Using Political Party Affiliation Data to Measure Civil Servants Risk of Corruption

Ricardo S. Carvalho
and Rommel N. Carvalho
Department of Strategic Information
Brazilian Office of the Comptroller General
SAS, Quadra 01, Bloco A, Edifício Darcy Ribeiro
Brasília, Distrito Federal, Brazil
ricardo.carvalho@cgu.gov.br and rommel.carvalho@cgu.gov.br

Marcelo Ladeira
Department of Computer Science
University of Brasília
Campus Universitário Darcy Ribeiro
Brasília, Distrito Federal, Brazil
mladeira@unb.br

***Abstract*—This paper presents a case study of machine learning applied to measure the risk of corruption of civil servants using political party affiliation data. Initially, a statistical hypothesis test verified the dependency between corruption and political party affiliation. Then, we constructed datasets with standardization and three different discretization techniques. Using Weka environment, this work shows the application and statistical evaluation of four classification algorithms to build models for predicting risk of corruption: Bayesian Networks, SVM, Random Forest, and Artificial Neural Networks with backpropagation. To evaluate the models we used data mining metrics such as precision, recall, kappa statistic and percent correct. Lastly, the case study compares the learned model with the best performance to the specialists model. The comparison not only confirms previous specialist affirmations, but also provides new assertions on the affiliation-corruptibility relation. CRISP-DM process model was the base reference for the data mining phases.**

I. Introduction

It is known that, nowadays, the theme of corruption has high priority in Brazil's agenda [17], being fundamentally necessary its efficient and intensive fight. Public corruption can be defined as a social relation (personal, extra market, and illegal) established between two agents or two groups of agents (corrupts and corruptors), whose goal is the illegal transfer of revenue, inside society or through public funds, for strictly private purposes [22]. Thereby, considering the high level of complexity involving social relations, there are numerous different aspects that can influence a corruption scenario.

In that manner, due to nonexistence of neutrality on both public and political spheres, as they carry out strong moral weight [2], it is perfectly possible to suggest a link between corruption and political affiliation. Keeping this in mind, the objective of this paper is to verify if there is a dependency between party affiliation and civil servants corruption risk. Finally, another objective of this work is to create a model capable of explaining the corruption-affiliation relation and classifying a given civil servant as corrupt given his/her political party affiliation data. For that, four machine learning algorithms are evaluated using Weka [28], namely: Bayesian Networks [24] (BN), Random Forests [3], Support Vector Machines (SVM) [8] and Artificial Neural Networks (ANN) [27]. This model is developed as part of a broader system and it will be used in the daily activities of the Department of Research and Strategic Information (DIE) at the Brazilian Office of the Comptroller General (CGU).

Currently, DIE already uses a model built by specialists. Through empirical knowledge, gained from experience, they created affirmations that sustain the model. The specialists model uses two attributes: number of parties one is affiliated and motive of affiliation cancellation. Specialists model is considered by DIE as conservative and exhibiting high precision. Nonetheless they have never done statistical tests to determine metrics like precision. Thus, an important step in this case study is comparing the learned model with the best performance to the specialists model. This comparison aims to assess both models and define the most suited to predict corruption from affiliation data. The evaluation intends to confirm or decline previous specialist affirmations. Furthermore, its purpose is providing previously unknown assertions on the affiliation-corruptibility relation. Along theses lines, this work also attempts to justify the use of automated generated models instead of the usual subject matter expert manually designed model.

This research followed the Cross Industry Standard Process for Data Mining (CRISP-DM) [21] in order to learn the desired model. This process has six major phases. The Business Understanding phase is responsible for determining business objectives and assessing the situation. The Data Understanding phase is responsible for exploring the data and drawing initial conclusions about it. The Data Preparation phase involves the construction and formating of data. The Modeling phase supports the construction of models. The Evaluation phase will cover methods to validate the model, to assess results, and to define the final model. Finally, in the Deployment phase the generated and selected model will be deployed, a final report will be written, and a plan to improve the model might be devised.

Section II discusses DIE's responsibilities regarding corruption. It defines business goals in terms of corruption prediction precision. We also describe the main objectives and assess the current data gathering scenario. Section III depicts the

understanding of the main dataset used in this work. This data is explored using a Chi-Square Hypothesis Test for Independence. We verify if the hypothesis that affiliation and corruption are dependent for civil servants is acceptable. Section IV displays the steps taken to retrieve and manipulate data. It describes the process of manually cleaning affiliation data and aggregating them into six attributes. Additionally, this sections goes through discretization steps in order to enable the use of algorithms that only work with categorical data, such as Bayesian Networks. Three different discretization methods implemented in Weka are used.

Section V exhibits the use of classification algorithms to build models that explains the corruption-affiliation relation. It shows the implementation of four classifiers using Weka: Bayesian Networks, Random Forests, SVM, and Artificial Neural Networks with backpropagation. We briefly describe the algorithms and outline the parameters used. Section VI details the evaluation of the models built from the 10 datasets with the four classifiers using 10-fold cross-validation. It uses the metrics precision, recall, kappa statistic, Mean Absolute Error and percent correct. We show the assessment to chose the most suited dataset. Then this is used to compare the machine learning algorithms and define the final model. This section also depicts the comparison between the specialists model from DIE and final model using a separate dataset. Additionally, it analyzes specialist affirmations and previously unknown assertions on the affiliation-corruptibility relation. Lastly this paper ends in Section VII with Conclusion.

II. Business Understanding

One of CGU's main responsibilities is inspecting and detecting frauds in the use of federal public funds. DIE is responsible for activities related to the investigation of possible irregularities committed by civil servants. Nowadays, there are approximately 600 thousand active civil servants, who are subject to investigation. Since DIE has a reduced staff, it is extremely important to prioritize its activities based on who presents the higher risk of being involved in corruption. That being said, it's imperative that the results of the model built to predict corruptibility are approximately 90% precise, to avoid wasting workforce investigating an innocent person.

Therefore, this case study aims to create a classification model of civil servants affiliated to political parties, where the analyzed class is an individual's corruptibility. For that, it's necessary to obtain information about civil servants, specifically those affiliated to political parties, and to separate them into corrupts and non corrupts. The government keeps servant's registrations in the Integrated Human Resources Management System (SIAPE, in Portuguese) [1]. While the information about political party affiliation is available at Superior Electoral Court (TSE, in Portuguese)[2]. Regarding corrupt servants, it's possible to access the Registry of Expelled from the Federal Administration (CEAF, in Portuguese) [3] and, after filtering its data, to obtain those expelled only due to corruption.

However, information about non corrupt servants does not exist anywhere. Mainly because characteristics related to non corrupts, like ethics and moral, involve value judgment [26]. Hence it is impossible to assess this kind of information objectively. To attend this issue, in this work we initially took a random sample of civil servants. Then, to try to minimize their corruptibility, we discarded from the sample those present in selected databases. This selection picked databases containing any information that could lead to corruption. That includes those keeping penalty registries, like Accounts Deemed Irregular from the Federal Court of Accounts (TCU) [4] and Administrative Disciplinary Processes from CGU [5]. As well as others that only have possible indirect effects on corruption, such as Corporate Relationships from Secretariat of the Federal Revenue [6] and Government Bank Orders from SIAFI [7].

III. Data Understanding

As the focus of this case study is treating affiliated civil servants, the main dataset is generated intersecting SIAPE and TSE. This is possible by connecting them using the unique number that identifies citizens in Brazil: the Individual Taxpayer Registry (CPF, in Portuguese) [8]. With the main dataset in place, its intersection with CEAF results in corrupt affiliated civil servants data. Whereas joining the main dataset to the created non corrupt database, described on Section II, produces non corrupt affiliated civil servants information.

In order to verify the dependency between party affiliation and risk of corruption of civil servants, this paper uses a random sample of civil servants to generate quantities about affiliation and corruption, as shown in Table I. For that purpose, we apply a Chi-Square Hypothesis Test for Independence [29] using the numbers presented. The null hypothesis states that knowing the affiliation does not help you predict corruption, that is, they are independent. The analysis uses the sample to accept or reject the null hypothesis. Considering Table I, the chi-square random variable calculated is $\chi^2_{calc} \approx 158$, whereas the chi-square table value for one degree of freedom and significance level of 1% is $\chi^2_{tab} \approx 6.64$. Since the calculated value is bigger than the table value, we can not accept the null hypothesis. Thus, the conclusion is that the data supports the hypothesis that affiliation and corruption are dependent for civil servants.

IV. Data Preparation

For the data preparation, we used Microsoft SQL Server Management Studio [9] to retrieve and to manipulate the data.

[1] http://www.siapenet.gov.br/Portal/Servico/Apresentacao.asp
[2] http://www.tse.jus.br/partidos/filiacao-partidaria/relacao-de-filiados
[3] http://www.portaldatransparencia.gov.br/expulsoes/entrada
[4] https://contas.tcu.gov.br/cadirreg/CadirregConsultaNome
[5] http://www.cgu.gov.br/cguPad/
[6] http://www.receita.fazenda.gov.br/
[7] https://www.serpro.gov.br/
[8] http://www.receita.fazenda.gov.br/pessoafisica/cpf/cadastropf.htm
[9] http://www.microsoft.com/pt-br/download/details.aspx?id=7593

TABLE I
SAMPLE FOR AFFILIATION AND CORRUPTION

Class	Corrupts	Non Corrupts	Total
Affiliated	161	5.742	5.903
Non Affiliated	411	45.158	45.569
Total	572	50.900	51.472

Information from SIAPE's database was not in the scope for this case study. Therefore we only used it to get civil servants by CPF and did not get any functional data. We made this choice because the paper aims to assert corruption relation to affiliation data alone, not including any other information. After obtaining data for each CPF, as stated in Section III, the initial count was 1994 different corrupts affiliated. Then, we needed random undersampling on the non corrupts affiliated to get approximately the same count. This way we obtained a dataset with balanced classes.

At first, we manually cleaned affiliation data removing records with invalid values for fields like dates. Later, we aggregated data for each CPF into six attributes using counting or maximum functions. Then, we selected[10] the three most representatives attributes with correlation-based feature selection [14] implemented in Weka. The first attribute is the sum of the number of days each CPF was affiliated to parties. In the case of existing disaffiliation, the calculation is the disaffiliation date minus the affiliation date. However, we need to specify definitions for the case of still regular affiliation, since disaffiliation date does not exist. To fill the gap, in the group of corrupts the attribute is the difference between the sanction date[11] and the affiliation date. In the non corrupts group, a recent date is used instead of the sanction date. For that matter, we evaluate the information until the day of sanction or a recent day. Therefore, for corrupts, we discarded all affiliation data after the sanction date.

The second attribute used is the maximum number of days each CPF was affiliated to a party. We used the same definitions as in the first feature to compute the number of days in the second one. Finally, the third feature selected is the highest value of a code that represents the affiliation cancellation motive for each CPF. The codes are the following:

0 for the absence of motive;
1 for the elector's choice;
2 for the party's choice;
3 for system's automatic cancellation; and
4 for judicial cancellation.

Additionally, we also took discretization steps in order to enable the use of algorithms that only work with categorical data, such as Bayesian Networks. We used three different methods implemented in Weka. One of them is the Multi-interval discretization of continuous valued attributes for classification learning [11]. This method uses an entropy minimization heuristic derived from minimum description length principle for deciding the intervals. Another method is the Equal-Frequency Binning [20]. This one determines the minimum and maximum values of the discretized attribute, sorts all values in ascending order, and divides the range into a user-defined number of intervals so that every interval contains the same number of sorted values. We divided the continuous attributes in this case study in 10 intervals. The third method was the Proportional k-Interval Discretization [32]. It uses equal-frequency binning but it adjusts the number and size of discretized intervals to the number of training instances – *i.e.*, the number of bins is equal to the square root of the number of non-missing values.

We used all three discretization methods on the two numeric attributes that sums days, resulting in nine different datasets. Also all numeric attributes were standardized to have zero mean and unit variance to reduce the chances of overfitting. Therefore, we produced 10 final different datasets.

V. MODELING

This work used the Waikato Environment for Knowledge Analysis (Weka) software as a data mining tool. Weka reads Attribute Relation File Format (ARFF) [12] and provides an useful GUI to explore and model data [13]. We used the tool to execute the selected machine learning algorithms on datasets generated in Section IV. We defined all three selected attributes[12] as numeric. While the class, named corrupt, has two categorical values: Corrupt and Non-Corrupt.

To begin with machine learning, we randomly separated data in training and test datasets following an approximate 90-10% proportion and keeping representativity of classes. Quantities of instances per dataset are shown on Table II. Also, for model selection, we chose 10-fold cross-validation to be used on the training set [19]. It randomly divides the dataset into k mutually exclusive subsets (the folds) of approximately equal size. Each time it uses nine folds to train and one to test. Finally, it estimates the overall accuracy with the average accuracy calculated for each train/test step.

TABLE II
NUMBER OF INSTANCES IN THE TRAINING AND TEST DATASETS

Dataset	Corrupts	Non Corrupts	Total
Training	1.855	1.855	3.710
Test	139	142	281
Total	1.994	1.997	3.991

This paper evaluates four machine learning algorithms using Weka: Bayesian Networks, Random Forests, SVM, and Neural Networks with backpropagation. They are used to select the model that explains the corruption-affiliation relation and generates corruption classification for any civil servant from its political party affiliation data.

Bayesian Networks [16] was one of the chosen algorithms since it is a competitive classifier [7]. Besides, another advantage is the semantics of the generated graph with its

[10]Activity not in the scope of this case study
[11]Obtained from CEAF
[12]Number of days affiliated; Maximum number of days affiliated to one party; and Highest motive of cancellation code

random variables, since it is easy to understand for a human being what the variables and their relation mean. Thus, it can provide important information for specialists working with corruption, giving them new insights. Random Forests [3] has been one of the most used algorithms. It had excellent performance compared with many others on several metrics [6], including precision, which is in the main business goals of this case study. SVM [25] has been a successful modeling and prediction tool for a variety of applications. Also it was identified as one of the top 10 data mining algorithms by the IEEE International Conference on Data Mining (ICDM) in December 2006 [31]. Neural networks with backpropagation [15] has been one of most widely applied algorithms for classification. Studies using it include recent work in areas with risk evaluation [1], [18], such as this case study.

As stated in Section II, approximately 90% precision must be attained to result in a satisfactory model for DIE. With that in mind, we wrapped all classifiers in Weka's MetaCost [10]. This wrapper does cost-sensitive classification by explicitly defining a cost matrix, forcing the algorithms to minimize the expected cost. This way, a higher cost is assigned to predictions for corrupts that are false positive. That is, the true instance is non-corrupt but the model predicts it as corrupt. Thus, we used the cost matrix [0.0 1.0; 5.0 0.0] in MetaCost with bag size of 100%. We picked the weight 5 for false positives after manual trials showing that higher values overfitted the model and lower did not achieve approximately 90% precision.

For Random Forest we used a number of one hundred trees, as suggested by its author on his paper [3]. Additionally, the size of the feature set considered was the maximum, *i.e.*, three, since that is a small number. For Bayesian Networks, we applied the most used search algorithm K2 [9] and eliminated initialization as a naive bayes network. Also we did not set a maximum number of parents for each node and randomized initial order of the nodes. These choices were made to allow a model without independence assumptions, such as Naive Bayes. Finally, we used Weka's default settings for SVM and Neural Networks. The only change was the removal of preprocessing steps on Neural Networks.

VI. Evaluation

Making use of the Weka Experiment Environment, we evaluated each of the 10 datasets with the four algorithms using 10-fold cross-validation. We ran each algorithm-dataset combination one hundred times to obtain averaged results. From this experiment, we exhibit mean and standard deviations values for a number of metrics. The following are the metrics selected for evaluation:

1) Precision: the probability of predicting as corrupts those that really are corrupts. It is used due to business needs of avoiding non-corrupts incorrectly classifieds as corrupts;
2) Recall: number of corrupts classified as a proportion of all corrupts. It has to be assessed, since the precision-recall relation has a trade-off [4] [23], and high precision could lead to very low recall;
3) Kappa statistic: proportion of agreement beyond what is expected by chance [5]. Although high precision will lower kappa, it is still important to take into account reliability considering results by chance;
4) Mean absolute error (MAE): used because it is considered as an unambiguous measure of average error [30];
5) Percent correct: Percentage of observations correctly classified.

First, to choose the dataset to be used, we obtained the results computed for all four algorithms and averaged each metric per dataset. Estimates produced are shown in Table III. Analyzing recall, datasets one, three, four and seven have slightly better results. While kappa, MAE, and percent correct results show that datasets one and four have advantages. Then, considering precision values, dataset four has the best score. Therefore it is used to compare the models from the machine learning algorithms.

TABLE III
AVERAGE RESULTS FOR ALL FOUR ALGORITHMS PER DATASET

DS[1]	A1[2]	A2[3]	M1[4]	M2[5]	M3[6]	M4[7]	M5[8]
1	PKI	PKI	0,87	0,32	0,28	0,64	0,64
2	PKI	MI	0,88	0,28	0,23	0,64	0,62
3	PKI	EQF	0,87	0,30	0,25	0,63	0,63
4	MI	PKI	0,90	0,32	0,28	0,64	0,64
5	MI	MI	0,90	0,29	0,25	0,63	0,63
6	MI	EQF	0,89	0,29	0,25	0,63	0,63
7	EQF	PKI	0,88	0,31	0,27	0,63	0,63
8	EQF	MI	0,90	0,23	0,20	0,61	0,60
9	EQF	EQF	0,87	0,21	0,18	0,60	0,59
10	-	-	0,88	0,27	0,22	0,62	0,61

[1] Dataset number
[2] Discretization for attribute Number of days affiliated
[3] Discretization for attribute Maximum number of days affiliated
[4] Precision [5] Recall [6] Kappa statistic [7] (1 - MAE) [8] Percent correct

Thus, by choosing dataset number four, we define the discretization method for the attribute number of days affiliated as the Multi-interval and for maximum number of days affiliated as Proportional k-Interval. So, these methods are also applied to the test dataset for later evaluation.

Finally, for the classification algorithms studied, we only compare results from training dataset four. Table IV presents the computed values for the metrics selected.

TABLE IV
FINAL AVERAGED RESULTS PER ALGORITHM

Metric	BN[1]	SVM	RF[2]	ANN[3]
Precision	0,90	0,78	0,88	0,85
Recall	0,28	0,74	0,36	0,35
Kappa	0,25	0,53	0,31	0,29
MAE	0,38	0,23	0,34	0,33
Correct	0,63	0,77	0,66	0,64

[1] Bayesian Networks [2] Random Forests
[3] Artificial Neural Networks with backpropagation

Observing results on Table IV, SVM does not attend business requirement of a model around 90% precise. Bayesian Networks has 90% precision but displays significantly lower values of kappa and recall relative to the other algorithms. Though very close to each other, Random Forest still produces slightly better results than Neural Networks with backpropagation on average. Therefore, we selected Random Forest to build the final model.

Completing the evaluation, we compared the conservative specialists model from DIE to the Random Forest model using the test dataset. The results for the comparison are shown in Table V. Observing the metrics produced, the Random Forest model clearly shows better results while keeping the same precision. The recall presents a substantial increase, indicating a 15% higher percentage of corrupts predicted. The kappa measure displays a 13% gain. The percentage of correct classification is 7% higher. Finally, the MAE is 12% lower. These improvements justify the use of automated generated models instead of the usual subject matter expert manually designed model.

TABLE V
RESULTS OF SPECIALIST AND RANDOM FOREST MODELS

Metric	Specialist	RandomForest
Precision	0,86	0,86
Recall	0,17	0,32
Kappa	0,14	0,27
MAE	0,48	0,36
Correct	0,57	0,64

Another important outcome of this case study is the general rules one can obtain from the models. Figure 1 shows one of the 10 tree classifiers produced by Random Forest[13]. Observing it, we can gain additional knowledge about affiliation and corruption. As seen in Section II, the specialists model takes into consideration number of parties affiliated and motive of cancellation code. It assumes that higher numbers of parties affiliated indicates higher corruptibility. Besides, the model supposes that higher motive of cancellation codes also implies higher corruptibility.

Observing Figure 1 we can already confirm the specialists' assumption on motive of cancellation codes. Also, the specialists model uses the attribute number of parties affiliated, but the Random Forest model does not. As stated in Section IV, that attribute was shown as unrepresentative and discarded by feature selection. Specialists can explain this due to the fact that corrupts do not affiliate to a party for political reasons. They will stay at a given party if the benefits are worthy or change it if other party offers more profit. Therefore, number of affiliations can not predict corruptibility with accuracy.

Additionally, the Random Forest model shows assertions not previously addressed by the specialists model. It reveals that on average, higher numbers of days affiliated denotes lower corruptibility. This fact is non obvious evidence, but the specialists could relate to it. For them corrupts civil servants normally are not affiliated for long, holding to actual political reasons. Actually they want to show trusting traces to political figures only while acting corruptly.

[13]We applied Random Forest to the training set without discretization for simpler viewing purposes

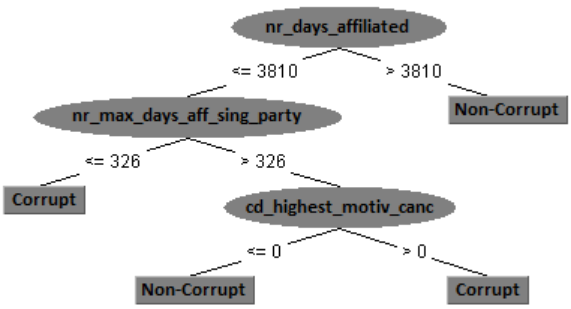

Fig. 1. Visualization of one of the trees of the model

VII. CONCLUSION

In this paper we presented a case study of machine learning applied to measure the risk of corruption of civil servants using political party affiliation data. We discussed the main business objectives for DIE specialists. They include predicting corruptibility with approximately 90% precision. Then, we applied a Chi-Square Hypothesis Test for Independence. Its conclusion is that the data supports the hypothesis that affiliation and corruption are dependent for civil servants. To construct datasets, we compared standardization and three different discretization methods implemented in Weka. That is, Multi-interval discretization, Equal-Frequency Binning, and Proportional k-Interval Discretization. Making use of the Weka Experiment Environment, we evaluated the datasets produced with four algorithms using 10-fold cross-validation. The classification algorithms used were Bayesian Networks, Random Forests, SVM, and Neural Networks with backpropagation. Additionally, to obtain approximately 90% precision, we wrapped all classifiers in Weka's MetaCost.

We evaluated the algorithms and datasets using the metrics precision, recall, kappa, Mean Absolute Error and percent correct. We obtained the results computed for all four algorithms and averaged each metric per dataset. After this, the dataset with Multi-interval discretization used on the first numeric attribute and Proportional k-Interval discretization on the second had the best metrics. We then compared the four classification algorithms only for the chosen dataset. Random Forest produced slightly better results than the others and, therefore, was selected as the final model.

Completing the evaluation, we compared the conservative specialists model from DIE to the Random Forest model using a test dataset. Random Forest model showed significant better results while keeping the same precision as the specialists model. The recall presented a substantial increase, indicating a 15% higher percentage of corrupts predicted. The kappa measure displayed a 13% gain. The percentage of correct classification was 7% higher. Finally, the Mean Absolute

Error was 12% lower. These improvements justify the use of automated generated models instead of the usual subject matter expert manually designed model.

Analyzing the final model, we were able to compare our Random Forest model with the previous model used by DIE. This specialists model supposed that higher motive of cancellation codes implies higher corruptibility. We managed to confirm this assumption observing the rules derived from the final model. Likewise, the specialists model assumed that higher numbers of parties affiliated indicates higher corruptibility. However, Random Forest model did not use that attribute. It was shown as unrepresentative and discarded by feature selection. On this assessment, we also gained additional knowledge about affiliation and corruption. Additionally, the Random Forest model showed non obvious assertions not previously addressed by the specialists model. It revealed that on average, higher numbers of days affiliated denotes lower corruptibility.

Therefore, the machine learning steps taken in this work were able to built a model with better results than the previous specialists model. Consequently, these improvements justify the use of automated generated models. This produced model was developed as part of a broader system and it will be used in the daily activities of the Department of Research and Strategic Information (DIE) at the Brazilian Office of the Comptroller General (CGU). As future work we intend to deploy the model into an application. The additional knowledge gained from this case study demonstrates the worth of the work. Thus, furthermore we will construct more corruptibility models with other civil servants data. The main purpose is to use all available data and have a bigger model to predict corruptibility of civil servants as efficiently as possible at DIE.

REFERENCES

[1] QEETHARA K. AL-SHAYEA and GHALEB A. EL-REFAE. Evaluating credit risk using artificial neural networks. 2011.

[2] Antônio Frederico Zancarano. A Corrupção Politico-Administrativa no Brasil. *Revista de Ciências Humanas da UNIPAR*, 3(10), 1995.

[3] Leo Breiman. Random Forests. Technical report, Technical Report 567, Department of Statistics, UC Berkeley, 1999. 31, 2001.

[4] Michael K. Buckland and Fredric C. Gey. The relationship between recall and precision. *JASIS*, 45(1):1219, 1994.

[5] Jean Carletta. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249254, 1996.

[6] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, page 161168. ACM, 2006.

[7] Jie Cheng and Russell Greiner. Learning bayesian belief network classifiers: Algorithms and system. In *Advances in Artificial Intelligence*, page 141151. Springer, 2001.

[8] Chih-Chung Chang and Chih-Jen Lin. LibSVM - A Library for Support Vector Machines, 2001. Available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm/.

[9] Gregory F. Cooper and Edward Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9(4):309347, 1992.

[10] Pedro Domingos. Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, page 155164. ACM, 1999.

[11] Usama Fayyad and Keki Irani. Multi-interval discretization of continuous-valued attributes for classification learning. 1993.

[12] Stephen R. Garner. Weka: The waikato environment for knowledge analysis. In *Proceedings of the New Zealand computer science research students conference*, pages 57–64. Citeseer, 1995.

[13] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[14] Mark A. Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

[15] Robert Hecht-Nielsen. Theory of the backpropagation neural network. In *Neural Networks, 1989. IJCNN., International Joint Conference on*, page 593605. IEEE, 1989.

[16] Finn V Jensen. *An introduction to Bayesian networks*, volume 210. UCL press London, 1996.

[17] José Matias Pereira. Reforma do Estado e transparência: estratégias de controle da corrupção no Brasil. Lisboa, Portugal, 2002.

[18] Adnan Khashman. Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. *Expert Systems with Applications*, 37(9):6233–6239, September 2010.

[19] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, volume 14, pages 1137–1145, 1995.

[20] Sotiris Kotsiantis and Dimitris Kanellopoulos. Discretization techniques: A recent survey. *GESTS International Transactions on Computer Science and Engineering*, 32(1):47–58, 2006.

[21] O. Marban, Gonzalo Mariscal, and Javier Segovia. A Data Mining & Knowledge Discovery Process Model. *Data Mining and Knowledge Discovery in Real Life Applications. IN-TECH*, 2009:8, 2009.

[22] Marcos Fernandes Gonçalves da Silva. A Economia Política da Corrupção, 2001.

[23] Michael Gordon and Manfred Kochen. Recall-precision trade-off: A derivation, 1988.

[24] Remco R. Bouckaert. Bayesian Network Classifiers in Weka, 2008. Available at http://www.cs.waikato.ac.nz/˜remco/weka.bn.pdf.

[25] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer, 2008.

[26] Herbert Lowe Stukart. *Ética e corrupção*. NBL Editora, 2003.

[27] The University of Waikato. Multilayer Perceptron Classifier in Weka, 1999. Available at http://weka.sourceforge.net/doc.stable/weka/classifiers/functions/MultilayerPerceptron.html.

[28] The University of Waikato. Weka: Waikato Environment for Knowledge Analysis, 2013. Available at http://www.cs.waikato.ac.nz/ml/weka/, version 3.6.10.

[29] University at Albany. Chi-Square Test for Independence, 2014. Available at http://omega.albany.edu:8008/mat108dir/chi2independence/chi2inm2h.html.

[30] Cort J. Willmott and Kenji Matsuura. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30(1):79, 2005.

[31] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37, January 2008.

[32] Ying Yang and Geoffrey I. Webb. Proportional k-interval discretization for naive-bayes classifiers. In *Machine learning: ECML 2001*, pages 564–575. Springer, 2001.