JURNAL BINA PRAJA
Journal of Home Affairs Governance

ARTICLE

# Transparency Prediction of Fraud Violations as an Anti-corruption Culture

## Experiment of Decision Tree

**Zico Karya Saputra Domas [1], Subagio [2], M. Rizkiawan [3]**

[1, 2, 3]Politeknik Keuangan Negara STAN

1401190046.zicoksd@gmail.com

**Abstract:** Several prominent reports have highlighted the unsatisfactory level of anti-corruption transparency for the private sector in Indonesia. Hence, the anti-corruption vision is still an aspect that deserves to be campaigned for to form an advanced and just civilization. This study aims to obtain a pattern of knowledge in predicting the level of transparency of disclosure of fraud violations based on a data mining approach. The classification function algorithm in this study is a decision tree which is then compared with other classification function algorithms, naive Bayes, and k-nn. The sample in this study is 141 companies combined in the construction, mining, and banking sectors, which are on the IDX for the 2019 period. As a result, the decision tree algorithm provides the second-best performance in predicting the level of corporate transparency, namely an accuracy of 70.92% and an AUC level of 0.740. Then in terms of different tests, the decision tree algorithm is in the same cluster as the algorithm with the best performance because the t-test results show no significant difference. Based on the pattern generated by the decision tree algorithm, the elements of opportunity, pressure, and arrogance are considered key factors in predicting the level of transparency of disclosure of fraud violations. One of them can be interpreted that a company that is supervised by a minimum of four independent commissioners means company tends to be predicted to be more daring in disclosing anti-corruption information in its annual report to the wider public. This study also recommends that every authorized institution in Indonesia can apply a data mining algorithm approach in utilizing the advantages of each agency's internal data volume to map anti-corruption cultural socialization strategies in private sector companies.

**Keywords:** corruption; anti-corruption; transparency; data mining; classification; decision tree

Check for updates

OPEN ACCESS

**Citation:** Domas, Z. K. S., Subagio, & Rizkiawan, M. (2022). Transparency Prediction of Fraud Violations as an Anti-corruption Culture: Experiment of Decision Tree. *Jurnal Bina Praja, 14*(2), 289–300. https://doi.org/10.21787/jbp.14.2022.289-300

**Received:** 20 May 2022

**Accepted:** 5 July 2022

**Published:** 21 September 2022

## 1. Introduction

Transparency International (2020) published less than a third of global countries that achieved a CPI (Corruption Perception Index) score above 50. This fact indicates that the risk of corruption is still a concern on a global scale. Regarding Indonesia, every year, Indonesia's CPI score tends to be consistently below the global country average score. Then, statistics on the official website of the Corruption Eradication Commission (accessed April 2021) show that the private sector has accumulated among the most dominant corruption perpetrators in Indonesia since 2004 as many as 314 perpetrators. Statistics on the official website of the Corruption Eradication Commission (accessed as of April 2021) also confirm the evidence that the mode of bribery dominates the practice of corruption. This means that the fact that corruption is proven is not limited to state officials but can also occur from the initiation of private sector individuals. Transparency International Indonesia (2017, 2018) and Salim (2018) also considered that anti-corruption transparency in private corporations in Indonesia is something that must also be addressed as a form of support for anti-corruption culture.

The construction infrastructure, mining-oil, and banking financial services sectors were chosen in this research for three main reasons. First, Syarif (2021) revealed that the construction and mining sectors were recorded to dominate the practice of payoffs in Indonesia. Second, the report of ACFE (2020) on a global scale revealed that 386 cases were recorded in the first position, far more than the 185 cases that occupy the second largest sector. Third, Pambudi's research (2020) observes that the construction sector and the financial sector in the period 2014 to 2018 are the sectors with the highest average Beneish score, which is a profile if it exceeds the score of -2.25 then it is considered more vulnerable to committing fraud. Then, the 2019 period was chosen because the year the election contestation in Indonesia was related to the neutrality policy of political donations by private sector corporations.

This study refers to the Hexagon Fraud theory that fraud is generally susceptible to being triggered by six elements: pressure, opportunity, rationality, competence, arrogance, and collusion. Hexagon Fraud Theory (Vousinas, 2019) is relevant to be used as the main reference in corruption research to be analyzed with a data mining approach. In substance, data mining can be defined as the process of determining usage patterns and trends from large data sets (Banerjee et al., 2018) so that the results of data processing from the data mining process can be used to improve future decision-making. Classification is an important function in data mining (Li et al., 2018; Nguyen et al., 2012) because it targets the category or class to be predicted accurately on each input data (Banerjee et al., 2018). There are various types of classifier algorithms with a data mining approach that is recommended in making predictions because it is proven to have a broad scope (Bujlow et al., 2012); some of them are decision tree, naive bayesian, and K-NN (Itoo et al., 2021; Wahono et al., 2014).

The decision tree's first algorithm is often used to solve classification and pattern recognition problems in machine learning (Shaheen et al., 2020). The main advantage is providing an illustrative way of representing all classification patterns that are easier to understand (Kirkos et al., 2007). In addition, the decision tree is also capable of non-linear processing data (Malini & Pushpa, 2017), so the decision tree algorithm is usually applied in the process of data mining and data analysis (Singh et al., 2013). For example, a study by Wirawan (2020) utilized the decision tree algorithm to predict the graduation status of UIN Syarif Hidayatullah students on time for 2012–2014.

The second algorithm, Naive Bayes or NB, is a classification method based on Bayes' theorem to calculate posterior probabilities. NB is also widely applied in classification problems because it performs highly (Chen et al., 2020; Khadafy & Wahono, 2015). However, this assumption can be invalid because of the dependencies between attributes vulnerable to occur (Kirkos et al., 2007). Sumanto et al. (2021) utilized the Naive Bayes algorithm in predicting the creditworthiness status of PT Pramtra Perumahan Queen Residence, while Niazi et al. (2019) predicted hotspot diagnosis for the solar module.

The third algorithm, K-Nearest Neighbor or k-nn, is a grouping technique that predicts the attributes of a data point based on its position relative to other data points (Bablani et al., 2018; Banerjee et al., 2018). According to Malini and Pushpa (2017), the advantage of applying k-nn is that there is no need for a certain requirement in making predictions on the classification. Tarjo and Herawati's research (2017) explained using k-nn to add value to make predictions more reliable. Soni et al. (2021) utilize the k-nn algorithm to predict credit card fraud detection in Europe. Kück and Freitag (2021) predict the level of customer demand, which is helpful for production level planning.

In Indonesia, research on corruption themes for the private sector with a data mining approach is still rarely done (Argandoña, 2005); one of them is the difficulty of data availability in proving corrupt practices in the private sector. This research is expected to be a pioneer for research on corruption themes in the private sector based on the usual algorithmic approach in data mining, namely decision tree, k-nn, and naive Bayes, to find patterns or insights related to the prediction of transparency in disclosure of fraud violations. Thus, this research is expected to contribute scientific references to the authorities as one of the considerations in determining the anti-corruption active socialization policy strategy.

## 2. Methods

This study observed as many as 141 companies in a combination of three sectors, namely the construction infrastructure, mining-oil, and banking financial services sectors listed on the Indonesia Stock Exchange (IDX). Data collection was carried out by utilizing secondary data in the form of financial statements and company annual reports for the 2019 period, which were obtained from the official website of the Indonesia Stock Exchange. The framework of this research refers to the CRISP-DM (Cross Industries Standard Process for Data Mining) model, which consists of six stages starting from the business understanding stage, data understanding, data preparation, modeling, evaluation, to the deployment stage (Larose, 2006) as shown in Figure 1.

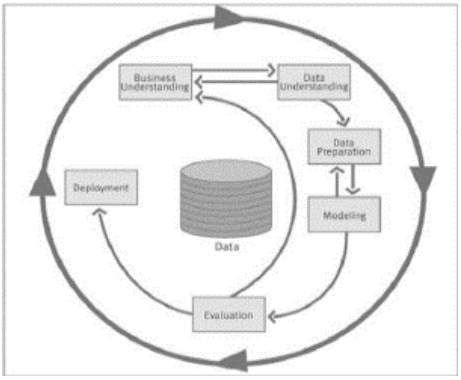


**Figure 1.** CRISP-DM

*Source: Larose (2006)*

From these six stages, it is enough to carry out this research until the evaluation stage because the main purpose of this research is to measure the performance level of the decision tree algorithm and find out the resulting classification pattern. In the business understanding phase, a basic understanding is carried out, which aims to find patterns in predictions regarding the extent of anti-corruption information disclosure in the private sector using a classification algorithm. In the data understanding phase, 141 secondary data samples were collected in the form of financial statements and annual reports of private companies in the 2019 period to collect most of the data attributes, as well as financial statements for the 2020 period to collect proxy attributes for the change of directors.

Tables and fields are selected as raw data materials in the data preparation phase. The attributes of this research data are divided into seven parts, namely the attributes of transparency in the disclosure of fraud violations, pressure, opportunity, rationality, competence, arrogance, to collusion attributes. Data label, transparency attribute of fraud violation disclosure, is measured according to the criteria presented in Table 1.

**Table 1.** Attributes of the Broad Transparency of Fraud Violation Disclosure

| Transparency Criteria | | |
|---|---|---|
| 1. Disclosure of the number of internal audit findings, including findings that have not been followed up by the corporation | | |
| 2. Disclosure of employee retention rates and details of human capital outflows | | |
| 3. Disclosure of the neutrality of political donations, either comply or otherwise to disclose these contributions publicly | | |
| 4. Disclosure of the number of violations of the code of ethics committed by employees | | |
| 5. Disclosure of the number of violations of administrative sanctions committed by the company | | |
| 6. Disclosure of the number of whistle-blower complaints, including follow-up on whistle-blower complaints | | |
| 7. Disclosure of the number of corruption violations committed by employees | | |
| 8. Disclosure of the implementation of a transparent, competitive, and objective procurement system for goods/services, especially the selection of vendors based on an anti-corruption reputation | | |
| 9. Disclosure of violations of tax obligations or commitments to comply with taxation | | |
| 10. Disclosure of the existence of external supervisors, not limited to the financial auditors of the Public Accounting Firm | | |
| 11. Disclosure of the amount or details of the settlement of legal disputes related to the court decision process | | |
| 12. Disclosure of organizational support service fees (e.g., bond rating service fees, legal service fees, accountant fees, and so on) | | |
| 13. Disclosure of the number of insider trading violations | | |
| 14. Disclosure of the number of complaints of customer dissatisfaction or customer satisfaction index | | |
| Scoring Stage I | For each of these 14 criteria, a score of 0 for adequate disclosure, a score of 1 if incomplete, and a score of 2 if not disclosed. | Accumulative Score = Total Score |
| Scoring Stage II | Calculate the average value of the accumulative score over the entire sample to be used as a reference value | |
| Scoring Stage III (final) | - Score 0 if a sample has a stage I value < a stage II value<br>- Score 1 if a sample has a stage I value > a stage II value | |

*Source: Argandoña (2005), Domas and Subagio (2022), Sopian et al. (2020), and Transparency International Indonesia (2016)*

Based on Table 1, companies that disclose information criteria can be quantified into whole numbers (both number 0 and number n), for example, "0 cases of whistleblower complaints, no violations of the code of ethics by employees this year, or 55 cases of whistleblower complaints" are given a score 0. While companies that disclose information in Table 1 criteria but cannot be quantified into whole numbers will be given a score of 1, companies that do not disclose information in Table 1 criteria will be given a score of 2.

Then, the cumulative score is compared with the average label value or data class for the entire sample. Then a dummy technique is used so that the dependent variable only has two categories. Where the score is 0 if the cumulative score of a sample is smaller than the average value of the entire sample, and a score of 1 if the cumulative score of a sample is greater than the average value of the entire sample, based on this measurement technique, it can be assessed that companies categorized as score one is perceived as relatively less transparent. On the other hand, companies categorized as 0 are perceived as more transparent because anti-corruption information dares to be disclosed to the public more thoroughly and widely.

Furthermore, the mapping of the entire data, as well as the measurement of the data attributes of pressure, opportunity, rationality, competence, arrogance, and collusion, are described in Table 2.

**Table 2.** Data Mapping

| Role | Name of Data Attribute | Data Description |
|---|---|---|
| Prediction label or target | The extent of transparency of disclosure of fraud violations | Binominal<br>0 : its annual report is more extensive in disclosing anti-corruption information.<br>1 : the annual report is more minimal in disclosing anti-corruption information. |
| Regular | Pressure | Binominal, modification of Lokanan and Sharma's research (2018),<br>0 : experienced a profit.<br>1 : incur a loss. |
| Regular | Opportunity | Integer, modification of research of Christian et al. (2019),<br>It is measured by proxy for the number of independent commissioners (the more independent commissioners, the lower the chance, and vice versa). |
| Regular | Rationality | Binominal, implementation of Abdullahi and Mansor research (2015) and Transparency International Indonesia (2017),<br>0 : banking business sector.<br>1 : non-banking business sector. |
| Regular | Competence | Binominal, a study from Novitasari and Chariri (2018) and Sasongko and Wijayantika (2019),<br>0 : there was no change of directors in the following year.<br>1 : there is a change of at least one board of directors in the following year. |
| Regular | Arrogance | Binominal, modification of Widodo and Fanani's research (2020) and Transparency International Indonesia (2017),<br>0 : the government acts as one of the owners of voting rights.<br>1 : the government does not act as the owner of the voting rights (purely private). |
| Regular | Collusion | Binominal, a modification of the research of Lo et al. (2010) and Fitri et al. (2019),<br>0 : disclose sales transaction information to a privileged party.<br>1 : do not disclose sales transaction information to privileged parties. |

Table 2 explains that the attributes of pressure, opportunity, rationality, competence, arrogance, and collusion will be used to predict the pattern of testing labels, namely the broad attribute of disclosure of fraud violations.

In the modeling phase, it is carried out using a rapid miner application that applies three classification algorithms: decision tree, naive bayesian, and k-nn. The decision tree algorithm in this study refers to the C4.5 method, which is the development of ID3 (Mienye et al., 2019; Pradana, 2018; Sitorus et al., 2021). Select the root of the attribute by calculating the gain value of all attributes, where the first root is the highest gain value. First, calculate the entropy value using the formula:

$$Entropy(S) = -\sum^{n} i = 1(-pi * Log2pi)$$

*Description:*
*S : Case Collection*
*n : Number of Partitions S*
*pi : Proportion of Si to S*

Next, calculate the gain value with the formula:

$$Gain\ (S, A) = Entropy\ (S) - \sum^{n} i = 1(|Si| / |S|) * Entropy(Si)$$

*Description:*
*S : Case Collection*
*A : Attribute*
*n : Number of Attribute Partitions A*
*|Si| : Number of Cases on Partition i*
*|S| : Number of Cases in S*

The naive Bayes algorithm is a classification algorithm based on Bayes' theorem. According to Banerjee et al. (2018), this method calculates the posterior probability (P(A|B)), namely the probability of the outcome (A) under certain conditions (B). Bayes' theorem calculates the posterior probability by relating it to the previous probability P(A), i.e., the probability of an outcome without knowledge of the condition

having an effect through the likelihood ratio P(B|A)/P(B). The Naive Bayes theorem is based on the assumption that each factor affects a result independently, so it is called naive.

$$P(A|B) = \{P(B|A)*P(A)\} / P(B)$$

The k-nn algorithm has three key elements: the label or object class, the distance between the objects, and the value of k, which is the number of nearest neighbors. To find unknown attributes or factors on a test data point, the Euclidean distance to each other data point must be determined (Banerjee et al., 2018; Bermúdez et al., 2020; Triguero et al., 2019).

$$Ed = \sqrt[2]{\{\triangle x + \triangle y + ... + \triangle n\}}$$
$$Input: D, \text{ the set of } k \text{ training objects, and test object } z = (x', y')$$
$$Process: \text{Compute } d(x', x), \text{ the distance between } z \text{ and every object, } (x, y) \in D$$
$$\text{Select } Dz \subseteq D, \text{ the set of } k \text{ closest training objects to } z.$$
$$Output: y' = \underset{v(xi, yi) \in Dz}{argmax} \sum I(v = yi)$$

In the evaluation phase, an analysis is carried out that interprets the pattern to predict which companies are classified as less transparent and which are classified as transparent. In this phase, the accuracy level between algorithms is also compared using the cross-validation technique to assess the feasibility of predictions. Cross-validation is a method that divides the dataset into two parts, where 90% of the part acts as training data while the other 10% acts as testing data. This process is repeated up to 10 times, so it is also known as ten-fold cross-validation. Researchers widely use this technique because it is proven to produce a more stable algorithm performance (Pradana, 2018). According to Itoo et al. (2021), four basic matrices in evaluating the performance of classification algorithms consist of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). Then, the level of accuracy is defined as the ratio of the total number of correctly predicted transactions, sensitivity is defined as the proportion of the positive observations correctly predicted as positive, and specificity is defined as how accurate the negative observations are correctly predicted as negative, so the Area Under Curve (AUC) describes the level of separability measurement that a model can distinguish between labels or classes.

$$Accuracy = (TP + TN) / (TP + FP + TN + FN)$$
$$AUC = 1/2*(Sensitivity + Specificity)$$

Furthermore, determining the Area Under Curve (AUC) performance of this study refers to the research (Gorunescu, 2011) with criteria:

1. 0.90–1.00 = Excellent Classification
2. 0.80–0.90 = Good Classification
3. 0.70–0.80 = Fair Classification
4. 0.60–0.70 = Poor Classification
5. 0.50–0.60 = Failure

## 3. Results and Discussion

### 3.1. Results of Comparison of Performance Levels Between Classification Function Algorithms

The design is carried out using the rapid miner version 9.9 application. The whole process of applying data mining to the three types of classification function algorithms is presented in Figure 2.

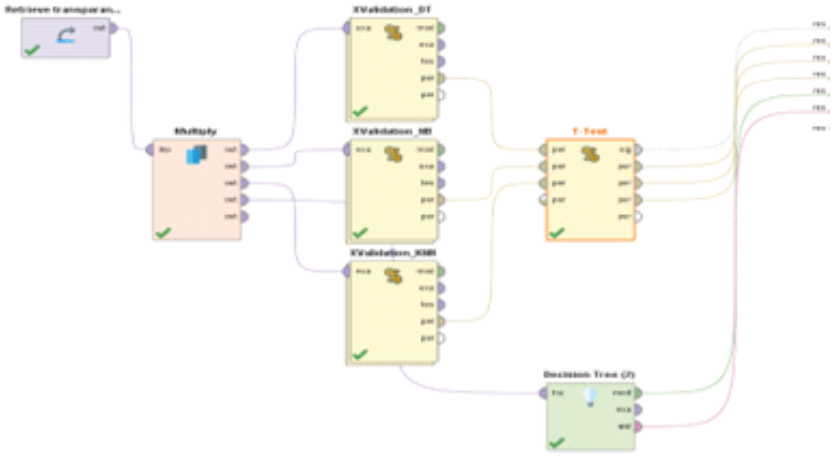

**Figure 2.** Evaluation Process of the Three Types of Classification Algorithms

Table 2 shows two main points in designing the cross-algorithm comparison process: applying the ten-fold stratified cross-validation process and the T-test. To validate the classification algorithm model, a ten-fold stratified cross-validation is used, namely the repetition of a combination of training data for ten repetitions of the learning process tested randomly (Perols, 2011; Singh et al., 2013). Then the pattern on the results of this training data is automatically applied to 10% of the testing data so that the performance evaluation of these three classification models can be measured objectively, as presented in Table 3.

**Table 3.** Recapitulation of Comparative Evaluation of the Three Types of Classification Algorithms

| Algorithm | Accuracy | AUC |
|---|---|---|
| Decision tree | 70.92%** | 0.740** |
| Naive Bayesian | 71.63%* | 0.827* |
| K-NN | 65.96% | 0.731 |

** algorithm with the best score;*
*** algorithm with the second best value*

Based on Table 3, the decision tree algorithm has the second-best performance level in terms of accuracy and AUC score. However, the decision tree cannot be concluded that it is in a worse feasibility cluster when compared to the Naive Bayesian algorithm, which briefly appears to have the best level of accuracy in these three types of classification algorithms. For this reason, it is necessary to carry out a different T-test to know the level of statistical differences, as presented in Table 4.

**Table 4.** T-test Result

| | Decision Tree | Naive Bayes | K-NN |
|---|---|---|---|
| Decision Tree | | 0.880 | 0.401 |

** significant above 0.05*

Table 4 shows that these three classification algorithms have no statistically significant difference because none of them has an alpha value of less than 0.05, so it can be concluded that the three main algorithms are in the appropriate eligibility cluster to be applied even though there are generally differences in accuracy and AUC scores. Then, the performance evaluation results on the AUC score for the decision trees algorithm are 0.74, so it can be concluded that there is a fairly good classifier category (Gorunescu, 2011). Thus, the decision tree algorithm can be concluded that it is also feasible to be applied as a method in predicting companies classified as less transparent and classified as transparent in disclosing anti-corruption information.

## 3.2. Discussion of the Transparency Pattern of Disclosure of Fraud Violations Using the Decision Tree Algorithm

The application of the decision tree algorithm shows the results of data processing as presented in Table 5.

**Table 5.** The Result of Attribute Weights on the Decision Tree

| Attribute | Opportunity | Pressure | Arrogance | Collusion | Rationality | Competence |
|---|---|---|---|---|---|---|
| Weight | 0.335 | 0.19 | 0.18 | 0.118 | 0.109 | 0.069 |

Table 5 shows that the elements of the hexagon model that can be a key factor in predicting the transparency of the disclosure of fraud violations are the elements of opportunity, pressure, and arrogance, where these three elements have the highest gain value of all attributes in this study. Furthermore, the pattern visualization of the results of the decision tree data processing is presented as shown in Figure 3.

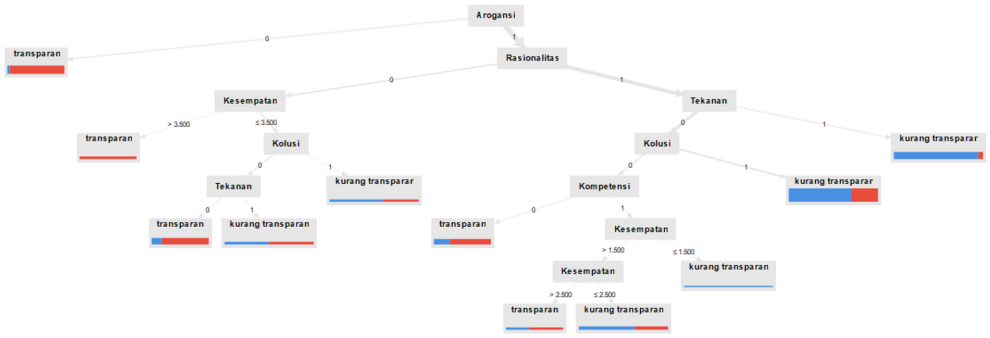


**Figure 3.** Pattern Visualization Based on Decision Tree Algorithm

Figure 3 shows that all elements in the hexagon model, ranging from pressure, opportunity, rationality, competence, and arrogance to collusion, in general, can be used to predict private companies' lack of transparency. If observed in-depth, three elements of the hexagon model can be used as key predictive factors, namely arrogance, opportunity, and pressure.

In layer one visualization of the decision tree diagram, the lack of arrogance can produce a predictive pattern of transparent companies disclosing anti-corruption information (shown by a dominant red symbol). This means that a company whose share ownership is owned by the government as the holder of influential voting rights, then the company will be predicted as a transparent company. Implicitly, this pattern is in line with the research of Chi et al. (2015), Matoussi and Gharbi (2011), and Wu et al. (2016), who view that companies with a purely private nature tend to be more limited in disclosing information where reports (Transparency International Indonesia, 2017) also explicitly assess the things that are in line.

Then in layer three, the lack of opportunity can produce a predictive pattern of transparent companies disclosing anti-corruption information (indicated by a dominant red symbol). This means that a company that has a minimum of four independent commissioners on the structure of the board of commissioners, then the company will be predicted as a transparent company. Implicitly, this pattern is in line with Kusumosari's research (2020) and Lokanan and Sharma (2018), who views that the lower the ratio of the independent board, the risk of fraud will also increase so it can also be interpreted that the higher the number of commissioners tends to increase the extent of disclosure of anti-corruption information. Furthermore, at layer six, the element of opportunity can also produce a predictive pattern of less transparent companies (shown by a dominant blue symbol). This means that a company that has only one independent member of the board of commissioners in the structure of the board of commissioners, then the company will be predicted as a less transparent company.

Still, at the third layer, the pressure element can also produce a predictive pattern for companies that are less transparent in disclosing anti-corruption information (shown by a very dominant blue symbol). This means that a company that is experiencing a loss in its financial statement report, then the company will be predicted as a less transparent company. Implicitly, Artiach et al. (2010) also view that companies with achieving stable financial targets tend to pay more attention to the non-operational vision of long-term sustainability than companies whose financial conditions are not yet stable. Then at layer five, the element of lack of pressure is quite capable of producing a predictive pattern of transparent companies (indicated by a red symbol which is quite dominant). This means that a company that is experiencing a profit in its financial statement report, then the company tends to be predicted as a transparent company.

In the fourth layer of visualization, the element of collusion is quite capable of producing a predictive pattern of companies that are less transparent in disclosing anti-corruption information (indicated by a blue symbol which is quite dominant). This means that companies have not explicitly disclosed the proportion of revenue from the special water relationship to overall turnover. The company tends to be predicted as a less transparent company. Implicitly, Fitri et al. (2019) also prove that sales transactions to a privileged party have a positive effect on fraud, so it can also be interpreted that the indecisiveness in disclosing transaction information originating from a special relationship can prevent the widespread disclosure of anti-corruption information.

In layer five of the decision tree diagram visualization, the competency element is quite capable of producing a predictive pattern of transparent companies disclosing anti-corruption information (indicated by a red symbol which is quite dominant). This means that a company whose General Meeting of Shareholders or its GMS decides not to change its board of directors in the following year (at least the GMS still maintains the board of directors in the normal cycle of a five-year term), then the company tends to be predicted as a transparent company in the regime the directors are still in office. Implicitly, Puspitha and Yasa (2018) and Sasongko and Wijayantika (2019) also views that the change of directors has a significant positive effect on fraud, so it can also be interpreted that the absence of a change of directors in the following year could be one of the indicators that a company tends to be transparent.

Thus, an analysis of the decision tree algorithm pattern recommends that a more massive campaign is needed regarding the disclosure of anti-corruption information that needs to be published by private companies. This can be seen in the attitude of private companies supervised by a minimum of four independent commissioners, so these companies tend to be more transparent in disclosing anti-corruption information to the public at large. Currently, the regulation regulates the number of independent commissioners of private companies based on the ratio of one-third of the board members, not based on the number of members of the board of commissioners that effective socialization techniques are needed so that companies with less than four independent commissioners also have an awareness of the importance of an anti-corruption information disclosure culture.

## 4. Conclusion

The results of the study conclude two main points. First, the accuracy of the decision tree algorithm is proven to provide the second-best accuracy performance. However, based on the statistical difference test aspect, the decision tree algorithm is proven to have no significant difference from the Naive Bayes algorithm, which provides the best performance so that it can be concluded that the decision tree algorithm is also feasible to apply. Second, based on the pattern generated by the decision tree algorithm, the elements of opportunity, pressure, and arrogance can be concluded as key factors related to the prediction of the wide transparency of disclosure of fraud violations of a company.

Then, the study also stated some limitations. First, the number of observations of the research sample is not too large. Second, this study does not at all discuss the occurrence of corrupt practices in the private sector but only focuses on the extent of disclosure of anti-corruption information due to the unavailability of data in measuring indications of corruption. Third, at the data input and training data stages, this study measures the broad attribute of disclosure with non-mandatory criteria so that companies still do not have an obligation to disclose the information.

This study also recommends several suggestions for further research on anti-corruption in data mining. First, further research can increase the number of observations close to the general election contestation agenda, for example, the 2017–2019 observation period or the 2022–2024 observation period in the future. Second, further research can add some data attributes that should be suspected of having relevance to the transparency of anti-corruption information disclosure, for example, the size of foreign party share ownership, the ratio of the incentive costs of the board of directors and commissioners to the salary costs of all employees, the increase in stock prices around the moment after announcements of financial statements, as well as various other data attributes.

Finally, the author tries to provide input to the competent authorities, such as the Corruption Eradication Commission, the Financial Services Authority, BI, the Capital Market Supervisory Agency, the Attorney General's Office, the Police, and all other regulators, so that they can carry out routine mapping every year based on data mining by taking advantage of the advantages of data mining attribute the internal data of each agency to disseminate anti-corruption active culture to private corporations.

Acknowledgment

We would like to express our deepest gratitude to all parties involved in this research process.

References

Abdullahi, R., & Mansor, N. (2015). Fraud Triangle Theory and Fraud Diamond Theory. Understanding the Convergent and Divergent for Future Research. *International Journal of Academic Research in Accounting, Finance and Management Sciences, 5*(4), 30–37. https://doi.org/10.6007/IJARAFMS/v5-i4/1823

ACFE. (2020). *Report to the Nations: 2020 Global Fraud Study on Occupational Fraud and Abuse*. https://legacy.acfe.com/report-to-the-nations/2020/

Argandoña, A. (2005). Private-to-Private Corruption. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.685864

Artiach, T., Lee, D., Nelson, D., & Walker, J. (2010). The Determinants of Corporate Sustainability Performance. *Accounting & Finance, 50*(1), 31–51. https://doi.org/10.1111/j.1467-629X.2009.00315.x

Bablani, A., Edla, D. R., & Dodia, S. (2018). Classification of EEG Data Using K-nearest Neighbor Approach for Concealed Information Test. *Procedia Computer Science, 143*, 242–249. https://doi.org/10.1016/j.procs.2018.10.392

Banerjee, R., Bourla, G., Chen, S., Kashyap, M., & Purohit, S. (2018). Comparative Analysis of Machine Learning Algorithms through Credit Card Fraud Detection. *2018 IEEE MIT Undergraduate Research Technology Conference (URTC)*, 1–4. https://doi.org/10.1109/URTC45901.2018.9244782

Bermúdez, J. R., López-Estrada, F. R., Besançon, G., Torres, L., & Santos-Ruiz, I. (2020). Leak-Diagnosis Approach for Water Distribution Networks based on a k-NN Classification Algorithm. *IFAC-PapersOnLine, 53*(2), 16651–16656. https://doi.org/10.1016/j.ifacol.2020.12.795

Bujlow, T., Riaz, T., & Pedersen, J. M. (2012). Classification of HTTP Traffic Based on C5.0 Machine Learning Algorithm. *2012 IEEE Symposium on Computers and Communications (ISCC)*, 000882–000887. https://doi.org/10.1109/ISCC.2012.6249413

Chen, S., Webb, G. I., Liu, L., & Ma, X. (2020). A Novel Selective Naïve Bayes Algorithm. *Knowledge-Based Systems, 192*(xxxx), 105361. https://doi.org/10.1016/j.knosys.2019.105361

Chi, C. W., Hung, K., Cheng, H. W., & Tien Lieu, P. (2015). Family Firms and Earnings Management in Taiwan: Influence of Corporate Governance. *International Review of Economics & Finance, 36*, 88–98. https://doi.org/10.1016/j.iref.2014.11.009

Christian, N., Basri, Y. Z., & Arafah, W. (2019). Analysis of Fraud Pentagon to Detecting Corporate Fraud in Indonesia. *International Journal of Economics, Business and Management Research, 3*(08), 1–13. https://www.researchgate.net/publication/335060762

Domas, Z. K. S., & Subagio. (2022). Fraud Hexagon Analysis on the Less-Transparent Anti-corruption Disclosures. *3rd National Conference Accounting and Fraud Auditing*.

Fitri, F., Syukur, M., & Justisa, G. (2019). Do the Fraud Triangle Components Motivate Fraud in Indonesia? *Australasian Accounting, Business and Finance Journal, 13*(4), 63–72. https://doi.org/10.14453/aabfj.v13i4.5

Gorunescu, F. (2011). *Data Mining* (12th ed., Vol. 12). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-19721-5

Itoo, F., Meenakshi, & Singh, S. (2021). Comparison and Analysis of Logistic Regression, Naïve Bayes and KNN Machine Learning Algorithms for Credit Card Fraud Detection. *International Journal of Information Technology, 13*(4), 1503–1511. https://doi.org/10.1007/s41870-020-00430-y

Khadafy, A. R., & Wahono, R. S. (2015). Penerapan Naive Bayes untuk Mengurangi Data Noise pada Klasifikasi Multi Kelas dengan Decision Tree. *Journal of Intelligent Systems, 1*(2), 136–142. https://journal.ilmukomputer.org/index.php?journal=jis&page=article&op=view&path%5B%5D=78

Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data Mining Techniques for the Detection of Fraudulent Financial Statements. *Expert Systems with Applications, 32*(4), 995–1003. https://doi.org/10.1016/j.eswa.2006.02.016

Kück, M., & Freitag, M. (2021). Forecasting of Customer Demands for Production Planning by Local k-nearest Neighbor Models. *International Journal of Production Economics, 231*, 107837. https://doi.org/10.1016/j.ijpe.2020.107837

Kusumosari, L. (2020). *Analisis Kecurangan Laporan Keuangan Melalui Fraud Hexagon pada Perusahaan Manufaktur yang Terdaftar di Bursa Efek Indonesia Tahun 2014–2018* [Universitas Negeri Semarang]. http://lib.unnes.ac.id/40840/

Larose, D. T. (2006). *Data Mining Methods and Models*. John Wiley & Sons, Inc.

Li, T., Li, J., Liu, Z., Li, P., & Jia, C. (2018). Differentially Private Naive Bayes Learning Over Multiple Data Sources. *Information Sciences, 444*, 89–104. https://doi.org/10.1016/j.ins.2018.02.056

Lo, A. W. Y., Wong, R. M. K., & Firth, M. (2010). Tax, Financial Reporting, and Tunneling Incentives for Income Shifting: An Empirical Analysis of the Transfer Pricing Behavior of Chinese-Listed Companies. *Journal of the American Taxation Association, 32*(2), 1–26. https://doi.org/10.2308/jata.2010.32.2.1

Lokanan, M., & Sharma, S. (2018). A Fraud Triangle Analysis of the Libor Fraud. *Journal of Forensic & Investigative Accounting, 10*(2), 187–212. https://doi.org/10.25316/IR-1573

Malini, N., & Pushpa, M. (2017). Analysis on Credit Card Fraud Identification Techniques Based on KNN and Outlier Detection. *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, 255–258. https://doi.org/10.1109/AEEICB.2017.7972424

Matoussi, H., & Gharbi, I. (2011). *Board Independence and Corporate Fraud: The Case of Tunisian Firms* (No. 620; Economic Research Forum). https://ideas.repec.org/p/erg/wpaper/620.html

Mienye, I. D., Sun, Y., & Wang, Z. (2019). Prediction Performance of Improved Decision Tree-Based Algorithms: A Review. *Procedia Manufacturing, 35*, 698–703. https://doi.org/10.1016/j.promfg.2019.06.011

Nguyen, L. T. T., Vo, B., Hong, T.-P., & Thanh, H. C. (2012). Classification Based on Association Rules: A Lattice-Based Approach. *Expert Systems with Applications, 39*(13), 11357–11366. https://doi.org/10.1016/j.eswa.2012.03.036

Niazi, K. A. K., Akhtar, W., Khan, H. A., Yang, Y., & Athar, S. (2019). Hotspot Diagnosis for Solar Photovoltaic Modules Using a Naive Bayes Classifier. *Solar Energy, 190*(July), 34–43. https://doi.org/10.1016/j.solener.2019.07.063

Novitasari, A. R., & Chariri, A. (2018). Analisis Faktor-Faktor yang Mempengaruhi Financial Statement Fraud dalam Perspektif Fraud Pentagon. *Diponegoro Journal of Accounting, 7*(4), 1–15. https://ejournal3.undip.ac.id/index.php/accounting/article/view/25572

Pambudi, A. S. (2020). *Analisis Pendeteksian Financial Statement Fraud Menggunakan Beneish M-Score Model dan Data Mining*.

Perols, J. (2011). Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms. *AUDITING: A Journal of Practice & Theory, 30*(2), 19–50. https://doi.org/10.2308/ajpt-50009

Pradana, E. (2018). *Analisis Penerapan Adaptive Boosting (Adaboost) dalam Meningkatkan Performasi Algoritma C4.5*. Sekolah Tinggi Teknologi Pelita Bangsa.

Puspitha, M. Y., & Yasa, G. W. (2018). Fraud Pentagon Analysis in Detecting Fraudulent Financial Reporting (Study on Indonesian Capital Market). *International Journal of Sciences: Basic and Applied Research, 42*(5), 93–109. https://www.gssrr.org/index.php/JournalOfBasicAndApplied/article/view/9628

Salim, F. (2018). *Anti Corruption Measurements in Business: Transparency in Corporate Reporting (TRAC) and Beyond*.

Sasongko, N., & Wijayantika, S. F. (2019). Faktor Resiko Fraud terhadap Pelaksanaan Fraudulent Financial Reporting (Berdasarkan Pendekatan Crown's Fraud Pentagon Theory). *Riset Akuntansi dan Keuangan Indonesia, 4*(1), 67–76. https://doi.org/10.23917/reaksi.v4i1.7809

Shaheen, M., Zafar, T., & Ali Khan, S. (2020). Decision Tree Classification: Ranking Journals Using IGIDI. *Journal of Information Science, 46*(3), 325–339. https://doi.org/10.1177/0165551519837176

Singh, D., Choudhary, N., & Samota, J. (2013). Analysis of Data Mining Classification With Decision Tree Technique. *Global Journal of Computer Science and Technology Software & Data Engineering, 13*(13), 1–5.

Sitorus, Z., Saputra S., K., & Sulistianingsih, I. (2021). C4.5 Algorithm Modeling for Decision Tree Classification Process Against Status UKM. *Proceedings of the 3rd International Conference of Computer, Environment, Agriculture, Social Science, Health Science, Engineering and Technology - ICEST*. https://doi.org/10.5220/0010046105360540

Soni, K. B., Chopade, M., & Vaghela, R. (2021). Credit Card Fraud Detection Using Machine Learning Approach. *Applied Information System and Management (AISM), 4*(2), 71–76. https://doi.org/10.15408/aism.v4i2.20570

Sopian, Pratama, R. S., & Subagio. (2020). The Indonesia's Anti Corruption Strategies: A Gap Analysis to the UNCAC'S Preventive Measurarements. *Test Engineering and Management, 83*, 12087–12108. http://www.testmagzine.biz/index.php/testmagzine/article/view/5824

Sumanto, S., Marita, L. S., Mazia, L., & Ratnasari, T. W. (2021). Analisis Kelayakan Kredit Rumah Menggunakan Metode Naïve Bayes untuk Mengurangi Kredit Macet. *Applied Information System and Management (AISM), 4*(1), 17–22. https://doi.org/10.15408/aism.v4i1.20274

Syarif, L. M. (2021). *Memaknai CPI 2020 yang Menurun*. Kemitraan Partnership.

Tarjo, T., & Herawati, N. (2017). The Comparison of Two Data Mining Method to Detect Financial Fraud in Indonesia. *GATR Accounting and Finance Review, 2*(1), 01–08. https://doi.org/10.35609/afr.2017.2.1(1)

Transparency International. (2020). *Corruption Perceptions Index*. https://www.transparency.org/en/cpi/2020

Transparency International Indonesia. (2016). *Transparency in Corporate Reporting*.

Transparency International Indonesia. (2017). *Transparency in Corporate Reporting: Perusahaan Terbesar Indonesia*. https://ti.or.id/transparency-in-corporate-reporting/

Transparency International Indonesia. (2018). *Transparency in Corporate Reporting*.

Triguero, I., García-Gil, D., Maillo, J., Luengo, J., García, S., & Herrera, F. (2019). Transforming Big Data Into Smart Data: An Insight on the Use of the K-nearest Neighbors Algorithm to Obtain Quality Data. *WIREs Data Mining and Knowledge Discovery, 9*(2), 1–24. https://doi.org/10.1002/widm.1289

Vousinas, G. L. (2019). Advancing Theory of Fraud: The S.C.O.R.E. Model. *Journal of Financial Crime, 26*(1), 372–381. https://doi.org/10.1108/JFC-12-2017-0128

Wahono, R. S., Herman, N. S., & Ahmad, S. (2014). Neural Network Parameter Optimization Based on Genetic Algorithm for Software Defect Prediction. *Advanced Science Letters, 20*(10), 1951–1955. https://doi.org/10.1166/asl.2014.5641

Widodo, A., & Fanani, Z. (2020). Military Background, Political Connection, Audit Quality and Earning Quality. *Jurnal Akuntansi, 24*(1), 84–99. https://doi.org/10.24912/ja.v24i1.658

Wirawan, C. (2020). Teknik Data Mining Menggunakan Algoritma Decision Tree C4.5 untuk Memprediksi Tingkat Kelulusan Tepat Waktu. *Applied Information System and Management (AISM), 3*(1), 47–52. https://doi.org/10.15408/aism.v3i1.13033

Wu, W., Johan, S. A., & Rui, O. M. (2016). Institutional Investors, Political Connections, and the Incidence of Regulatory Enforcement Against Corporate Fraud. *Journal of Business Ethics, 134*(4), 709–726. https://doi.org/10.1007/s10551-014-2392-4