10$) of cross-validations on the whole dataset. On each round of the 10-fold cross-validation, the model is trained in all-but-one folds and tested on the excluded fold, which is the testing subset for that specific round. The average of the result measures from all ten rounds is then compiled for final analyses. We followed Olson and Delen (2008), who point out that using stratified cross-validation tends to reduce bias when compared to regular cross-validation.

$$CV = \frac{1}{k}\sum_{i=1}^{k} A_i \tag{1}$$

The overall accuracy from the cross-validation procedure is calculated using the average of each individual *k* accuracy measure. In Eq. (1), CV represents cross-validation accuracy, A represents the accuracy measure of the *k* folds, while *k* is the number of splitting rounds (in this case, $k = 10$) (Delen et al., 2012). Fig. 2 is a pictorial representation of

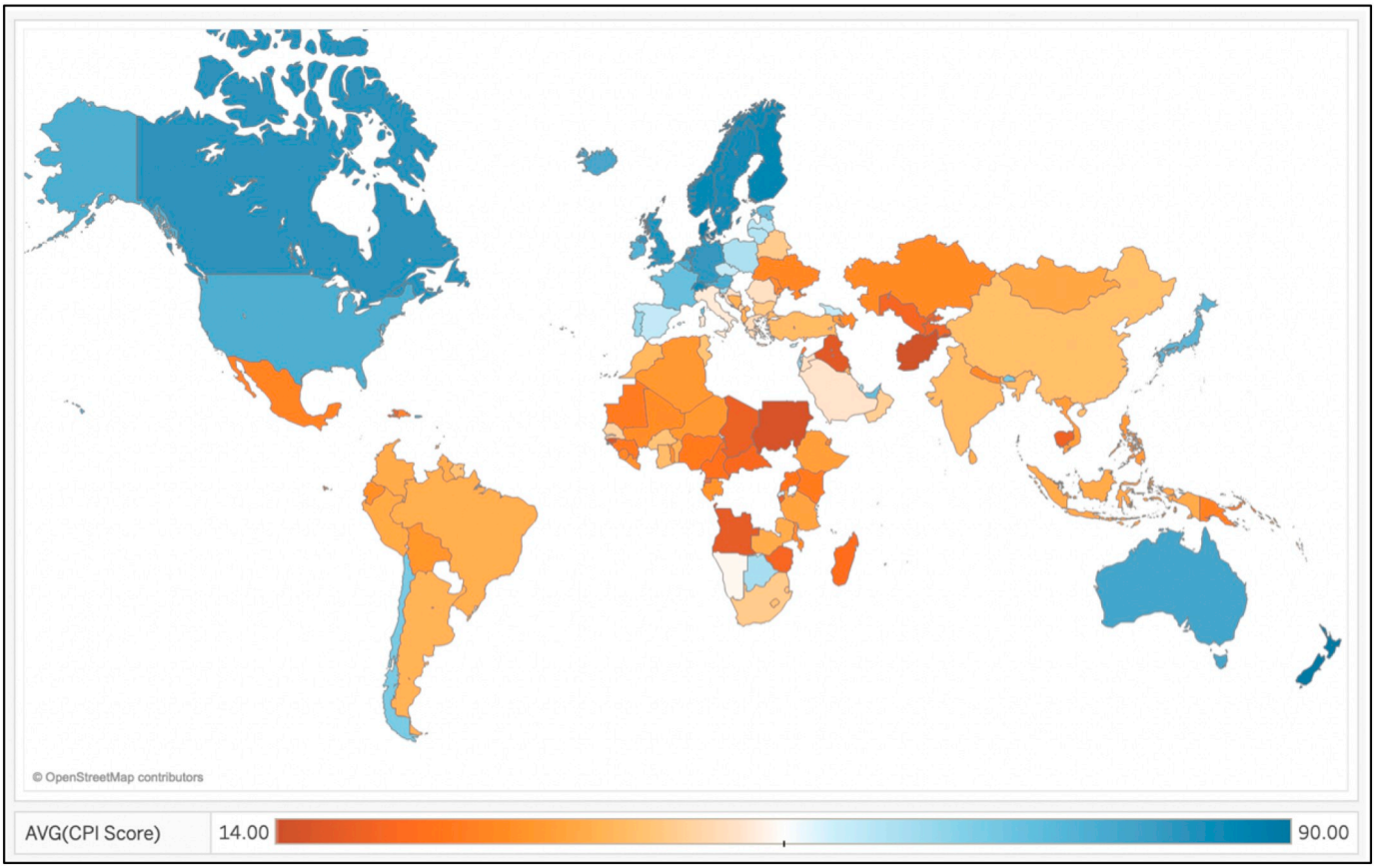

Fig. 1. Corruption Perception Index across countries. Dark blue countries (higher CPI scores) relate to less corrupt economies while dark red countries (low CPI scores) relate to more corrupt economies. (For interpretation of the references to color in this figure legend, the reader is referred to the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

the methodology employed for the present study. It follows an outline proposed and adopted by a number of modern and relevant machine learning studies (Delen et al., 2017; Sharda et al., 2017).

3.4. To compare model performances

To assess and compare prediction performances between the three machine learning algorithms, we employ the concepts of overall accuracy and per-class accuracy. Eqs. 2 and 3 demonstrate the relationships between truly/correctly classified samples (TC) and false/incorrectly classified samples (FC) to compute the measures of model performance. The Accuracy measure assesses the models' capacity to correctly classify all possible CPI classes (i.e., High Corruption, Low Corruption, Very High Corruption, Very Low Corruption). In addition to the overall accuracy, and because there are four possible output classes, we examine the individual (i.e., per-class) accuracy measure as a complementary comparative method. The individual class accuracy measure uses TC and FC to assess the models' capacity to predict each class individually and separately.

$$\textit{Overall Accuracy} = \frac{TC_1 + TC_2 + TC_3 + TC_4}{\textit{All Samples}} \ (\textit{true classification rate of all cases}) \quad (2)$$

$$\textit{Individual Class Accuracy} = \frac{TC_n}{TC_n + FC_n} \ (\textit{true class accuracy of class } n) \quad (3)$$

where TC_n represents the truly/correctly classified samples of class n whereas FC_n reresents the false/incorrectly classified samples of class n.

Table 2 provides a schematic representation of a confusion matrix related to the four-class classification approach employed by this research. The labels displayed on diagonal refer to correctly classified CPI classes where models' predicted CPI classes matched actual CPI classes.

We employ the most prevailing methodological techniques that have been empowering the machine learning community to better assess differences in models' ability to predict relevant outcomes (Lord & Mannering, 2010; Sharda et al., 2017) across a number of fields.

3.5. Variable importance

To assess the relevant order of predictors, we assess the “actual splitting rate” (ASR) from the Random Forest algorithm. The Random Forest is an extension of the decision tree classifiers, where the algorithm randomly selects a subset of input variables and data for each simple tree while building the forest with no pruning/stopping rule (Breiman, 2001). The ASR method builds on the notion of reduced entropy (lack of predictability) at each node—each time a certain variable is used as a splitting point. Basically, the decrease in the Gini impurity criterion is computed and then collectively averaged for a particular variable (see Fig. 3 for a pictorial illustration of this process). Gini impurity generates a split in the forest and determines the splitting hierarchy (Archer & Kimes, 2008). The number of times that each input variable is chosen to be split (candidate to split), and the number of times that particular variable is actually split (split) are computed. This Random Forest-based heuristic method assesses variable importance by calculating the ratio of the number of actual splits on a certain variable to the number of times that particular variable was selected as a candidate to split within the forest. Eq. (4) demonstrates the Gini impurity measurement where j represents possible classes and p_i represent the classification probability of class.

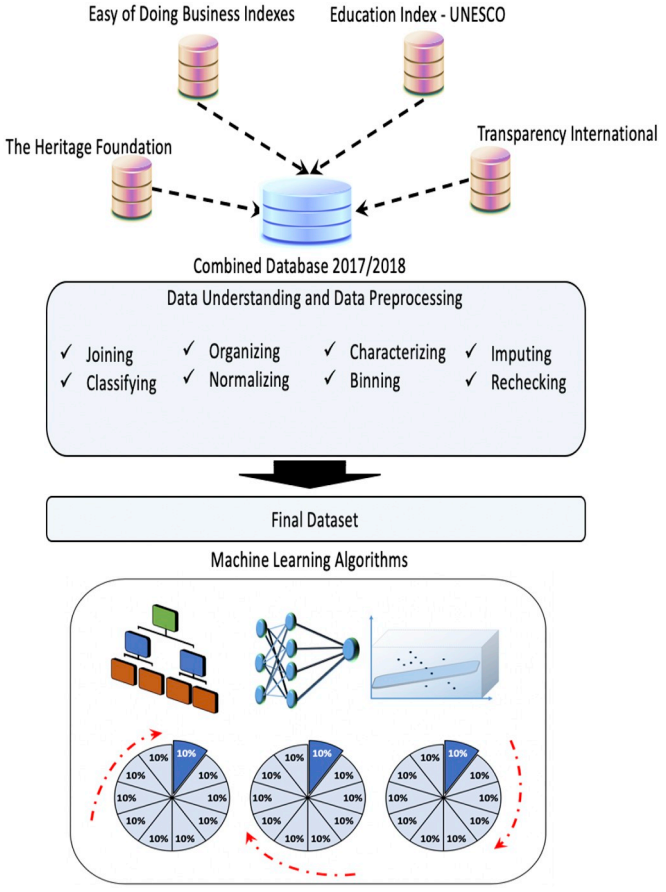

Fig. 2. A Graphical depiction of the methodological approach: the *k*-fold cross-validation methodology used to train and tests 30 models (10 different models for each algorithm type).

Table 2
A confusion matrix for a four-class classification problem.

Predicted/Actual*	Class 1	Class 1	Class 3	Class 4
Class 1	TC_1	FC_2	FC_3	FC_4
Class 2	FC_1	TC_2	FC_3	FC_4
Class 3	FC_1	FC_2	TC_3	FC_4
Class 4	FC_1	FC_2	FC_3	TC_4

* Predicted classes are presented on rows and actual classes are presented on columns.

$$\mathrm{I\,G}\,(p) = 1 - \sum_{i=1}^{J} p_i^2 \tag{4}$$

Determining which predictors exert the largest influence on the outcomes can potentially save time, money, and effort (Dreiseitl & Ohno-Machado, 2002). In this sense, using machine learning techniques to reveal the factors that directly impact perceptions of corruption among citizens across the world represents a unique opportunity for public administrators to prioritize resources and improve government services.

4. Results and discussion

Across 30 prediction models (i.e., 10-fold cross-validations for the three algorithm types), the results on overall accuracy and per-class accuracy are shown in Table 3. The Random Forest algorithm was the most accurate classification technique, achieving 85.77% overall accuracy. All individual class accuracy results also favored the Random Forest model. The Support Vector Machine algorithm was the next best model with 76.15% overall accuracy, followed by Artificial Neural Networks with 73.84%. Although the Random Forest model showed superior predictive accuracy, all model types' accuracy measures are considered decent and relevant.

In addition to exploring the predictive accuracy of the three model types, another major goal of our work relates to revealing the most relevant factors that influence perceptions of corruption across countries. To do so, we conducted the "actual splitting rate" analysis on the Random Forest attribute statistics output as this algorithm was revealed as the most accurate model for our purpose. The number of splits, number of times each variable was chosen as the candidate to split, and the Actual Splitting Rate (ASR) were calculated and graphically presented in Figs. 4 and 5.

The ASR analysis allowed for a reasonable separation of the input variables in three distinct groups. The first group, the most relevant, includes Government Integrity, Property Rights, Judicial Effectiveness, and Education Index. Among the variables within the first group of predictors, Government Integrity appears at the top. To measure Government Integrity scores on topics such as public trust in politicians, irregular payments, and transparency of government policy making are weighted equally, averaged, and computed. Resonating with the work of Gong (2015) and reflecting the common sense that government integrity is intimately related to corruption perceptions by citizens across countries, our results point to Government Integrity as the most important predictor for CPI. Within the Random Forest attribute statistics and through the ASR analysis, the Government Integrity variable was split 90.47% of the times it was chosen as the splitting point candidate (see Fig. 4). Importantly, as opposed to the broad concepts of the first and the fourth relevant input variables, Property Rights and Judicial Effectiveness offer more clear and specific insights.

According to the Heritage Foundation, Property Rights refers to the assessment of how a country's legal system permits individuals or organizations to accumulate private property without restrictions and protection of clear laws. In other words, it measures the extent to which a country's laws secure private property rights (e.g., physical property rights, strength of investor protection, and risk of expropriation). Our results are consistent with the extant literature that suggests the influence of property rights on corruption. For example, Acemoglu and Verdier (1998) infer that the enforcement of property rights by the government, if not accompanied by an effective judicial system, may actually foster corruption. Angulo-Guerrero, Pérez-Moreno, and Abad-Guerrero (2017) suggest the beneficial effect of property rights on economies by exploring its positive influence on entrepreneurship motivation. Our results show through the ASR analysis that the Property Rights variable was split 78.84% of the times it was chosen as a candidate to split, while Judicial Effectiveness was split 77.94% of the time. Judicial Effectiveness measures the extent to which a country's legal system secures citizens' and organizations' rights by ensuring that the law is fully respected with adequate legal procedures. The specific literature found evidence that unstable political environments with ineffective judicial systems may set the base for a rapid spread of bureaucratic corruption (Damania et al., 2004). Consistent with the literature, the ASR analysis pointed out that Judicial Effectiveness should be considered one of the most important predictors of how social entities perceive corruption levels. Arguably, highlighting Property Rights and Judicial Effectiveness among the most important predictors for CPI with seamless predictive accuracy is one of the important contributions of our study. According to the ASR, the fourth most important CPI predictor is the Education Index. This, to some extent, resonates with the work of Glaeser and Saks (2006), who encountered a positive relationship between education and corruption. According to the United Nations, the Education Index reflects the average of years of schooling of adults and probable years of schooling of children. The Education Index was split 53.93% of the times within the Random Forest algorithm, which led it to the fourth place in the order of importance. As the results indicate, the ASR distance between the Education Index and the

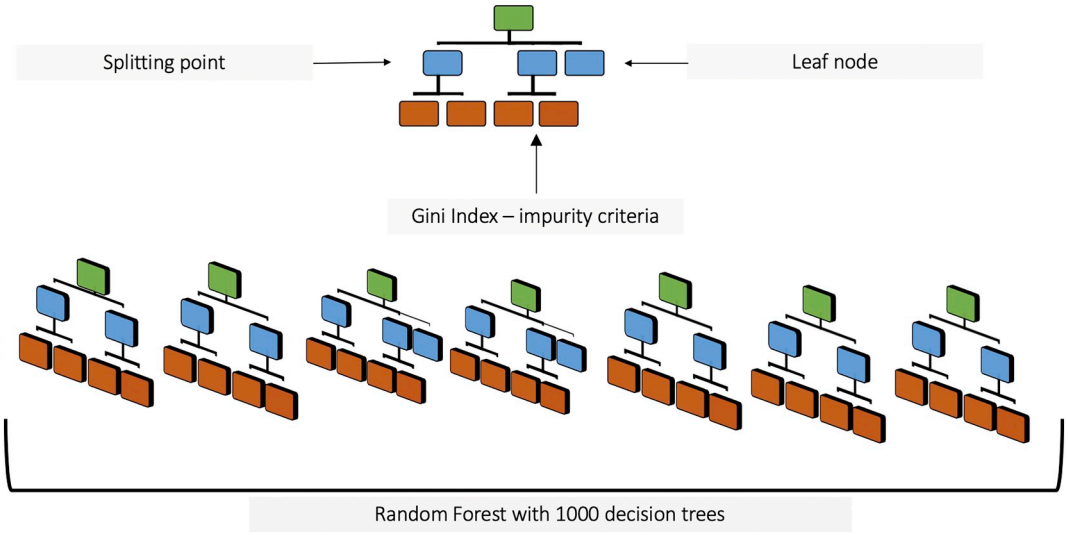

Fig. 3. A pictorial representation of the Random Forest algorithm and the Gini Index splitting criteria.

third, second, and first important variables is clearly substantial (see Fig. 4.)

On the second group of important CPI predictors, the Overall Economic Freedom score may be thought of as a deliberately broad concept that has been used by a number of studies to capture the overall economic environment of nations across the globe. There has been evidence highlighting the potential influence of Economic Freedom on corruption measures (Graeff & Mehlkop, 2003). Although not entirely surprising, the fact that this variable appeared in fifth place among the most important CPI predictors may indicate that the direction of the causal relationship may be different than what has been accepted. For example, according to the Heritage Foundation, corruption tends to corrode Economic Freedom by spoiling confidence among economic players, including government actors. Conversely, and in light of the impressive predictive accuracy results of our study (85%), it is probable that Economic Freedom influences corruption across countries and not the opposite. The Overall Economic Freedom variable was split 46.37% of the times it was chosen as a candidate to split (see Fig. 4.) The second group of variables include Trading across Borders - Time to Export: Documentary Compliance (Hours) (DB16–18 methodology), Business Freedom, Registering Property - Quality of Land Administration Index (0-30) (DB16 methodology), Registering Property - Quality of Land Administration Index (0-30) (DB17-18 methodology), Getting Electricity - DTF - Getting Electricity (DB16-18 methodology), Getting Electricity - Cost (% of Income per Capita), Paying Taxes - DTF - Paying Taxes (DB17-18 methodology), Resolving Insolvency – DTF - Resolving Insolvency (DB15-18 methodology), Resolving Insolvency - Recovery Rate (Cents on the Dollar), Trading across Borders - Time to Import: Documentary Compliance (Hours) (DB16-18 methodology), Trade Freedom and Dealing with Construction Permits - DTF - Dealing with Construction Permits (DB16–18 methodology). Although these factors influence CPI, their contribution is rather less substantial than the group 1 variables. It is worth noting that to some extent, our results echo the regression-based work of Murphy (2015) by suggesting that Trade Freedom may have only a moderate level of importance when predicting macroeconomics.

A non-exhaustive list of the last group of predictors includes Starting a Business - Cost - Men (% of Income per Capita), Financial Freedom, Trading across Borders - DTF - Trading across Borders (DB16-18 methodology), Trading across Borders - Time to Export: Border Compliance (Hours) (DB16-18 methodology), Getting Electricity - Total Duration and Frequency of Outages per Customer a Year (0-3) (DB16–-18 methodology), Investment Freedom, Starting a Business - Cost - Women (% of Income per Capita), Registering Property - Geographic Coverage Index (0-8) (DB16-18 methodology), Getting Electricity - Reliability of Supply and Transparency of Tariff Index (0-8) (DB16-18 methodology), Trading across Borders - Time to Import: Border Compliance (Hours) (DB16-18 methodology) and Paying Taxes - DTF-

Table 3
Tabulation of prediction results based on the *10*-fold cross-validation methodology.

Model Type	Confusion Matrices*				Accuracy	VH CPI Class Accuracy	H CPI Class Accuracy	L CPI Class Accuracy	VL CPI Class Accuracy
	VH CPI	H CPI	L CPI	VL CPI					
Artificial Neural Networks (ANN)	VH CPI 41	9	8	1	73.84%	69.49%	63.24%	74.24%	83.58%
	H CPI 8	43	13	4					
	L CPI 5	7	49	5					
	VL CPI 0	4	7	56					
Random Forest (RF)	VH CPI 53	6	0	0	**85.77%**	**89.83%**	**80.88%**	**78.79%**	**94.03%**
	H CPI 6	55	7	0					
	L CPI 0	8	52	6					
	VL CPI 0	1	3	63					
Support Vector Machines (SVM)	VH CPI 49	10	0	0	76.15%	83.05%	64.71%	74.24%	82.09%
	H CPI 15	44	8	1					
	L CPI 1	9	49	7					
	VL CPI 1	2	9	55					

* The number of predicted classes are stated on the rows and the number of actual classes are stated on the columns. The boldface in the last two columns indicate the best prediction results.

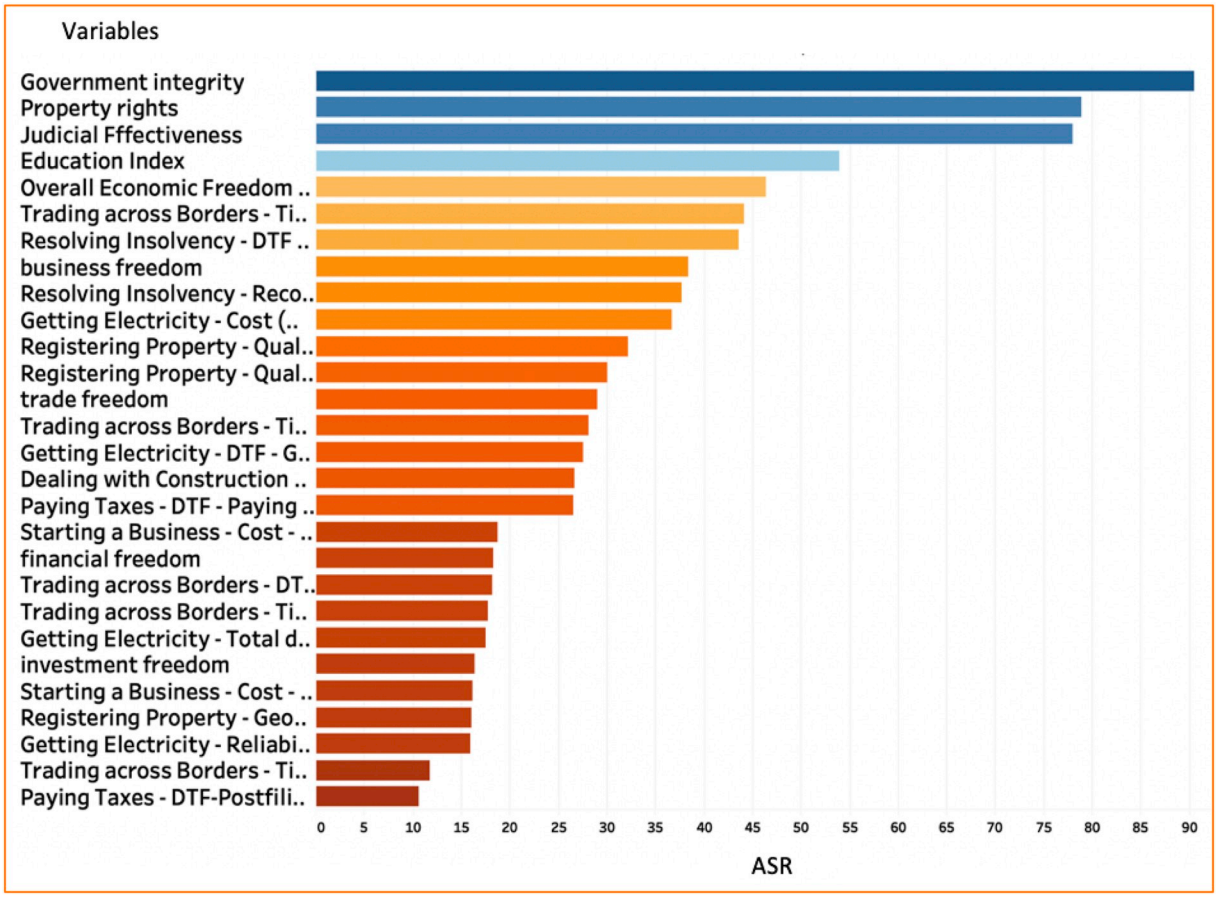

Fig. 4. Variables that were split at least 10% of the times when they were randomly selected as part of the candidates list.

Postfiling Index (0-100) (DB17-18 methodology), Monetary Freedom, Government Spending, Paying Taxes – Time (Hours per Year), Tax Burden, Fiscal Health, Region and Labor Freedom. Interestingly, even though the Investment Freedom score is often used as part of the inputs to calculate the Overall Economic Freedom score (which is ranked next to the first group of predictors in Fig. 4), our results pointed out a rather marginal predictive contribution of this variable (i.e., Investment Freedom score). Moreover, we contrast extant field suggestions regarding a potentially relevant and significant influence of financial freedom, monetary freedom, and trade freedom as predictors of CPI. These findings may be associated with the fact that these variables were never studied holistically as a group, through the lenses of machine learning, classification, and pattern recognition techniques. These modern-day machine learning methods use the variable space as a collective information source, where some of the obvious variables with direct individual influence on CPI may in fact turn out to be less significant contributors to the collective predictive power.

5. Summary and conclusion

In this study, we explored several potential predictors for Corruption Perception Indexes across 132 countries. To better capture variable influence and to ensure credibility, we opted to employ input variables that are provided by well-recognized and respected institutions such as the World Bank, Transparency International, and the Heritage Foundation for the years of 2017 and 2018. After various experiments with diverse model types and modeling parameters, we chose Random Forest, ANN, and SVM to explore and test 30 prediction models (3 model types, each with ten folds). In contrast with the majority of classification problems that use the binary class output option (e.g., win-lose, yes-no), we imposed a methodological challenge by assessing four levels of ordinal classes (Very Low Corruption, Low Corruption, High Corruption, Very High Corruption) for better granularity. Despite using multiple class options, the predictive accuracy achieved was notably high. The cross-validation results highlighted Random Forest as the most accurate classification method to predict CPI. SVM and ANN were, respectively, the second and the third best models showing satisfactory prediction performances.

We know that when eroding Government Integrity through bribery, fraud, nepotism, and misappropriation, corruption can become systemic and difficult to revert (Damania et al., 2004). Interestingly, given the notable predictive accuracy of our results, we point out that Government Integrity is the most significant predictor of corruption, and largely based in the previous theoretical works, we can say that Government Integrity bolsters corruption and not the other way around. This potential inversion of the causal relationship between variables, in this case, Corruption and Government Integrity, is not rare in regression-based studies, causing ambiguity for scholars and practitioners. 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. As previously explained, by decreasing the degree of impurity (Gini impurity) between categorical classes within a decision tree and improving predictability, the generation of splitting points within the forest provides meaningful and actionable insights that are not necessarily based on linear relationships.

That said, it is worth mentioning that most machine learning algorithms are considered predictive instruments with reduced descriptive capabilities. Although the predictive accuracy of machine learning models has been consistently higher than traditional regression models, their functioning process has been called a “black box” by some

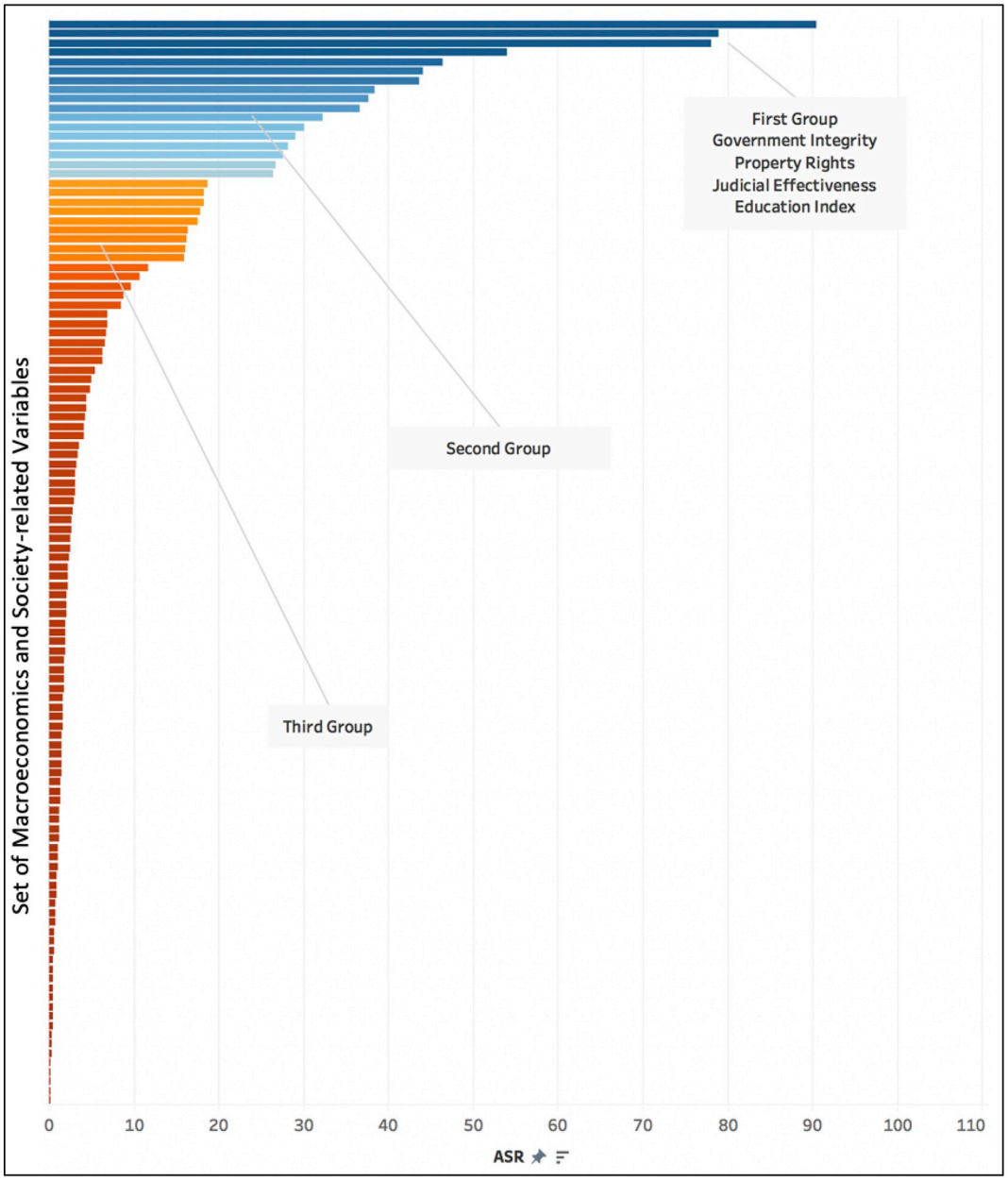

Fig. 5. A pictorial representation of the three groups of variables (related to the order of importance according to the ASR from the Random Forest algorithm).

scholars because a machine learning model trains itself, assuming that data is generated in complex ways that are not necessarily correlated to produce accurate predictions of a certain outcome. In other words, in machine learning, the relationship between input and output variables can only be inferred through heuristic methods of experimentation, keeping the main focus on predictive accuracy. Therefore, the lack of explicit theoretical explanations regarding relationships between variables (strength and direction of influence) may be stated as one of the limitations of predictive analytics studies, including the one proposed in this paper. To further explore a potential positive or negative effect of input variables, we encourage future research to couple predictive analytics with the use of powerful descriptive and theoretical studies available in the extant literature. Also, data-related constraints should be pointed out as part of our study limitations. For example, for the vast majority of the economies assessed by the Doing Business research team, the data was drawn from the largest business cities in each country, which may not be representative of other regions of the same country. As a potential limitation, the collection of data on all regions of all countries does not seem to be feasible given the enormous amount of time and resources that would need to be applied by related institutions. Although we explore a reliable and complete dataset, addressing 132 countries across the globe, it is worth mentioning that data from economies like Russia, Egypt, and Pakistan were not entirely available. In addition, although the CPI measure has been widely used by academics and organizations in general, it reflects corruption perceptions to the neglect of factual corruption cases. In this sense, future research would benefit from collecting, merging, and assessing datasets that reliably report actual corruption cases across countries through the leans machine learning algorithms.

Corruption is a ubiquitous, challenging, thought-provoking, and research-worthy social phenomenon. However, as the previous research efforts suggested, using limited number of well-known macroeconomics indicators to predict corruption may not yield reliable results and practical implications. The real value lays on the use of rich dataset to actually reveal and understand the important factors that exert influence on corruption perceptions across societies. In this sense, the

important corruption predictors that we identified and described in the previous section may serve as valuable insights for public administrators and policy makers who focus their efforts on efficiently combating corruption in their countries. As previously stated, due to its focus on predicting results with the highest possible accuracy, data mining and machine learning have unearthed hidden patterns of data that can help society in many ways. Some examples of similar machine learning uses include targeting tax fraud, credit card fraud detection, medical and health care research, manufacturing processes, and banking risk management (Sharda et al., 2017). Similarly, the current results reveal meaningful knowledge that should be properly explored by social scientists, public agencies, and individuals who seek better societal structures for themselves and future generations. The Big Data era in which we live comes with unprecedented levels of accelerated knowledge discovery that in turn, pushes efficient and effective decision making towards the desired and needed levels of accuracy. Pragmatically embracing the insights produced by these powerful machine learning techniques adopted for this research can be a key strategical initiative for government administrators, individuals, and scholars to understand and improve ethics to reduce corruption within societies.

References

Acemoglu, D., & Verdier, T. (1998). Property rights, corruption and the allocation of talent: A general equilibrium approach. *The Economic Journal, 108*, 1381–1403. https://doi.org/10.1111/1468-0297.00347.

Angulo-Guerrero, M. J., Pérez-Moreno, S., & Abad-Guerrero, I. M. (2017). How economic freedom affects opportunity and necessity entrepreneurship in the OECD countries. *Journal of Business Research, 73*, 30–37. https://doi.org/10.1016/j.busres.2016.11.017.

Apergis, N., & Cooray, A. (2017). Economic freedom and income inequality: Evidence from a panel of global economies—A linear and a nonlinear long-run analysis. *The Manchester School, 85*, 88–105. https://doi.org/10.1111/manc.12137.

Apergis, N., Dincer, O. C., & Payne, J. E. (2010). The relationship between corruption and income inequality in U.S. states: Evidence from a panel cointegration and error correction model. *Public Choice, 145*, 125–135. https://doi.org/10.1007/s11127-006-9557-1.

Archer, K. J., & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis, 52*, 2249–2260. https://doi.org/10.1016/j.csda.2007.08.015.

Blanshard, O., & Schleifer, A. (2000). Federalism with and without political centralization: China versus Russia. *IMF Staff Papers, 48*, 171–179. https://doi.org/10.2307/4621694.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on computational learning theory* (pp. 144–152). New York: Association of Computer Machinery. https://doi.org/10.1145/130385. 130401.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A: 1010933404324.

Chang, E., & Golden, M. (2007). Electoral systems, district magnitude and corruption. *British Journal of Political Science, 37*, 115–137. https://doi.org/10.1017/S0007123407000063.

Clark, S. D., Morris, M. A., & Lomax, N. (2018). Estimating the outcome of UKs referendum on EU membership using e-petition data and machine learning algorithms. *Journal of Information Technology & Politics, 15*(4), 344–357. https://doi.org/10.1080/19331681.2018.1491926.

Cortes, C., Gonzalvo, X., Kuznetsov, V., Mohri, M., & Yang, S. (2016). Adanet: Adaptive structural learning of artificial neural networks. *Proceedings of the 34th international conference on Mechine learning. Vol. 70. Proceedings of the 34th international conference on Mechine learning* (pp. 874–883). JMLR ORG.

Damania, R., Fredriksson, P. G., & Mani, M. (2004). The persistence of corruption and regulatory compliance failures: Theory and evidence. *Public Choice, 121*, 363–390. https://doi.org/10.1007/s11127-004-1684-0.

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.

Delen, D., Oztekin, A., & Kong, Z. J. (2010). A machine learning-based approach to prognostic analysis of thoracic transplantations. *Artificial Intelligence in Medicine, 49*, 33–42. https://doi.org/10.1016/j.artmed.2010.01.002.

Delen, D., Sharda, R., & Bessonov, M. (2006). Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. *Accident Analysis & Prevention, 38*, 434–444. https://doi.org/10.1016/j.aap.2005.06.024.

Delen, D., Tomak, L., Topuz, K., & Eryarsoy, E. (2017). Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods. *Journal of Transport & Health, 4*, 118–131. https://doi.org/10.1016/j.jth.2017.01.009.

Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. *Artificial Intelligence in Medicine, 34*(2), 113–127. https://doi.org/10.1016/j.artmed.2004.07.002.

Depken, C. A., & La Fountain, C. L. (2006). Fiscal consequences of public corruption: Empirical evidence from state bond ratings. *Public Choice, 126*, 75–85. https://doi.org/10.1007/s11127-006-4315-0.

Dixit, A. (2009). Governance institutions and economic activity. *American Economic Review, 99*, 5–24. https://doi.org/10.1257/aer.99.1.5.

Donchev, D., & Ujhelyi, G. (2008). *What Do Corruption Indices Measure?* Working PaperUniversity of Houston Economics Department.

Dong, B., & Torgler, B. (2013). Causes of corruption: Evidence from China. *China Economic Review, 26*, 152–169 (DOI: 10.106/j.chieco.2012.09.005).

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics, 35*, 352–359. https://doi.org/10.1016/S1532-0464(03)00034-0.

Duro, D. C., Franklin, S. E., & Dubé, M. G. (2012). A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. *Remote Sensing of Environment, 118*, 259–272. https://doi.org/10.1016/j.rse.2011.11.020.

Fisman, R., & Gatti, R. (2002). Decentralization and corruption: Evidence from U.S. federal transfer programs. *Public Choice, 113*, 25–35. https://doi.org/10.1023/A:1020311.

Glaeser, E. L., & Saks, R. E. (2006). Corruption in America. *Journal of Public Economics, 90*, 1053–1072. https://doi.org/10.1016/j.jpubeco.2005.08.007.

Goel, R. K., & Nelson, M. A. (1998). Corruption and government size: A disaggregated analysis. *Public Choice, 97*, 107–120. https://doi.org/10.1023/A:1004900.

Gong, T. (2015). Managing government integrity under hierarchy: Anticorruption efforts in local China. *Journal of Contemporary China, 24*, 684–700. https://doi.org/10.1080/10670564.2014. 978151.

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.

Grimes, M. (2013). The contingencies of societal accountability: Examining the link between civil society and good government. *Studies in Comparative International Development, 48*, 380–402. https://doi.org/10.1007/s12116-012-9126-3.

Hillard, D., Purpura, S., & Wilkerson, J. (2008). Computer-assisted topic classification for mixed-methods social science research. *Journal of Information Technology & Politics, 4*(4), 31–46. https://doi.org/10.1080/19331680801975367.

Huang, X., Shi, L., & Suykens, J. A. (2014). Support vector machine classifier with pinball loss. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 36*, 984–997.

Hussmann, K. (2011). Addressing corruption in the health sector: Securing equitable access to health care for everyone. *U4 Issue, 2011*, 1.

Johannesen, J. K., Bi, J., Jiang, R., Kenney, J. G., & Chen, C. M. A. (2016). Machine learning identification of EEG features predicting working memory performance in schizophrenia and healthy adults. *Neuropsychiatric Electrophysiology, 2*, 3. https://doi.org/10.1186/s40810-016-0017-0.

Kaufmann, D., Kraay, A., & Mastruzzi, M. (2006). Measuring governance using cross-country perceptions data. *International handbook on the economics of corruption* (pp. 52). .

Kaufmann, D., Kraay, A., & Mastruzzi, M. (2009). *Governance matters VIII: Aggregate and individual governance indicators, 1996–2008.* World Bank Working Paper.

Klitgaard, R. (1988). *Controlling corruption.* Berkeley, CA: University of California Press.

Lambsdorff, J. G. (2003). *Background paper to the 2003 corruption perceptions index.* Transparency International Working Paper.

Lee, K. J., & Carlin, J. B. (2010). Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology, 171*, 624–632. https://doi.org/10.1093/aje/kwp425.

Lessig, L. (2011). *Republic, lost: How money corrupts congress—And a plan to stop it.* UK: Hatchette.

Li, H., Xu, L. C., & Zou, H. F. (2000). Corruption, income distribution, and growth. *Economics and Politics, 12*, 155–182.

Lipset, S. M. (1960). *The social bases of politics.* Baltimore MA: The Johns Hopkins University Press.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science, 5*, 161–171. https://doi.org/10.1111/1467-8721.ep11512376.

López-Iturriaga, F. J., & Sanz, I. P. (2017). Predicting public corruption with neural networks: An analysis of Spanish provinces. *Social Indicators Research, 140*, 1–24. https://doi.org/10.1007/s11205-017-1802-2.

Lord, D., & Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy Practice, 44*, 291–305. https://doi.org/10.1016/j.tra.2010.02.001.

Mauro, P. (1995). Corruption and growth. *The Quarterly Journal of Economics, 110*, 681–712. https://doi.org/10.2307/2946696.

Miller, S. (2011). Corruption. In E. N. Zalta (Ed.). *Stanford encyclopedia of philosophy*http://plato.stanford.edu/archives/spr2011/entries/corruption/.

Miller, T., Kim, A. B., & Holmes, K. R. (2015). *2015 index of economic freedom.* Washington D.C.: The Heritage Foundation.

Mitchell, D. T., & Campbell, N. D. (2009). Corruption's effect on business venturing within the United States. *American Journal of Economics and Sociology, 68*, 1135–1152. https://doi.org/10.1111/j.1536-7150.2009.00665.x.

Mocan, N. (2008). What determines corruption. International evidence from microdata. *Economic Inquiry, 46*, 493–510. https://doi.org/10.1111/j.1465-7295.2007.00107.x.

Murphy, R. H. (2015). The impact of economic inequality on economic freedom. *Cato Journal, 35*, 117.

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.

Olken, B. (2007). Monitoring corruption: Evidence form a field experiment in Indonesia. *Journal of Political Economy, 115*, 200–249. https://doi.org/10.1086/517935.

Olson, D., & Delen, D. (2008). *Advanced data mining techniques.* New York: Springer.
Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2*, 559–572. https://doi.org/10.1080/14786440109462720.
Persson, T., Tabellini, G., & Trebbi, F. (2003). Electoral rules and corruption. *Journal of the European Economic Association, 1*, 958–989. https://doi.org/10.1162/154247603322493203.
Philp, M., & Dávid-Barrett, E. (2015). Realism about political corruption. *Annual Review of Political Science, 18*, 387–402 (DOI: 10.116/annurev-polisci-092012-134421).
Power, T. J., & Taylor, M. M. (2011). *Corruption and democracy in Brazil: The struggle of accountability.* Notre Dame, IN: University of Notre Dame Press.
Prud'homme, R. (1994). On the dangers of decentralization. *The World Bank Research Observer, 10*, 201–220. https://doi.org/10.1093/wbro/10.2.201.
Razafindrakoto, M., & Roubaud, F. (2010). Are international databases on corruption reliable? A comparison of expert opinion surveys and household survey in Sub-Saharan Africa. *World Development, 38*, 1057–1069. https://doi.org/10.1016/j.worlddev.2010.02.004.
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.
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.
Rose-Ackerman, S. (1996). Altruism, nonprofits, and economic theory. *Journal of Economic Literature, 34*, 701–728.
Schäfer, J., & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology, 4*. https://doi.org/10.2202/1544-6115.1175.
Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33*, 545–571. https://doi.org/10.1207/s15327906mbr3304_5.
Seifert, J. W. (2004). Data mining and the search for security: Challenges for connecting the dots and databases. *Government Information Quarterly, 21*(4), 461–480.
Seligson, M. A. (2006). The measurement and impact of corruption victimization: Survey evidence from Latin America. *World Development Journal, 34*, 381–404. https://doi.org/10.1016/j.worlddev.2005.03.012.
Shah, A. (1998). *Balance, accountability, and responsiveness: Lessons about decentralization.* Washington, D.C.: The World Bank.
Sharda, R., Delen, D., & Turban, E. (2017). *Business intelligence, analytics, and data science: A managerial perspective* (4th ed.). London: Pearson.
Shleifer, A., & Vishny, R. (1993). Corruption. *Quarterly Journal of Economics, 108*, 599–617. https://doi.org/10.2307/2118402.
de Sousa, W. G., de Melo, E. R. P., Bermejo, P. H. D. S., Farias, R. A. S., & Gomes, A. O. (2019). How and where is artificial intelligence in the public sector going? A literature review and research agenda. *Government Information Quarterly* (in press number 101392).
Stockemer, D. (2018). The internet: An important tool to strengthening electoral integrity. *Government Information Quarterly, 35*(1), 43–49.
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.
Svetnik, G., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: A classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences, 43*, 1947–1958. https://doi.org/10.1021/ci034160g.
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).
Tanzi, V. (1995). *Fiscal federalism and decentralization: A review of some efficiency and macroeconomic aspects.* Washington, D.C.: CDC World Bank295–316.
Thompson, D. (2018). Theories of institutional corruption. *Annual Review of Political Science, 21*, 495–513. https://doi.org/10.1146/annurev-polisci-120117-110316.
Thompson, D. F. (2005). Two concepts of corruption. *George Washington Law Review, 73*, 1036–1069.
Varma, M. K. S., Rao, N. K. K., Raju, K. K., & Varma, G. P. S. (2016). Pixel-based classification using support vector machine classifier. *Advanced computing (IACC), 2016 IEEE 6th international conference* (pp. 51–55). .
Wolfe, P. J. (2013). Making sense of big data. *Proceedings of the National Academy of Sciences, 110*, 18031–18032. https://doi.org/10.1073/pnas.1317797110.
Wu, D., Jennings, C., Terpenny, J., Gao, R. X., & Kumara, S. (2017). A comparative study on machine learning algorithms for smart manufacturing: Tool wear prediction using random forests. *Journal of Manufacturing Science and Engineering, 139*, 071018. https://doi.org/10.1115/1. 4036350.
Zhang, Y., Dong, Z., Liu, A., Wang, S., Ji, G., Zhang, Z., & Yang, J. (2015). Magnetic resonance brain image classification via stationary wavelet transform and generalized eigenvalue proximal support vector machine. *Journal of Medical Imaging and Health Informatics, 5*, 1395–1403. https://doi.org/10.1166/jmihi/2015.1542.

Marcio Salles Melo Lima received his doctoral degree from Spears School of Business at Oklahoma State University. He is the director of Research and Development at Metalsider Inc. in Betim, MG, Brasil. Dr. Lima's research focus application of big data and analytic methods to social problems including prediction and explanation of corruption, use of data analytics in human resource management, and prediction of time series data.

Dursun Delen is the holder of Spears and Patterson Endowed Chairs in Business Analytics, Director of Research for the Center for Health Systems Innovation, and Regents Professor of Management Science and Information Systems in the Spears School of Business at Oklahoma State University. He has authored/co-authored more than 150 peer-reviewed articles. His research has appeared in major journals including Decision Sciences, Journal of Production Operations Management, Decision Support Systems, Communications of the ACM, Computers and Operations Research, Computers in Industry, Journal of the American Medical Informatics Association, Artificial Intelligence in Medicine, International Journal of Medical Informatics, Health Informatics Journal, among others. He has published ten books/textbooks in the broad area of Business Intelligence and Business Analytics. He is often invited to national and international conferences and symposiums for keynote addresses, and companies and government agencies for consultancy projects on data science and analytics related topics. Dr. Delen served as the general co-chair for the 4th International Conference on Network Computing and Advanced Information Management (held in Soul, South Korea), and regularly chairs tracks and mini-tracks at various information systems and analytics conferences. He is currently serving as the editor-in-chief, senior editor, associate editor and editorial board member of more than a dozen academic journals.