BOHR International Journal of Advances in Management Research
2022, Vol. 1, No. 1, pp. 7–14
https://doi.org/10.54646/bijamr.002
www.bohrpub.com

Fraud Detection in E-Commerce Using Machine Learning

Samrat Ray

ISMS Sankalp Business School, Pune, India
E-mail: samratray@rocketmail.com

Abstract. A rise in transactions is being caused by an increase in online customers. We observe that the prevalence of misrepresentation in online transactions is also increasing. Device learning will become more widely used to avoid misrepresentation in online commerce. The goal of this investigation is to identify the best device learning calculation using decision trees, naive Bayes, random forests, and neural networks. The realities to be utilized have not yet been modified. Engineered minority over-testing stability information is made utilizing the strategy framework. The precision of the brain not entirely settled by the disarray network appraisal is 96%, trailed by naive Bayes (95%), random forest (95%), and decision tree (92%).

Keywords: AI, fraud identification, algorithms, matrix, web-based.

INTRODUCTION

According to research on web clients in Indonesia published in the October 2019 issue of *Free Marketeers Magazine*, the country's 132 million web clients in 2019 alone represented an increase from the 142.3 million clients depicted in Figure 1 from the previous year. There were far too many people using the web-based system and conducting web-based transactions during COVID-19, but where there are inventions, there are also many problems. There are numerous methods for growing an e-commerce business [1, 3].

Based on information from many datasets, it is predicted that by 2022, the amount of retail online business transactions in Indonesia will expand from its current position to 134.6% of US$ 15.3 million, or almost 217 trillion. Rapid technical advancements that make it easier for customers to shop are supporting this growth.

Numerous e-commerce transactions present a variety of challenges and new problems, particularly the e-commerce fraud shown in Figure 2. The number of Internet business-related scams has also continuously climbed since around 1993. According to a 2013 survey, 5.65 pennies out of every $100 in web-based business exchanges' total turnover were overstated. More than 70 trillion dollars will have been stolen by 2019 [4, 5]. Fraud identification is one method to cut down on misrepresentation in online transactions.

The technology for detecting credit card fraud has advanced quickly, moving from machine learning to deep learning [6]. But regrettably, the amount of research on e-commerce fraud detection is still tiny, and it is only now focused on identifying the traits or qualities [7] that will be used to identify whether an e-commerce transaction is fraudulent or not.

The datasets used in this study had a combined 140,130 insights, 11,150 data points, and a 0.093 rate for extortion measures. Datasets with very small proportions produce lopsided information. When compared to minority data, irregularity data produces more accurate results that are more heavily weighted toward bigger portions of insights. The categorization of mainly non-extortion as opposed to misrepresentation produced more remarkable findings from the dataset studied. Using the destroyed (synthetic minority oversampling) strategy to adapt to data irregularities worsens the class outcomes [8, 9].

This study aims to identify the most effective model for identifying deception in an online transaction. Extraction is included in recent research on where to find fraud in e-commerce [10, 11]. This paper concentrates on fraud detection in e-commerce. It concentrates on the use of datasets from Kaggle, upgrade grouping AI, the use of SMOTE, and SMOTE utilization taking care of unbalanced records. After the use of SMOTE, the dataset will be trained on the use of contraption dominating. Decision trees, naive Bayes,

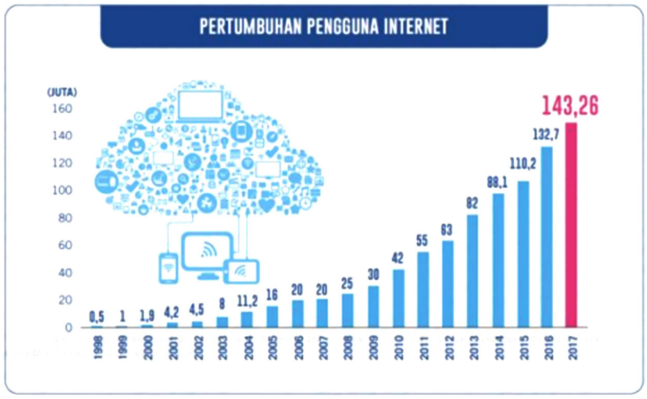

Figure 1. Growth of internet users [2].

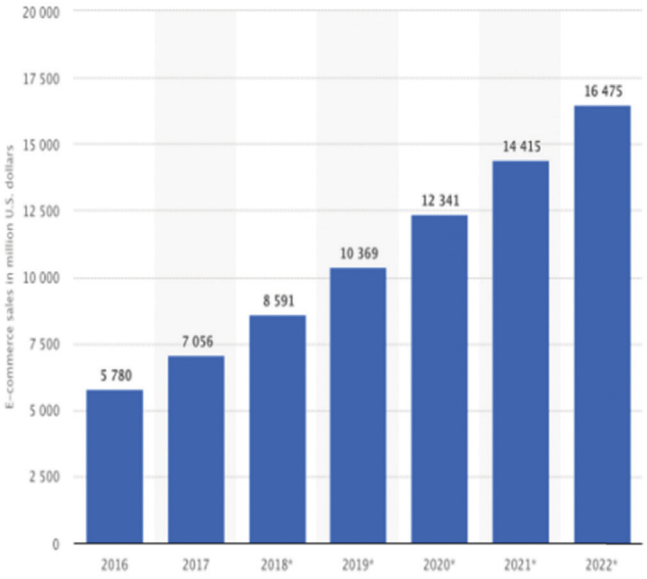

Figure 2. Sales of e-commerce, statista.com [4].

irregular woods, and brain network machine examinations are used to determine the exactness, correctness, and consideration of F1 -rating, and G-mean.

MATERIALS AND METHODS

Using computations from decision tree, naive Bayes, random forest, and neural networks, this study investigates extortion and non-misrepresentation in online business transactions. The cycle has ended, as seen in Figure 3.

The dataset's component determination process serves as the starting point for the collection framework. Change, normalization, and scale of the characteristics are employed to express the relationship so that they may be used for the game plan once the SMOTE procedure has finished the depiction cycle. After that, there is no permanent setup, which is accomplished by preprocessing data using principal component analysis (PCA). The importance of destroyed is essential for balancing faulty data.

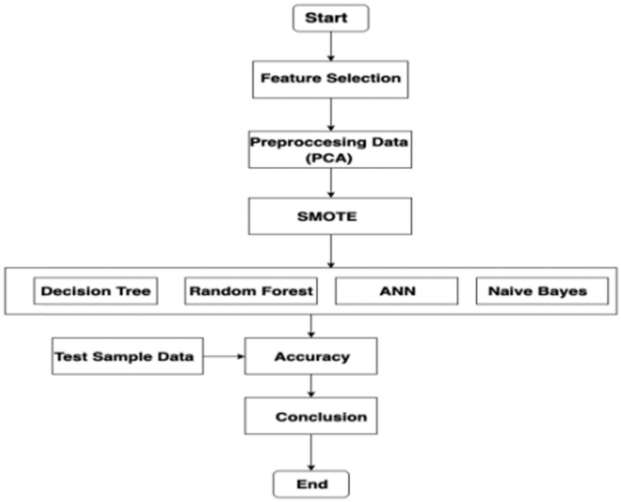

Figure 3. Research steps.

Since misrepresentation situations are typically about 2%, the SMOTE technique is useful for reducing the greater portion of the class in the dataset and addressing information discomfort issues. The implications of the SMOTE dataset exchange misrepresentation cycle will be altered if the bigger part class causes the grouping to be more coordinated to the larger part class such that the predictions of the order are not accurate [12, 15].

In the characterization cycle, AI utilized a decision tree, irregular woodland, counterfeit brain organization, and credulous Bayes. The web-based firm uses these AI calculations to take into account and then locate the exchange dataset's greatest accuracy outcomes.

Preprocessing Data

New elements that will be employed in the AI computation cycle are subject to preprocessing, which removes, modifies, scales, and standardizes them. Unreliable data are converted into reliable data through preprocessing. The highlights of the PCA preprocessing in this study include extraction, modification, normalization, and scaling.

In order to isolate highlights from information at a high-layered scale, PCA is a direct modification that is typically applied in information pressure. Furthermore, PCA can reduce complex information to more modest aspects to show obscure parts and improve the construction of information. PCA computations include computations of covariance frameworks to limit decreases and boost change.

Decision Tree

Decision trees are valuable for investigating extortion information and finding secret connections between various likely factors and an objective variable. The decision tree [20] consolidates misrepresentation information investigation and displaying, so it is generally excellent as the most important phase in the displaying system in any

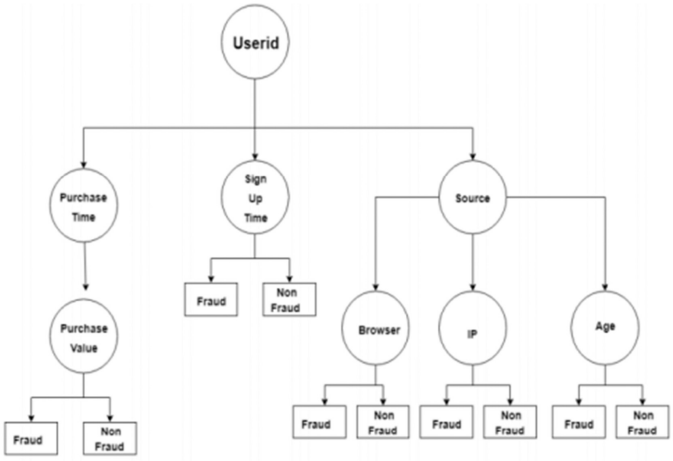

Figure 4. Architecture of decision trees.

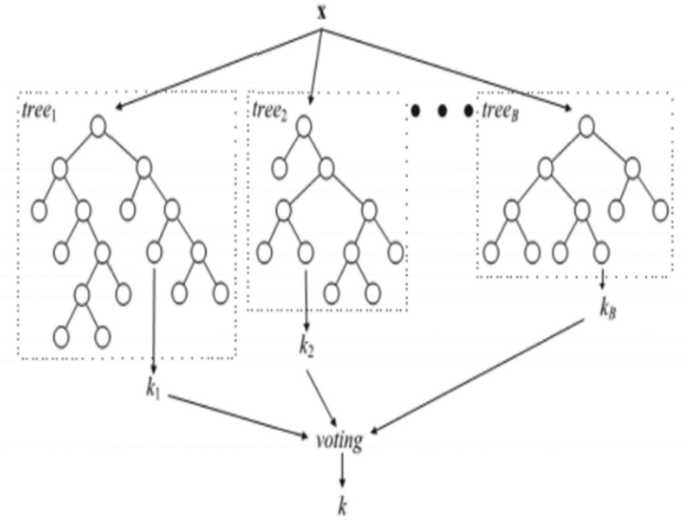

Figure 5. Architecture of random forest.

event, when utilized as the last model of a few different procedures [16, 18].

Decision trees are excellent for ordering computations and are a type of controlled learning calculation. The decision tree organizes the dataset into a few increasing segments in line with choice principles by emphasizing the connection between information and result credits.

- Root node: This addresses the whole population or test, and this is additionally separated into at least two.
- Parting: This is the most common way of separating a hub into two or, on the other hand, more sub-hubs.
- When a sub-center point splits into a few smaller sub-center points, the decision node is activated.
- Leaf/Terminal node: Unspecified center points are called leaf or terminal center points.
- Pruning: When a decision's sub-center point is removed.
- Branch/Sub-Tree: Subdivisions of all trees are called branches or sub-trees.
- Parent and child node: A center point that is divided into sub-centers [19].

As shown in Figure 4, the fraud detection employs a decision tree with a root hub, inner hub, and leaf hub.

Naive Bayes

Naive Bayes predicts open doors because of experience [23]. It involves the estimation equation as beneath.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \quad (1)$$

Where

B: cope with the statistics with obscure training
A: specific splendor is the statistical hypothesis
P(A | B): speculation possibility given conditions (returned opportunity)
P (A): probability of the hypothesis (prior possibility)
P(B | A): Probability—taking into account the speculative conditions
P(B): Possibility A

The aforementioned equation can be used to access both fraudulent and lawful transactions.

Random Forest

When a lot of data is required, the random forest (RF) algorithm is used. The classification and regression tree (truck) system has evolved into RF by including the bootstrap hoarding (firing) method and unexpected element determination architecture. In Figure 5, the RF is displayed.

A model called a "random forest" is made up of all intelligent group action fraud trees. The maximum depth call trees in the e-commerce fraud detection system depends on RF and employs a random vector distribution that is the same across all trees. The decision tree produces the top categories, and they are used to select the classification method's category.

Neural Network

A neural network system with nodes connected, such as the architectural neural network seen in Figure 6, is applied in the human body as part of the algorithm neural network artificial intelligence technique.

Before preparing, there were 11 information layers. After preprocessing, there were 17 information layers. The secret layer was decided on the neural network by hereditary calculations on the secret layer notwithstanding the number of info layers [18]. This forecasting procedure uses the GA-NN [19] algorithm, which is as follows:

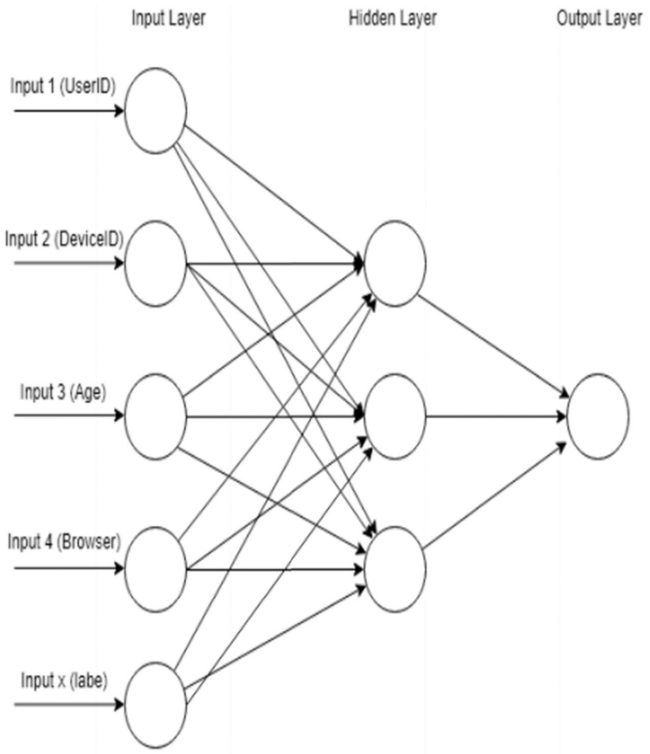

Figure 6. Architecture of neural network.

These predictions are as follows:

- Initialization count is zero, fitness is one, and there are no cycles.
- Early stages of population growth. Each consecutive gene sequence that makes up chromosome codes for the input.
- Suitable network architecture.
- Give weights.
- Train your backpropagation skills. examinations of fitness metrics and accumulated errors. then assessed according to the worth of fitness. If the current value of fitness is greater than the prior value of fitness.
- Count = count +1.
- Selection: A roulette wheel mechanism is used to choose the two mains. Crossover, mutation, and reproduction are examples of genetic operations that create new capabilities.
- Assuming the number of cycles rises to the count, return to number 4.
- Network guidance with picked attributes.
- Look at execution utilizing test results.

Confusion Matrix

A technique that may be used to assess categorization performance is the confusion matrix. A dataset with just two different class categories is shown in Table 1 [20].

False Positive and False Negative count the number of positively and negatively categorized objects, respectively, whereas True Positive and True Negative count the number of positively and negatively classed objects, respectively (FN).

Table 1. Confusion matrix.

Class	Predictive Positive	Predictive Negative
Actual Positive	TP	TN
Actual Negative	FP	FN

The most popular metric for assessing classification abilities is accuracy, but if you operate in an unequal setting, this assessment is flawed since the minority class will only make up a very small portion of the accuracy metric.

The F-1 score, G-mean, and recall evaluation criteria are advised. The G-mean list is utilized to quantify by and large execution (in general arrangement execution), though the F-1 score is utilized to evaluate how minority classes are ordered in imbalanced classes.

Recall, precision, F-1 score, and G-mean categorization ability were examined in this study.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TN} + \text{FP}} \tag{3}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{G-Mean} = \sqrt{\text{TP} - \text{TN}} \tag{5}$$

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

RESULTS

Dataset

This study utilizes a Kaggle-obtained online business fraud dataset. The dataset has 151,112 records. Of these, 14,151 records are classified as deceitful movement, and the extent of false action information is 0.094. The extortion exchange dataset results in 152,122 full records, 14,152 records classified as misrepresentation, and a misrepresentation information fraction of 0.094, as shown in Figures 7 and 8. SMOTE reduces class lopsidedness by blending information.

The image has been oversampled.

Decision Trees

Data that have undergone preprocessing are prepared for the experimental phase using the decision tree model. Subsequent to preprocessing, the information will be oversampled before an order utilizing a decision tree is performed. Moreover, the decision tree will likewise be performed using information that has not been oversampled. The findings of these two experiments will be utilized to

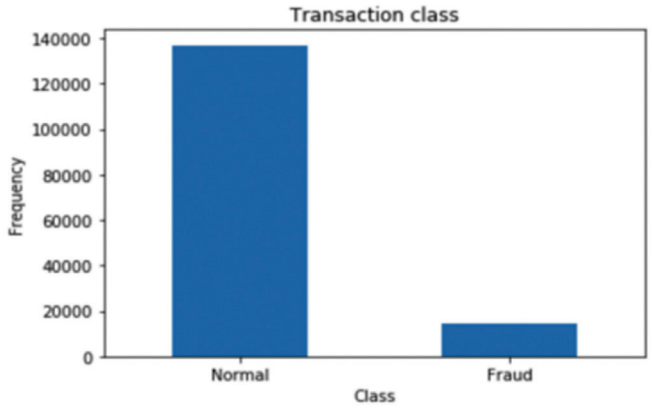

Figure 7. Ratio fraud.

Figure 8. Ratio fraud after over sampling.

Table 2. Confusion matrix decision tree without SMOTE.

Class	Predictive Positive	Predictive Negative
Actual Positive	38782	38782
Actual Negative	1746	2595

Table 3. Confusion matrix decision tree with SMOTE.

Class	Predictive Positive	Predictive Negative
Actual Positive	38651	2342
Actual Negative	1724	2617

analyze decision trees and demonstrate the classification outcomes utilizing the SMOTE oversampling technique.

The decision-making process without SMOTE precision is 53.2%, F1-score is 56.8%, accuracy is 90%, recall is 57.7%, and G-mean is 76.3%. Results from the confusion matrix decision tree without SMOTE are shown in Table 2.

Decision tree that produces SMOTE recall is 61.4%, precision is 90.5%, F1-score is 90.2%, and G-mean is 72.2%. Accuracy is 90%. Results from the confusion matrix decision tree with SMOTE are shown in Table 3.

Naive Bayes

Getting ready information that has recently been handled during preprocessing is the manner in which the naive Bayes model test is done. Following preprocessing, the information will be oversampled utilizing the two sorts of information: information that has been oversampled and

Table 4. Confusion matrix Naïve Bayes without SMOTE.

Class	Predictive Positive	Predictive Negative
Actual Positive	40764	229
Actual Negative	1993	2348

Table 5. Confusion matrix Naïve Bayes with SMOTE.

Class	Predictive Positive	Predictive Negative
Actual Positive	40760	233
Actual Negative	1988	2353

Table 6. Confusion matrix random forest without SMOTE.

Class	Predictive Positive	Predictive Negative
Actual Positive	40881	112
Actual Negative	1954	2387

information that has not, as well as naive Bayes arrangement will be finished utilizing the two sorts of information. Through a side-by-side comparison of naive Bayes and the oversampling approach, the findings of these two research methods will be utilized to demonstrate the grouping outcomes.

Without SMOTE generation, naive Bayes recall is 52.1%, precision is 90.2%, F1-score is 67.9%, and G-mean is 72.3%. Accuracy is 95%. Table 4 displays the conclusions from the confusion matrix naive Bayes without SMOTE.

Simple Bayes using SMOTE output recall is 53.1%, precision is 93.8%, F1-score is 95.4%, and G-mean is 72.2%. Accuracy is 95%. Results from the confusion matrix naive Bayes with SMOTE are shown in Table 5.

Random Forest

The Random Forest model trial procedure is carried out by preparing data that has already been processed during the pretreatment step, the Random Forest model trial procedure is carried out. In the wake of preprocessing, the information will be exposed to arrangement oversampling utilizing random forest. Both oversampled and non-oversampled data will be used in the random forest process. Utilizing the SMOTE oversampling approach and the random forest comparison, the classification findings from these two studies will be shown.

The random forest result is 54%, precision is 93.3%, F1-score is 62.7%, and G-mean is 73.1% without SMOTE generation. Accuracy is 95%. The results of a confusion matrix random forest without SMOTE are shown in Table 6.

Precision is 80%, F1-score is 94.3%, SMOTE result is 58.1%, and G-mean is 75.7%. These results were generated via random forest. Accuracy is 95%. The results of the confusion matrix random forest utilizing SMOTE are shown in Table 7.

Neural Network

Data that have previously undergone preprocessing are prepared for searching using the neural network

Table 7. Confusion matrix random forest with SMOTE.

Class	Predictive Positive	Predictive Negative
Actual Positive	40383	610
Actual Negative	1820	2521

Table 8. Confusion matrix neural network without SMOTE.

Class	Predictive Positive	Predictive Negative
Actual Positive	41113	24
Actual Negative	1932	2265

Table 9. Confusion matrix neural network with SMOTE.

Class	Predictive Positive	Predictive Negative
Actual Positive	38566	2539
Actual Negative	9585	31487

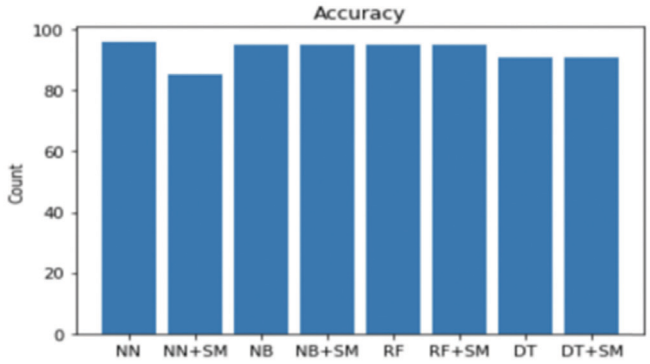

Figure 9. Accuracy result.

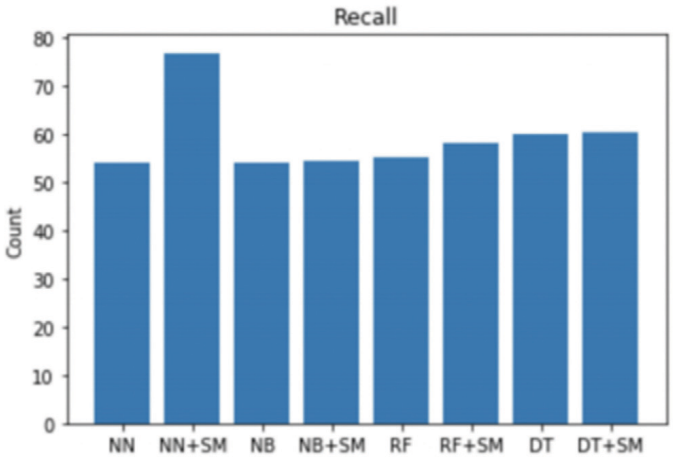

Figure 10. Recall result.

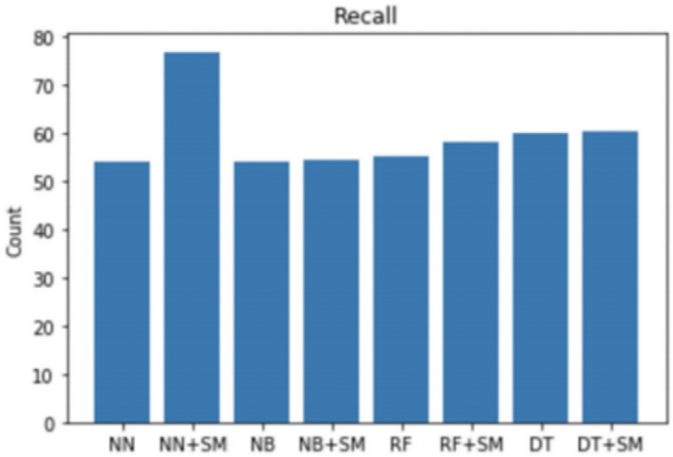

Figure 11. Precision result.

model. Following preprocessing, classification oversampling using a neural network and random forest will be performed on the data. Neural networks will be used with oversampled data, while random forests will be used with undersampled data. The findings of these two experiments will demonstrate how classification outcomes were attained utilizing neural network comparison and the synthetic minority oversampling technique (SMOTE) oversampling approach.

Neural network creation without SMOTE precision is 96.1%, F1-score is 95.1%, accuracy is 96%, recall is 56%, and G-mean is 74.5%. Results from a confusion matrix neural network without SMOTE are shown in Table 8.

The neural network that generates the SMOTE result has a 76.7% SMOTE, 92.5% precision, 85.1% F1-score, and 82.4% G-mean. The accuracy is 85%. Table 9 displays findings from the disorder framework brain network using SMOTE.

The accuracy numbers from experiments employing various methods are displayed in Figure 9. The neural network algorithm's best accuracy rating is 96%.

Review values are created by tests utilizing different calculations, as displayed in Figure 10. When AI computations and the SMOTE are utilized in place of only decision trees, random forests, naive Bayes, and brain networks, review values increase more quickly. The neural network computation and the SMOTE provided the biggest rise in review values.

As displayed in Figure 11, results from tests utilizing different calculations show that accuracy values decline while AI calculations and the SMOTE are utilized rather than just the commonly used algorithms, which we mention in the methodology, with the most noteworthy decline happening when neural network calculations and SMOTE are utilized.

As can be shown in Figure 12 from experiments using many algorithms, integrating machine-learning algorithms with the SMOTE results in higher F1-score values than just utilizing algorithms alone. The categorization of minority classes into imbalanced classes is evaluated using the F1-score.

Rather than using just the G-mean calculation to evaluate in general execution (by and large order execution), the G-mean value rose while utilizing AI calculation values as displayed in Figure 13.

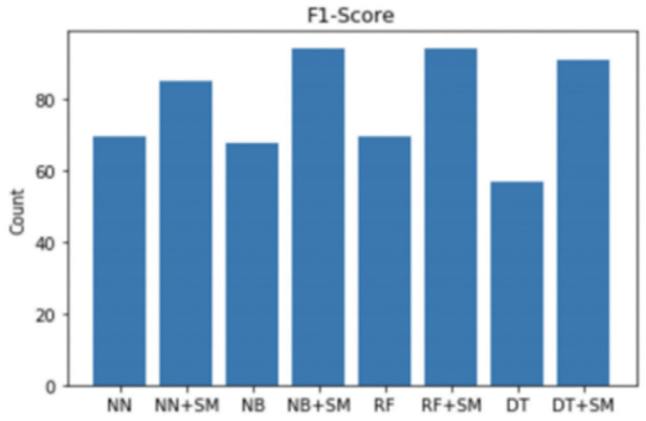

Figure 12. F1-score result.

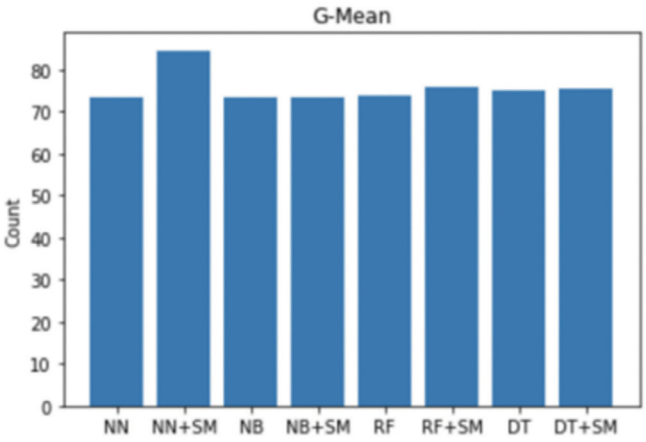

Figure 13. G-mean result.

CONCLUSION AND FUTURE WORK

A hereditary calculation can be used to determine the number of secret hubs and layers, as well as to select the appropriate qualities for brain organizations. The review, F1-score, and G-mean qualities were expanded in the analysis while utilizing the SMOTE approach. Memory utilizing brain networks rose from 52% to 74.6%, reviews utilizing gullible Bayes rose from 41.2% to 41.3%, reviews utilizing arbitrary woodlands rose from 54% to 57%, and reviews utilizing choice trees rose from 57.7% to 62.3%.

The value of the F1-score developer has increased for all AI techniques, rising from 69.8% to 85.1% for neural networks, 67.9% to 94.5% for naive Bayes, 69.8% to 94.3% for random forest, and 56.8% to 91.2% for decision trees. However, SMOTE increases the value.

In light of the discoveries of the previously mentioned try, it was resolved that SMOTE had the option to work on the exhibition of brain organizations, arbitrary timberlands, choice trees, and naive Bayes. Address the web-based business misrepresentation dataset's lopsidedness by expanding G-mean and F-1 scores in contrast with brain organizations, choice trees, irregular timberlands, and naive Bayes. This shows the viability of the SMOTE approach in raising the classification of imbalanced information execution.

Future research is anticipated to enable the use of additional computations or in-depth learning for the location of online business deception as well as other investigation to increase the accuracy of the brain network employing the SMOTE approach.

REFERENCES

[1] AsosiasiPenyelenggaraJasa Internet Indonesia, "Magazine APJI(AsosiasiPenyelenggaraJasa Internet Indonesia)" (2019): 23 April 2018.

[2] AsosiasiPenyelenggaraJasa Internet Indonesia, "Mengawaliintegritas era digital 2019 – Magazine APJI (AsosiasiPenyelenggaraJasa Internet Indonesia)" (2019).

[3] Laudon, Kenneth C., and Carol GuercioTraver. E-commerce: business, technology, society. 2016.

[4] statista.com. retail e-commerce revenue forecast from 2017 to 2023 (in billion U.S. dollars). (2018). Retrieved April 2018, from Indonesia: https://www.statista.com/statistics/280925/e-commerce-revenueforecast-in-indonesia/.

[5] Kiziloglu, M. and Ray, S., 2021. Do we need a second engine for Entrepreneurship? How well defined is intrapreneurship to handle challenges during COVID-19?. In *SHS Web of Conferences* (Vol. 120). EDP Sciences.

[6] Roy, Abhimanyu, et al. "Deep learning detecting fraud in credit card transactions." 2018 Systems and Information Engineering Design Symposium (SIEDS). IEEE, 2018.

[7] Zhao, Jie, et al. "Extracting and reasoning about implicit behavioral evidences for detecting fraudulent online transactions in e-Commerce." Decision support systems 86 (2016): 109–121.

[8] Zhao, Jie, et al. "Extracting and reasoning about implicit behavioral evidences for detecting fraudulent online transactions in e-Commerce." Decision support systems 86 (2016): 109–121.

[9] Pumsirirat, Apapan, and Liu Yan. "Credit card fraud detection using deep learning based on auto-encoder and restricted boltzmann machine." International Journal of advanced computer science and applications 9.1 (2018): 18–25.

[10] Srivastava, Abhinav, et al. "Credit card fraud detection using hidden Markov model." IEEE Transactions on dependable and secure computing 5.1 (2008): 37–48.

[11] Lakshmi, S. V. S. S., and S. D. Kavilla. "Machine Learning For Credit Card Fraud Detection System." International Journal of Applied Engineering Research 13.24 (2018): 16819–16824.

[12] Ray, S. and Leandre, D.Y., 2021. How Entrepreneurial University Model is changing the Indian COVID–19 Fight?. Entrepreneur's Guide, *14*(3), pp. 153–162.

[13] Bouktif, Salah, et al. "Optimal deep learning lstm model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches." Energies 11.7 (2018): 1636.

[14] Xuan, Shiyang, Guanjun Liu, and Zhenchuan Li. "Refined weighted random forest and its application to credit card fraud detection." International Conference on Computational Social Networks. Springer, Cham, 2018.

[15] Samrat, R., 2021. Why Entrepreuneral University Fails to Solve Poverty Eradication?. Herald Tuva State University. No. 1 Social and Human Sciences, (1), pp. 35–43.

[16] Zhao, Jie, et al. "Extracting and reasoning about implicit behavioral for detecting fraudulent online transactions in e-Commerce." Decision support systems 86 (2016): 109–121.

[17] Sharma, Shiven, et al. "Synthetic oversampling with the majority class: A new perspective on handling extreme imbalance." 2018 IEEE International Conference on Data Mining (ICDM). IEEE, 2018.

[18] Kim, Jaekwon, Youngshin Han, and Jongsik Lee. "Data imbalance problem solving for smote based oversampling: Study on fault detection prediction model in semiconductor manufacturing process." Advanced Science and Technology Letters 133 (2016): 79–84.

[19] Sadaghiyanfam, Safa, and Mehmet Kuntalp. "Comparing the Performances of PCA (Principle Component Analysis) and LDA (Linear Discriminant Analysis) Transformations on PAF (Paroxysmal Atrial Fibrillation) Patient Detection." Proceedings of the 2018 3rd International Conference on Biomedical Imaging, Signal Processing. ACM, 2018.

[20] Harrison, Paula A., et al. "Selecting methods for ecosystem service : A decision tree approach." Ecosystem services 29 (2018): 481–498.

[21] Ray, S., 2021. Are Global Migrants At Risk? A Covid Referral Study of National Identity. In Transformation of identities: the experience of Europe and Russia (pp. 26–33).