

Sampling methods in handling imbalanced data for Indonesia health insurance dataset

Felix Indra Kurniadi¹, Kartika Purwandari¹, Ajeng Wulandari¹, Syarifah Diana Permai²

¹Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia

²Statistics Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia

Article Info

Article history:

Received Nov 18, 2022

Revised Jan 20, 2023

Accepted Jan 30, 2023

Keywords:

Health insurance frauds

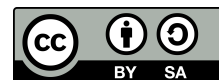
Machine learning

Sampling method

ABSTRACT

Health insurance fraud is one of the most frequently occurring fraudulent acts and has become a concern for every insurance. According to data from The Indonesian General Insurance Association or *Asosiasi Asuransi Umum Indonesia* (AAUI), the private insurance industry suffered losses up to billions rupiah throughout 2018 due to the fraudulent acts committed by the perpetrators. The problem in with the number of frauds in Indonesia is that the current system is highly vulnerable and they is still done manually. The other problem from this detection is imbalance data which often occurs in fraudulent cases. In this research, we used a sampling methods using several machine learning as the baseline. The result shows that the instance hardness thresholding algorithm and extreme gradient boosting gives the best performance for all the case. It shows the method can reduced the bias and can achieve better generalization.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Felix Indra Kurniadi

Computer Science Department, School of Computer Science, Bina Nusantara University

Jl. Kebon Jeruk Raya No. 27, Kebon Jeruk, Jakarta, Indonesia

Email: felix.indra@binus.ac.id

1. INTRODUCTION

Health insurance fraud is one of the most frequently occurring fraudulent acts and has become a concern for every insurance. Fraud in health insurance can be in the form of a claim against a case not covered by insurance or even making a claim for action on a health procedure that has never been carried out. Another problem in detecting fraud is the procedure always done in manually.

A public health expert said the Healthcare and Social Security Agency or *Badan Penyelenggara Jaminan Sosial* (BPJS Kesehatan) suffer a loss around Rp 1.86 trillion (USD 140.5 million). Furthermore, Rp 6.9 trillion required further investigation [1]. Based on a report provided by the Nations Association of Certified Fraud Examiners (ACFE), losses caused by fraud in Indonesian health services are around 5% of the total service cost. Based on observational data, around 175,000 claims are proven fraud with a total loss of 400 million rupiahs. Until 2019, there is a possibility of 1 million claims suspected of being a fraud [2].

According to Fatimah *et al.* [2], the prevalence of fraudulent activities in Indonesia has been exacerbated by a system that is extremely vulnerable to such deceptions. Currently, the majority of fraud detection mechanisms are manual, which substantially lengthens the time required to identify illegal processes. As a result, the rate of effective identification and resolution of fraud lags behind, creating an urgent need for enhanced automated systems to improve security.

It can be concluded from each of the problems mentioned above that a system that can detect fraud in health insurance data is needed. Several studies have carried out health insurance fraud detection [3]–[5]

using the extreme gradient boosting (XGBoost) method as a classifier to detect health insurance fraud. Other studies such as [5], [6] and used the Catboost method, and [7] used the logistic regression method for modeling. The other obstacle is the problem of imbalanced datasets that often occur in fraud cases. Several studies have attempted to overcome this problem. One of them is [8], which uses a random undersampler to solve the imbalanced problem. In this study, we propose several solutions to the problem of insurance fraud detection. One proposed solution is to use XGBoost [3]–[5], which will be compared by several algorithms such as K-nearest neighbor (KNN), support vector machine (SVM) and logistic regression. We also compared without using sampling and using outcome sampling. The main objectives for this research are:

- Examining for the best algorithm in the insurance fraud detection process.
- Result comparison between using the sampling method and not using the sampling method.
- Examining a sampling method that delivers optimal results on health insurance data from health insurance companies in Indonesia.

The rest of the paper is structured into five sections: section 2 provides previous research on health insurance fraud. Section 3 explains the dataset and the distribution of the data. Section 4 explains the methodology in this paper, and each step of the methodology will be explained thoroughly. Section 5 shows the experiment and the result of the experiment. Section 6 concludes all the finding.

2. RELATED WORKS

Research to overcome health insurance fraud has been carried out in various ways, one of which is the artificial intelligence approach. Research conducted by Akbar *et al.* [3], the methods used are XGBoost and random forest for the modeling process. These methods are also used to improve the results of conducting a random undersampling process to obtain balanced data. The conclusion provided in this study is that XGBoost with random undersampler and tuning parameters delivers good results. However, it should be understood that this method is prone to noise or outliers.

Shamitha and Ilango [9]. proposed the use of an artificial neural network (ANN) for modeling. Sharmita carried out several preprocessing stages: handling missing values, data filtering, label encoding, data transformation, and scaling. The following stage is the dimensional reduction process using principal component analysis (PCA). Besides that, Sharmita used the synthetic minority over-sampling technique (SMOTE) method to overcome the problem of imbalanced data before it was finally included in the modeling. This study concludes that ANN with the right process provides the best accuracy results compared to several other machine learning benchmark methods.

Hancock and Khoshgoftaar [6] proposed the Catboost method compared to the XGBoost method. This study found that Catboost was used to provide a baseline for the method to reduce the preprocessing process for imbalanced data. This is also proven by the higher mean area under of curve (AUC) of XGBoost for complex features. The same research was conducted by Hancock and Khoshgoftaar [5] with the same data but different experimental scenarios. The results provided confirm what was stated in the study [6]. Different from the approach of Hancock and Khoshgoftaar [5], [6] and Akbar *et al.* [3], the approach taken by Seo and Mendelevitch [10] and Georgakopoulos *et al.* [11] was to detect fraud with outlier detection analysis.

Based on the research done by previous researchers, it is known that fraud detection always has problems in handling imbalanced ones. Several studies have also tried to deal with this by providing samplings such as the random undersampler [3] and SMOTE [9]. Therefore, this research will focus on determining what the best sampling method is. Our modeling process will use XGBoost, KNN, logistic regression, and SVM methods.

3. DATA

The data used in this study was acquired from the data warehouse of one of the health insurance companies in Indonesia. Data has 11,882 data instances with 51 features. Table 1 and Figure 1 provide an overview of the distribution between fraud and non-fraud classes. We divided the data into independent and dependent variables from the given data. Of the 51 available features, only 12 features were used as independent variables, i.e. Total Claim, Total Claim Amount, Total LOS, TotalICU LOS, InvestigationFlg, BasicPremiumCollected, PolicyAge, InsuredAge(claim), InsuredSmoke, InsuredGender, InsuredWeight, InsuredHeight and one feature is used as class is HoldFlg.

Table 1. Data distribution frequency in health insurance dataset

Class	Frequency	Probability
Fraud	109	0.009
Non-Fraud	11,773	0.991
Total	11,882	1

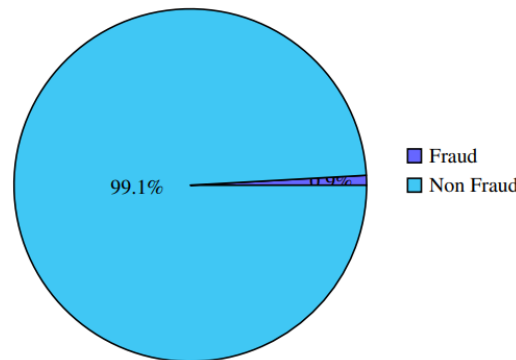


Figure 1. Data distribution in Indonesia health insurance dataset

4. METHOD

The process carried out at the methodology stage is as: the first process carried out is preprocessing the data. Preprocessing starts with selecting the features used using Pearson's correlation coefficient. Subsequently, we checked whether the data had a missing value and checked the distribution of the data. The last step is to change the data from text to categorical using a label encoding approach. Afterward, the normalization process is carried out for each selected feature. The normalization process used is scaling using min-max values. In the following process, we divided into two scenarios which will be explained in more detail in section 5. We selected four classifiers used as baselines for this study. The four classifiers used are SVM, logistic regression, XGBoost, and KNN. XGBoost selection in this study referred to the research by [4], [5], [12] and the logistic regression method selection was based on [7], [12], [13]. Research conducted by

In the following stage, the object assessed the model with measurements such as accuracy, precision, recall, and f1-score. The research methodology was better explained in Figure 2. Each subsection provides a complete description of the steps of the methodology used; starting from preprocessing, sampling methods, machine learning models and lastly, evaluation metric.

4.1. Preprocessing

The selected main features will be examined for correlation using Pearson's correlation coefficient in the first stage. The formulation of Pearson's correlation coefficient is as [14]:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 (y_i - \bar{y})^2}} \quad (1)$$

x, y is a feature. After understanding the correlation for each feature, we looked for a correlation value that was more than the threshold we set for the drop process for that feature.

The following preprocessing stage was the search for missing values. Subsequently, we handled missing values by dropping data that had missing values. Then, the label encoding process where each string category would be assigned a discrete value as one example is gender = ["Male", "Female"] to gender = [0, 1].

The last step of preprocessing was normalization. In the normalization process, we used feature scaling with min-max, where the formulation is [15]

$$scaling = \frac{f_i - \min(f)}{\max(f) - \min(f)} \quad (2)$$

f is the data value of the feature to be processed by the feature scaling.

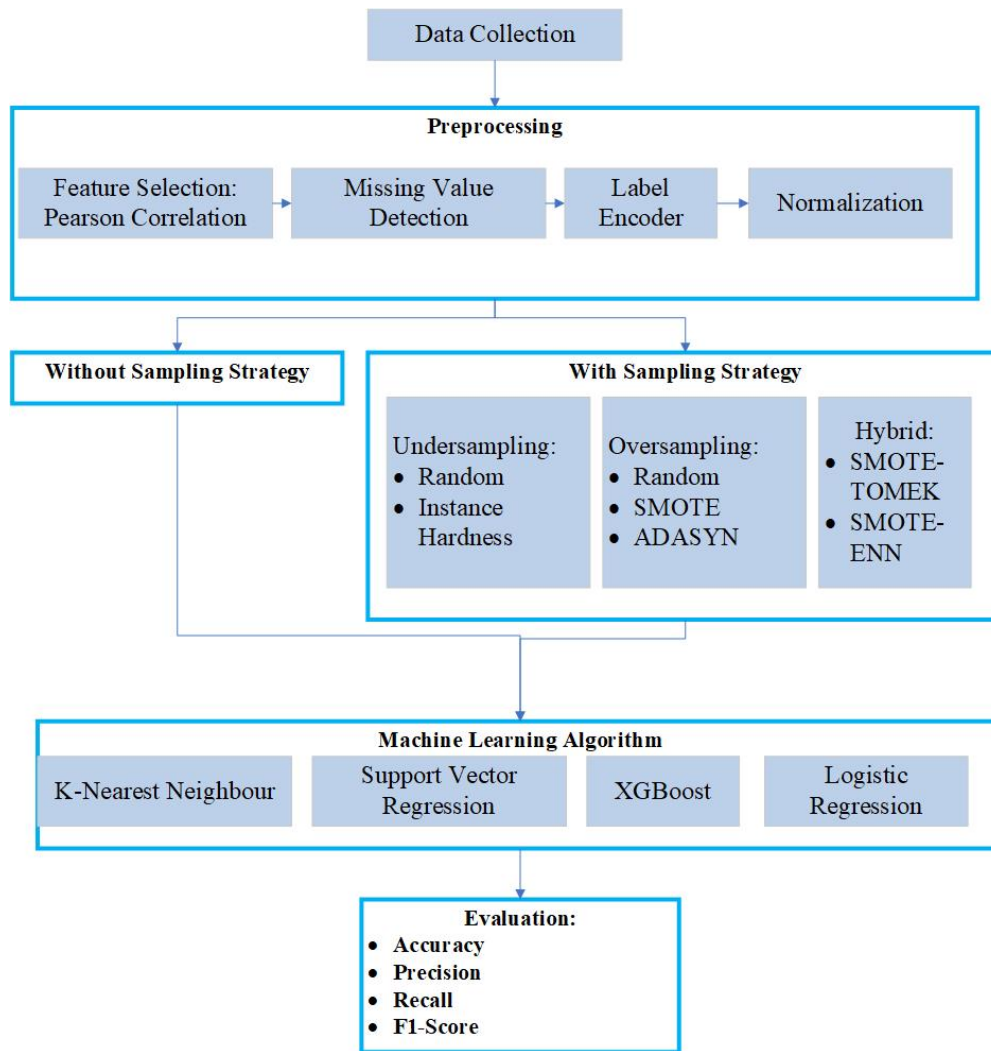


Figure 2. Workflow of Indonesia health insurance fraud detection

4.2. Sampling method

Random undersampling removes the majority class; thus, the number of data is equal to the number of data from the minority class. It is one of the easiest methods to deal with the imbalanced data problem [16]. Instance hardness is an undersampling method that removes "hard" samples. If the probability of the data is smaller than the specified threshold, then the data will be deleted [17]. Random oversampling by replicating the minority class by taking all values in the class and replicating these values; thus, the dataset increases. Random oversampling does not change the diversity of the sample and does not create new data [16]. SMOTE is one of the oversampling methods in which the data with the minority class is increased to equal the majority class. SMOTE sampling is done to minimize the expected risk [18]. Adaptive synthetic sampling (ADASYN) is an oversampling weight distribution method from minority class samples based on the difficulties in the learning process. The more synthetic samples generated for minority samples, the harder it is to learn [19]. Synthetic minority over-sampling technique-TomekLink (SMOTE-Tomek) is a hybrid learning method that combines SMOTE and TomekLink methods. The standard flow of TomekLink is; first, the imbalanced data (D) is used by SMOTE for the balancing process, and then the Tomek Link algorithm is used to carry out the undersampling process [20]. A similar process is carried out by synthetic minority over-sampling technique-edited nearest neighbor (SMOTE-ENN), where SMOTE carries out the oversampling process, and the undersampling process is carried out by the ENN method [21].

4.3. Machine learning method

KNN is a supervised learning method. KNN is a non-parametric method. KNN itself is a relatively simple machine learning method where this method performs classification by checking the surrounding k data, where the majority class will be the class for new data. This method is often called the lazy model because it does not do intensive learning during training [13]. In the calculating distance, the KNN method uses distance methods such as Euclidean distance with the formula. XGBoost is a method proposed by Chen and Guestrin to improve the gradient tree method [22]. XGBoost itself is an efficient algorithm and is one of the most frequently used methods for Kaggle competition. The main idea of this method is to form sub-trees of the original tree where each subsequent tree reduces the error of the previous tree [22]. Logistic regression is a supervised learning method that is usually used in multivariable. Unlike linear regression, which produces continuous results, logistic regression produces binary targets [23]. The SVM determines the optimum hyperplane for separating the data by assuming that the best decision boundary has the greatest distance and margin from both classes of data. SVM is also known as a maximum margin classifier [24].

4.4. Evaluation metric

This study evaluates model performance using accuracy, precision, recall, and F1-score. Accuracy is the model's positive and negative prediction accuracy. Precision measures the accuracy of positive identifications, while recall measures the accuracy of real positives. The F1-score balances precision and recall, making it beneficial when false positives and false negatives have distinct costs. The formulas for the four metrics are [13]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \quad (6)$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

5. EXPERIMENT AND DISCUSSION

The experiments in this paper were carried out using a dataset from one of Indonesia's largest health insurance companies. An explanation of the dataset and its distribution is explained in section 2. In solving the problem, this experiment was carried out using several python libraries such as scikit-learn [25], pandas [26], [27], and imblearn [28]. Figure 3 describes the distribution of data in a scatter plot using PCA to reduce the dimensions of the data held. Figure 4 provides an overview of the heatmap of the correlation coefficient for the features used.

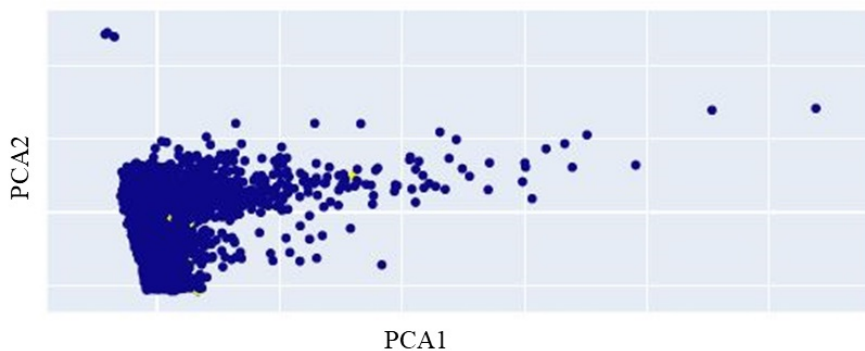


Figure 3. Scatterplot PCA for health insurance fraud dataset

The scatter plot of our health insurance fraud dataset showed the distribution between fraud and non-fraud is not easily to separate. In this research, we used a threshold for the Pearson correlation is 0.8. Based on the correlation heatmap from Figure 4, each features did not have more than 0.8 correlation values. Therefore, we used all the features for our research. Moreover, the Table 2 showed our data did not have any missing values.

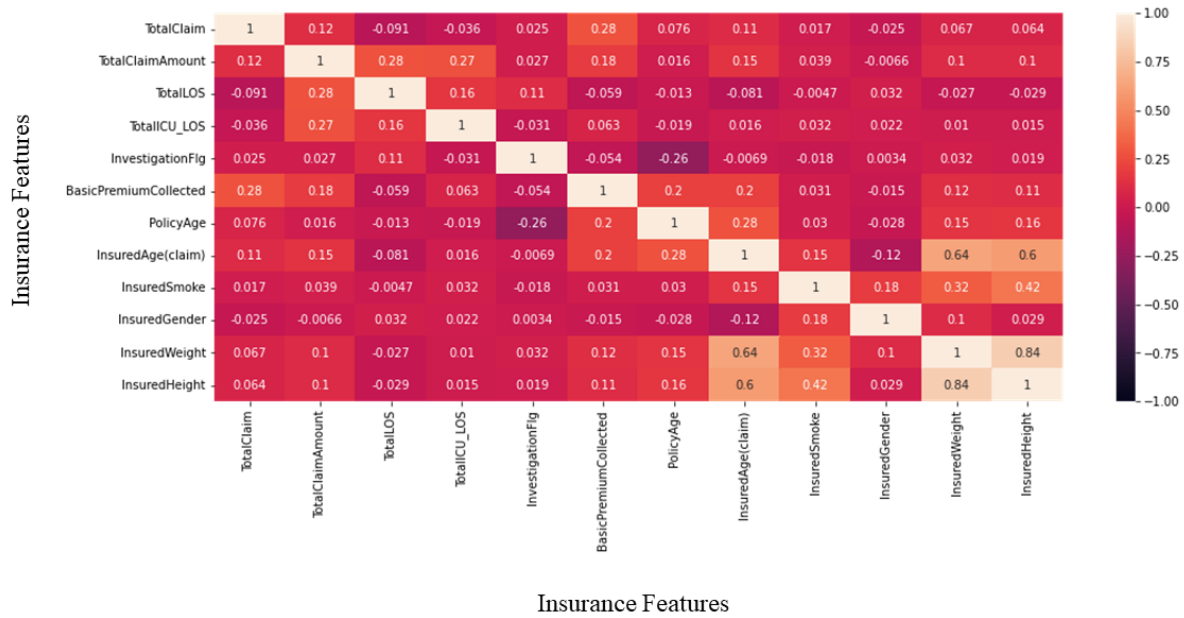


Figure 4. Pearson's correlation heatmap

Table 2. Total missing values each feature

Features	Total missing value
TotalClaim	0
TotalClaimAmount	0
TotalLOS	0
TotalICU_LOS	0
InvestigationFlg	0
BasicPremiumCollected	0
PolicyAge	0
InsuredAge(Claim)	0
InsuredSmoke	0
InsuredGender	0
InsuredWeight	0
InsuredHeight	0

Subsequently, we divided this research into four parts. In the first part, we researched health insurance fraud using only raw data or referred to in the following as the "Vanilla" approach. While in the second part, we experimented with using several over-sampling methods. The following part used several undersampling approaches, and lastly, we used a hybrid sampling approach. Each experiment will be conducted with 10-fold cross-validation to determine the best model. Table 3 illustrated the result of our experiment in vanilla approach, while Tables 4 to 10 illustrated the result of our experiment using several sampling methods.

Table 3. Result without sampling method (vanilla approach)

Methods	Accuracy	Precision	Recall	F1-score
Linear Svc	99.083	0	0	0
KNN	98.92	0.062	0.22	0.099
Logistic regression	99.07	0	0	0
XGBoost	99.09	0.025	0.25	0.04

Based on Table 3, the results provided illustrate that the accuracy results are unbalanced with the precision, recall, and f1-score values. This can be considered a problem because the accuracy value describes a reasonably high result, with the maximum on XGBoost being 99.09%. In reality, the precision, recall, and F1-score values are unbalanced with the accuracy value. This imbalance is because of the confusion matrix given on Figure 5. It can be seen that most predictions are centered on the method that has the largest training data. This is a problem in the bias towards the machine learning model created.

Table 4. Result with over-sampling method oversampler

Methods	Accuracy	Precision	Recall	F1-score
Linear Svc	95.38	0.92	0.95	0.96
KNN	98.79	0.98	0.99	0.99
Logistic regression	95.94	0.92	0.96	0.96
XGBoost	98.3	0.97	0.98	0.98

Table 5. Result with over-sampling method SMOTE

Methods	Accuracy	Precision	Recall	F1-score
Linear Svc	95.59	0.92	0.9	0.96
KNN	98.09	0.96	0.98	0.98
Logistic regression	96.21	0.93	0.99	0.96
XGBoost	98.41	0.97	0.99	0.98

Table 6. Result with over-sampling method ADASYN

Methods	Accuracy	Precision	Recall	F1-score
Linear Svc	95.61	0.92	0.99	0.95
KNN	98.08	0.96	0.98	0.98
Logistic regression	96.21	0.93	0.96	0.96
XGBoost	98.35	0.97	0.98	0.98

Table 7. Result with undersampling method undersampler

Methods	Accuracy	Precision	Recall	F1-score
Linear Svc	95.37	0.92	0.95	0.96
KNN	98.79	0.98	0.99	0.99
Logistic regression	95.88	0.92	0.96	0.96
XGBoost	98.29	0.97	0.98	0.98

Table 8. Result with undersampling method instance hardness

Methods	Accuracy	Precision	Recall	F1-score
Linear Svc	99.7	0.9	1	0.84
KNN	99.72	0.89	1	0.85
Logistic regression	99.71	0.91	1	0.84
XGBoost	99.87	0.98	1	0.93

Table 9. Result with hybrid sampling: SMOTE-ENN

Methods	Accuracy	Precision	Recall	F1-score
Linear Svc	97.77	0.96	0.98	0.98
KNN	99.59	0.99	1	1
Logistic regression	98.3	0.97	0.98	0.98
XGBoost	99.34	0.99	0.99	0.99

Based on Tables 4 to 10, the results given illustrate that the sampling process greatly affects the results and reduces the habit of machine learning models. The most suitable approach for our health insurance data is undersampling, especially using instance hardness sampling. This is because the instance hardness thresholding

is an algorithm that looks for the probability of no "hardness" in the dataset. In addition, from every classifier model used in this study, the XGBoost method almost beats every method compared. This result shows that the margin of difference between precision, recall, and f1-score is the smallest compared to other methods.

Table 10. Result with hybrid sampling: SMOTE-Tomek

Methods	Accuracy	Precision	Recall	F1-score
Linear Svc	95.58	0.92	0.96	0.96
KNN	98.08	0.96	0.98	0.98
Logistic regression	96.19	0.93	0.96	0.96
XGBoost	98.41	0.97	0.98	0.98

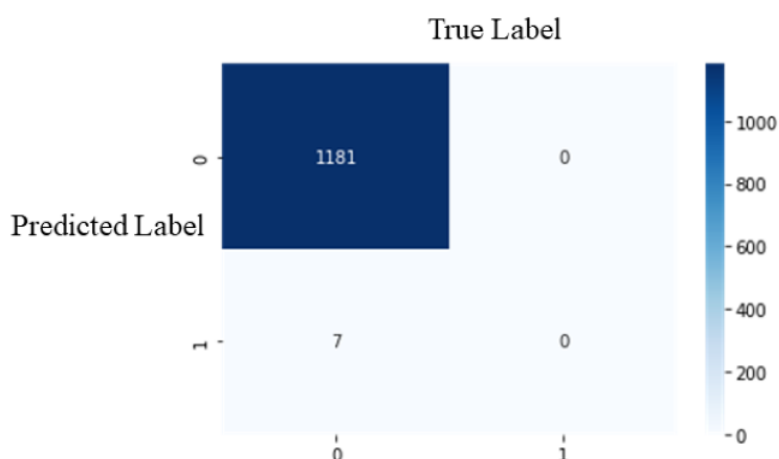


Figure 5. Confusion matrix using linear SVM

6. CONCLUSION

This paper discusses several sampling algorithms for reducing bias in the evaluation results of imbalanced data. The data used is data from one of Indonesia's largest health insurance companies. This research uses four machine learning methods as a baseline, i.e. SVM, KNN, logistic regression and XGBoost. The sampling used is in three categories, i.e. undersampling with random undersampling, instance hardness thresholding; oversampling with random oversampling, SMOTE, and ADASYN; and lastly, hybrid sampling using the SMOTEENN and SMOTE-Tomek methods. This study's results show that the sampling process, either undersampling, oversampling, or hybrid, can reduce habits in machine learning methods. This can be seen where the values of precision, recall, and f1-score are very low in the "Vanilla" approach, where the best for precision, recall, and f1-score are less than 0.3. Meanwhile, after using sampling, the values between accuracy, precision, recall, and the f1-score have similar values, with the biggest difference being 0.1. This study also found that the best classifier is XGBoost. It can be observed that, unlike other classifier methods, the XGBoost method provides unbiased results and has the smallest margin for each sampling method.




REFERENCES

- [1] "Hospital fraud costing BPJS Kesehatan, expert says," *The Jakarta Post*, 2016. <https://www.thejakartapost.com/news/2016/08/26/hospital-fraud-costing-bpjs-kesehatan-expert-says.html>.
- [2] R. N. Fatimah, Misnaniarti, and R. A. Syakurah, "Potential fraud in the implementation of national health insurance in the health sector: systematic review," *Journal: JMMR (Jurnal Medicoeticolegal dan Manajemen Rumah Sakit)*, vol. 10, no. 3, pp. 255–270, 2021, doi: 10.18196/jmmr.v10i3.10825.
- [3] N. A. Akbar, A. Sunyoto, M. R. Arief, and W. Caesarendra, "Improvement of decision tree classifier accuracy for healthcare insurance fraud prediction by using extreme gradient boosting algorithm," in *2nd International Conference on Informatics, Multimedia, Cyber, and Information System (ICIMCIS)*, Nov. 2020, pp. 110–114, doi: 10.1109/ICIMCIS51567.2020.9354286.
- [4] N. Dhiieb, H. Ghazzai, H. Besbes, and Y. Massoud, "A secure ai-driven architecture for automated insurance systems: fraud detection and risk measurement," *IEEE Access*, vol. 8, pp. 58546–58558, 2020, doi: 10.1109/ACCESS.2020.2983300.




- [5] J. Hancock and T. M. Khoshgoftaar, "Performance of CatBoost and XGBoost in medicare fraud detection," in *19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec. 2020, pp. 572–579, doi: 10.1109/ICMLA51294.2020.00095.
- [6] J. Hancock and T. M. Khoshgoftaar, "Medicare fraud detection using CatBoost," in *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*, Aug. 2020, pp. 97–103, doi: 10.1109/IRI49571.2020.00022.
- [7] J. H. Wilson, "An analytical approach to detecting insurance fraud using logistic regression," *Journal of Finance and accountancy*, vol. 1, p. 15, 2009.
- [8] R. A. Bauder and T. M. Khoshgoftaar, "Medicare fraud detection using random forest with class imbalanced big data," in *2018 IEEE international conference on information reuse and integration (IRI)*, Jul. 2018, pp. 80–87, doi: 10.1109/IRI.2018.00019.
- [9] S. K. Shamitha and V. Ilango, "A time-efficient model for detecting fraudulent health insurance claims using artificial neural networks," in *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)*, Jul. 2020, pp. 1–6, doi: 10.1109/ICSCAN49426.2020.9262298.
- [10] J. Seo and O. Mendelevitich, "Identifying frauds and anomalies in Medicare-B dataset," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul. 2017, pp. 3664–3667, doi: 10.1109/EMBC.2017.8037652.
- [11] S. V. Georgakopoulos, P. Gallos, and V. P. Plagianakos, "Using Big data analytics to detect fraud in healthcare provision," in *2020 IEEE 5th Middle East and Africa Conference on Biomedical Engineering (MECBME)*, Oct. 2020, pp. 1–3, doi: 10.1109/MECBME47393.2020.9265118.
- [12] M. Hanafy and R. Ming, "Machine learning approaches for auto insurance big data," *Risks*, vol. 9, no. 2, p. 42, Feb. 2021, doi: 10.3390/risks9020042.
- [13] M. Hanafy and R. Ming, "Classification of the insureds using integrated machine learning algorithms: a comparative study," *Applied Artificial Intelligence*, vol. 36, no. 1, pp. 1–32, Jan. 2022, doi: 10.1080/08839514.2021.2020489.
- [14] I. M. Nasir *et al.*, "Pearson correlation-based feature selection for document classification using balanced training," *Sensors*, vol. 20, no. 23, p. 6793, 2020, doi: 10.3390/s20236793.
- [15] V. N. G. Raju, K. P. Lakshmi, V. M. Jain, A. Kalidindi, and V. Padma, "Study the influence of normalization/transformation process on the accuracy of supervised classification," in *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2020, pp. 729–735, doi: 10.1109/ICSSIT48917.2020.9214160.
- [16] M. Khushi *et al.*, "A comparative performance analysis of data resampling methods on imbalance medical data," *IEEE Access*, vol. 9, pp. 109960–109975, 2021, doi: 10.1109/ACCESS.2021.3102399.
- [17] M. R. Smith, T. Martinez, and C. Giraud-Carrier, "An instance level analysis of data complexity," *Machine Learning*, vol. 95, no. 2, pp. 225–256, Nov. 2014, doi: 10.1007/s10994-013-5422-z.
- [18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.
- [19] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, 2008, pp. 1322–1328, doi: 10.1109/IJCNN.2008.4633969.
- [20] Z. Wang, C. Wu, K. Zheng, X. Niu, and X. Wang, "SMOTETomek-based resampling for personality recognition," *IEEE Access*, vol. 7, pp. 129678–129689, 2019, doi: 10.1109/ACCESS.2019.2940061.
- [21] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20–29, Jun. 2004, doi: 10.1145/1007730.1007735.
- [22] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
- [23] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 20, pp. 215–232, 1958, doi: 10.1111/j.2517-6161.1958.tb00292.x.
- [24] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, Sep. 1995, doi: 10.1007/bf00994018.
- [25] Pedregosa Fabian *et al.*, "Scikit-learn: machine learning in Python," *the Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [26] W. McKinney, "Data structures for statistical computing in Python," in *Proceedings of the 9th Python in Science Conference*, 2010, pp. 56–61, doi: 10.25080/majora-92bf1922-00a.
- [27] "pandas-dev/pandas: Pandas," *The pandas development team*. <https://zenodo.org/record/7344967>.
- [28] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research*, vol. 18, pp. 1–5, 2017.

BIOGRAPHIES OF AUTHORS






Felix Indra Kurniadi    received his Bachelor of Computing from Tarumanagara University, Indonesia, in 2014, and his Master of Computer Science, Depok, Indonesia, in 2017. He is currently a lecturer at Computer Science Department Bina Nusantara University, Jakarta, Indonesia. In the past 5 years, he worked on several university as lecturer and as an artificial intelligence Engineer at private company. His research interest includes machine learning, deep learning, computer vision, and data science. He can be contacted at email: felix.indra@binus.ac.id.






Kartika Purwandari    received the bachelor's degree in Information Technology from Brawijaya University and the master's degree in Computer Science from National Central University Taiwan. She is currently a lecturer at Computer Science Department in Bina Nusantara University, Jakarta, Indonesia. She is also a lecturer specialist S2 of basic programming in Bina Nusantara University since December 2021. In the past 2 years ago, she was become a research assistant at Bioinformatics and Data Science Research Center (BDSRC) Bina Nusantara University. She has developed programming based on artificial intelligence and bioinformatics by joining the colorectal cancer project since she joined BDSRC. Furthermore, she is also active in participating with AI projects in BDSRC to help in processing data about lidar, air quality, crowd counting, fishery image, text, and pap smear. She can be contacted at email: kartika.purwandari@binus.edu.



Ajeng Wulandari    received her Bachelor of Computing from Duta Wacana Christian University, Yogyakarta, Indonesia, in 2015, and her Master of Computer Science, Depok, Indonesia, in 2019. Since 2020, she is working as a Computer Science Lecturer in BINUS University, Jakarta, Indonesia. Her research interest includes computer vision, machine learning, deep learning, and artificial intelligence. She can be contacted at email: ajeng.wulandari@binus.ac.id.



Syarifah Diana Permai    currently works as a Lecturer at the Department of Computer Science and Statistics, Binus University, Indonesia since 2014. She holds a Bachelor's degree and Master's degree in Statistics at Institut Teknologi Sepuluh November (ITS), Indonesia in 2011 and 2013. She has published papers in several cases such as economics, health, finance, and chemistry. using statistical analysis. Her research interests include machine learning, Bayesian, time series analysis, survival analysis and spatial analysis. She can be contacted at email: syarifah.permai@binus.ac.id.