IMPLEMENTATION OF RANDOM OVERSAMPLING TECHNIQUE IN THE K-NEAREST NEIGHBOR METHOD FOR CREDITWORTHINESS ANALYSIS

Ayu Dhita Putri Wulandari[1)]
Shantika Martha[2)]
Wirda Andani[3)]
[1,2,3)] Statistics, Faculty of Mathematics and Natural Sciences, Tanjungpura University, Pontianak, Indonesia
e-mail: ayudhita291@student.untan.ac.id

ABSTRACT

Banks are financial institutions, one of whose main activities is providing credit to their customers. The existence of credit granting activities requires the bank to know the feasibility of prospective debtors in receiving credit. Because in practice, credit granting activities still often have bad credit problems. The problem of bad credit can be overcome by analyzing the feasibility of granting credit to prospective debtors. The data used in this study consists of 10 independent variables and 1 dependent variable is collectibility (kol). The collectibility (col) data consists of 500 data for the current debtor class and 26 data for the non-current debtor class, this indicates an imbalance class. So in this study, the application of the random oversampling (ROS) technique is used to overcome the imbalance class with the K-Nearest Neighbor (KNN) method in classifying current and non-current debtor data. ROS was chosen because it can generally provide better results and does not eliminate information from existing data. The analysis results obtained show that the use of the KNN method with the application of ROS is better than the KNN model without ROS, with an accuracy of 84.91% at data testing. The KNN model with ROS can improve the model's ability to classify noncurrent debtor data or the specificity value of the model increases by 25%. In the KNN model without ROS the model cannot classify non-current debtor data correctly at all, this can endanger the bank in making decisions.

Keywords: credit worthiness, KNN, imbalance class, ROS.

INTRODUCTION

Banks are financial institutions where one of their main activities is to channel funds to the public. This activity can take the form of providing credit. According to Law Number 10 of 1998, credit is a process where a bank provides money or loans to other parties based on a loan agreement, with the obligation for the other party to repay their debt after a certain period, usually with added interest.

In the implementation of credit provision activities, banks are required to determine whether borrowers have credibility and the ability to repay a loan within the established time frame. This is because extending credit, besides being a source of income, also poses a significant threat to business operations (Adi & Winarko, 2015). In practice, credit provision often leads to occurrences of non-performing loans. Non-performing loans are loans that encounter issues where borrowers fail to repay the loan as agreed upon with the bank. According to Winata et al., (2013), non-performing loans can occur due to several reasons, such as misuse of credit, deliberate non-payment by the borrower, inadequate credit analysis, and lack of supervision from the bank. Non-performing loan issues can result in losses for the bank, both financially and non-financially. Delinquent loan payments pose obstacles to the smooth operation of banking businesses.

One way to address this issue is by conducting proper credit analysis on prospective borrowers. Thus, before making a credit decision, the bank can determine whether the customer applying for credit is eligible or not. One of the analysis processes that can be conducted is by using data mining. Data mining is a concept in information technology related to data and information. It can be used as a method or modeling technique to identify patterns and relationships among data variations (Ginting, 2019). One aspect of data mining is classification. Classification involves grouping data to form a set of models aimed at predicting a class of objects whose class is unknown (Pramadhana, 2021). Classification falls under supervised learning, a method used to discover relationships between input attributes and target attributes (Hendrian, 2018). A commonly used classification method is K-Nearest Neighbor (KNN). KNN is a simple, efficient, and effective classification method in the field of object processing capable of handling large training data sets (Bhatia, 2010). For instance, in the study by Wajhillah et al., (2019), research was conducted on the accuracy of the K-Nearest Neighbor (KNN) method in analyzing non-performing loans. The accuracy of the KNN method showed satisfactory results. However, in reality, problems are often found in the classification process, such as when one class of response variables has an unbalanced number of data, also known as an imbalanced class. This can cause the classification process and the obtained model to become inaccurate. Therefore, a way to address the imbalance class issue in classification is needed. One approach that can be taken is by resampling the data to address this issue. 3 There are two resampling techniques that can be used to balance the dataset: random oversampling (ROS) by adding minority class instances, this method is used when the dataset is insufficient, and random undersampling (RUS) by reducing the size of the majority class, this method is used when the dataset is sufficient. According to Santoso et al., (2017), oversampling methods generally provide better results compared to undersampling methods because undersampling methods discard a lot of data (Wijayanti et al., 2021). This research aims to address the imbalance class in the data of bank x debtor credits. This study combines random oversampling technique to address the imbalance class and uses the K-Nearest Neighbor classification method to predict whether debtor data belongs to the current or non-current class. Thus, the results of this research can be used as a reference in credit decision-making by the bank and also can be used as an analysis handling when the existing data is imbalanced.

METHOD

Classification

Classification is a way of grouping objects based on the characteristics they possess. A classification is made from a set of training data with predetermined classes. Classification is one of the processes in data mining. It involves grouping features into appropriate classes. Classification falls under supervised learning methods, which are used to discover relationships between input attributes and target attributes (Hendrian, 2018).

K-Nearest Neighbor (KNN)

The K-Nearest Neighbor (KNN) is a method used for classifying a dataset. The working principle of KNN is to find the closest distance between the data to be evaluated and the k neighbors in the training data (Whidhiasih et al., 2013). KNN classifies by considering the nearest class of an object. This KNN method falls into the category of simple classification methods, effective in pattern recognition, text categorization, and so on. Determining the value of k used in this study involves finding the optimum accuracy value among the k values tested, ranging from k=1 to k=5. In the application of

the KNN method, it is important to know the units of data to be analyzed; differences in unit variability can lead to inaccurate analysis results. Therefore, before calculating the distance between data, data transformation is performed to standardize the scale and/or data range so that each input variable contributes relatively equally. The process that can be carried out is min-max normalization (Martha et al., 2022). The equation for min-max normalization is shown in equation 1.

$$x_i' = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \quad (1)$$

with
x_i'= New data resulting from normalization
x_i = The old value of each data entry at index-i
$\min(x_i)$ = The minimum value of the independent variable at index- i
$\max(x_i)$ = The maximum value of the independent variable at index- i

The calculation to find the nearest neighbors in the KNN method can use the Euclidean distance formula. The Euclidean distance is a commonly used distance to determine the nearest neighbors. The equation for the Euclidean distance is as follows (Wicaksono, 2017):

$$d_{hj} = \sqrt{\sum_{i=1}^{n} (x_{hi} - x_{ji})^2} \quad (2)$$

with
d_{hj}= The distance between the h-th training data and the j-th testing data
x_{hi} = The h-th independent variable in the i-th training data
x_{ji} = The j-th independent variable in the i-th testing data
n = The number of independent variables

After obtaining the distance values between the training data and the testing data, these distance values are sorted from smallest to largest. Subsequently, the process of selecting the nearest distances continues until reaching the optimal parameter k to be used. The results of selecting the nearest distances up to the optimal parameter k are used to determine the classification class. The classification class is determined by the most frequent class or the dominant class that appears in the results of selecting the k nearest distances. This dominant class is chosen as the classification class for the testing data.

Random Oversampling (ROS)

Imbalanced class is a condition where the class ratio in each dataset is not balanced, or when the number between the majority and minority classes differs significantly (Astuti & Lenti, 2021). If the data is not evenly distributed across each class, the possibility of classification errors is quite high in the minority class. Imbalanced class in the classification process can cause the classifier to learn less from the minority class, resulting in predictions leaning more towards the majority class (Dahlan, 2022).

Handling imbalanced classes can be done using resampling methods. Resampling methods work to balance the data so that the proportion of the resulting data is equal (balanced) (Saifudin & Wahono, 2015). The common resampling technique used is random oversampling (ROS).

The random oversampling technique is used when the number of datasets is insufficient. This technique balances the dataset by randomly selecting minority class data and adding it to the training data. The selection and addition process are repeated until the number of minority class data equals the number of majority class data. First, the difference between the majority and minority class data is calculated. Then, the process of randomly selecting minority class data and adding it to the training data is carried out while repeating the process as many times as the difference between the majority and minority classes.

Confusion Matrix

The confusion matrix is a table that depicts the performance of a specific model or algorithm. Each row of the matrix represents the actual class of the data, and each column represents the predicted class of the data (or vice versa) (Wicaksono, 2017). The confusion matrix can be used to measure performance in both binary and multiclass classification problems, the confusion matrix in binary classification is shown in Table 1 below:

Tabel 1. Confusion Matrix

	Predicted negative	**Predicted positive**
Actual negative	True negative (TN)	False positive (FP)
Actual positive	False negative (FN)	True positive (TP)

After obtaining the confusion matrix values, calculations for accuracy, sensitivity, specificity, and precision can be performed. Sensitivity is used to calculate the proportion of true positives in predicting actual positive values, while specificity is used to calculate the proportion of true negatives in predicting actual negative values. These values are calculated using the following equations:

$$Accurancy = \frac{TP + TN}{TP + TN + FN + FP} \; x \; 100\% \quad (3)$$

$$Sensitivity = \frac{TP}{TP + FN} \; x \; 100\% \quad (4)$$

$$Specificity = \frac{TN}{TN + FP} \; x \; 100\% \quad (5)$$

$$Precision = \frac{TP}{TP + FP} \; x \; 100\% \quad (6)$$

Data Types and Research Variables

This study is of a secondary research type with data sourced from bank x in the year 2022. The sample data used consists of 526 current and non-current borrower data with 10 attributes as

independent variables, comprising limit, rate, tenor, total installment, age, salary, premium and admin, type, credit type and institution, and 1 attribute as a dependent variable, specifically, collectability (col). These attributes are utilized to identify risks by assessing customers' financial capabilities and obligations.

Descriptive statistical analysis is conducted for attributes limit, rate, tenor, total installment, age, salary, along with premium and admin because they consist of numerical data. Descriptive statistical analysis of these attributes is as follows:

Table 2. Descriptive Statistics

	Min	Max	Mean	Std. Dev
Limit	5.000.000	343.000.000	165.539.087	82.285.869
Rate	11,3	18,58	14,24591	1,313343
Tenor	12	220	152	40,16556
Total Installment	69.814,18	292.952.021,8	4.258.752,219	20.727.196,2
Age	46	75	62	4,136993
Salary	750.000	4.722.400	3.339.444	1.011.255
Premium and Admin	257.428	38.524.500	10.904.507	7.126.751

Based on Table 4.1, it can be seen that the smallest value of the variable limit is Rp. 5,000,000,- which means the minimum loan amount provided by the bank to debtors. The largest limit value from the data is Rp. 343,000,000,-. The smallest debtor salary in the data is Rp. 750,000,- and the highest debtor salary is Rp. 4,722,400,-. The fastest credit payment duration is 12 months (minimum tenor value) and the longest is 220 months.

The description of the attributes used in this study is as follows:

Table 3. Research Data Attributes

Atributes	Information
Limit	The maximum amount of credit that can be given by the bank to prospective debtor customers, in order to fulfill the credit agreement
Rate	The interest rate used as the basis for determining the loan interest that will be charged by the bank
Tenor	Credit term in months
Total Installment	Total installments for each period to be paid
Age	Age of debtor
Salary	Income of debtors
Premium and Admin	Fees payable by the debtor
Types	The purpose of credit use, consisting of working capital and consumption
Type of Credit	The type of credit taken, consists of new, take over, top up and top up single account
Institution	Credit agency, consisting of *taspen* and *asabri*
Coll	Credit collectibility is a classification of the status of the state of credit payment by the debtor

Analysis Phases

The analysis phases in this research are as follows: 1. Conducting data preprocessing. Data preprocessing consists of data selection, data cleaning, and data transformation. 2. Dividing the data into two parts: training data and testing data. Training data is used for model formation, and testing data is used to assess the accuracy of the model. 3. Normalizing the training and testing data using min-max normalization. 4. Finding the optimal k value to form the classification model. 5. Forming the KNN model without ROS using the training data and determining the classification prediction of prospective borrowers using the KNN model without ROS on the testing data. 6. Forming the KNN model with ROS using the training data that has been resampled with ROS to balance and determining the classification prediction of prospective borrowers using the KNN model with ROS on the testing data. 7. Comparing the formed models to see if the KNN model with ROS yields better results than the KNN model without ROS. 8. Determining the best model for classifying prospective borrowers and drawing conclusions.

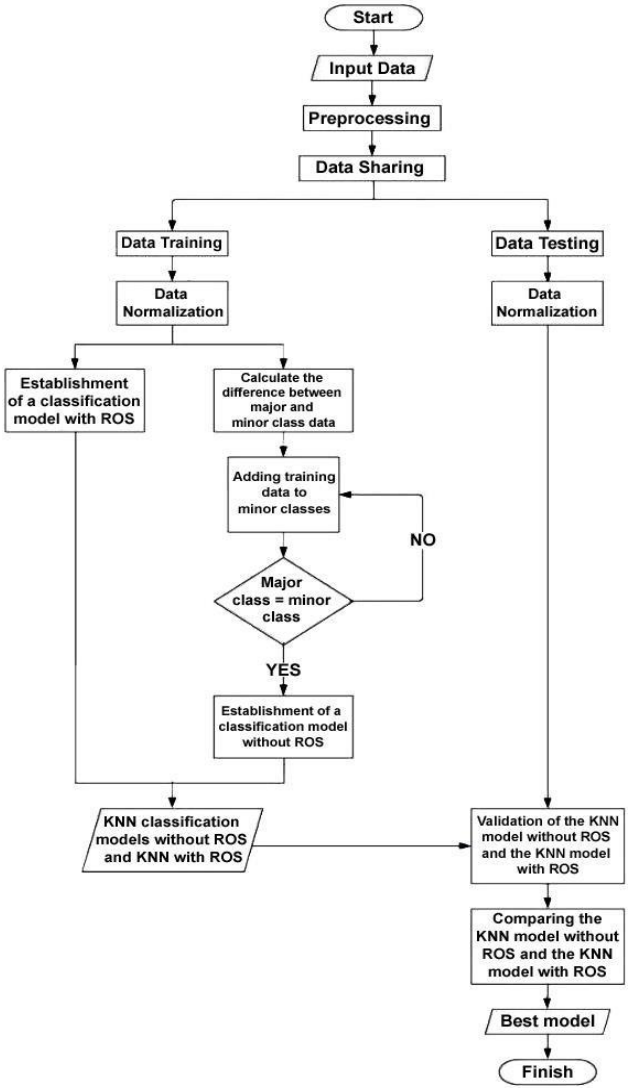

Figure 1. Research Flowchart

RESULTS AND DISCUSSION

In this study, a sample of 526 debtor data points was used, consisting of 500 current debtor data and 26 non-current debtor data.

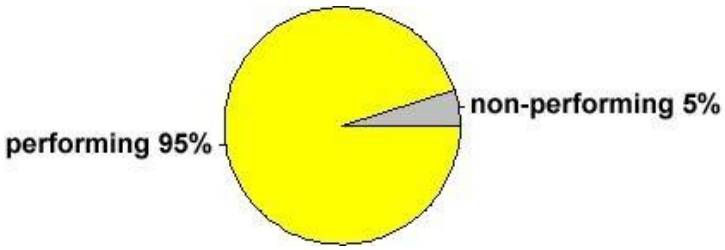

Figure 2. Debtor Status

Based on Figure 1, it can be observed that the status of current debtors is 95%, totaling 500 debtors, while the status of non-current debtors is 5%, totaling 26 debtors. The comparison of classes between current and non-current debtors in the collectibility data as the response variable indicates an imbalanced class.

Data Training and Testing Division

The initial step in the classification process is to divide the data into training data and testing data. Several proportions for data division are selected, namely 70% for training data and 30% for testing data, 80% for training data and 20% for testing data, and 90% for training data and 10% for testing data. This is aimed at determining the analysis results on the data proportions that yield the best results.

Before the data division process, a data randomization process is conducted to ensure that the data has an equal chance of being included in the training and testing datasets. Subsequently, a data normalization process is carried out with the aim of reducing the range of values in the data used in the analysis to prevent analysis errors (Martha et al., 2022).

K-Nearest Neighbor (KNN) without Random Oversampling (ROS)

The formation of the KNN model without ROS using a proportion of 90% training data and 10% testing data begins with determining the value of k to be used. The determination of the value of k is obtained by calculating the accuracy values from k=1 to k=5. The accuracy value with the optimum k is chosen to form the classification model. The results of these accuracy calculations are displayed in Table 3 below:

Table 4. Accuracy results from k=1 to k=5

k	Accuracy
1	90,566
2	88,679
3	**92,453**
4	92,453
5	92,453

Based on Table 3, the optimum value of k obtained is k=3. After obtaining the optimum value of k, the process continues by forming the KNN model without ROS with the training data. The confusion matrix of the KNN model without ROS is shown in the following Table 4.

Table 5. KNN classification results without ROS on training data with k=3

	Predicted Negative	**Predicted Positive**
Actual Negative	0	22
Actual Positive	1	450

The confusion matrix results in Table 4 show an accuracy value of 95%. The model can correctly classify current debtor data at a rate of 99% (sensitivity value). However, the model cannot correctly classify non-current debtor data, with a specificity value of 0%. Despite the excellent accuracy value obtained, the model can be considered inadequate. This is because the specificity value obtained is 0%, indicating that the KNN model without ROS cannot classify non-current debtor data at all.

Subsequently, the formed model using training data is used to predict the debtor status in the testing data. The confusion matrix results are shown in Table 5 as follows:

Table 6. KNN classification results without ROS on testing data with k=3

	Predicted Negative	**Predicted Positive**
Actual Negative	0	4
Actual Positive	0	49

The classification results of the model on the testing data in Table 5 show an accuracy value of 92%. The sensitivity value is 100% and the specificity value is 0%. This means that the KNN model without ROS is not suitable for classifying current and non- current debtor data even though it has an accuracy of 92%. The model's inability to classify non-current debtors may be due to an imbalance class issue in the data. This situation can worsen the classification results because the minority class (non-current debtors) is ignored by the model. As a result, the model only correctly classifies current debtor data. Results like these can pose risks to the bank as they can lead to losses and errors in policy-making.

K-Nearest Neighbor (KNN) with Random Oversampling (ROS)

K-Nearest Neighbor (KNN) with Random Oversampling (ROS) ROS implementation is used to enhance the minority data in the training dataset, as the imbalance between the minority class and the majority class exists. The random oversampling process is applied to the training data by randomly selecting data from the minority class and adding them to the training dataset. This process is repeated until the number of data in both the minority and majority classes in the training dataset becomes equal.

The results obtained with a 90% training data and 10% testing data split showed an equalization between the majority and minority classes, with a proportion of 0.5:0.5. The training data, which initially consisted of 473 entries, with 451 entries in the majority class and 22 entries in the minority class, changed to 902 entries after the random oversampling process, comprising 451 entries for current credit and 451 entries for non-current credit.

Following the ROS process, the formation of the KNN model with ROS was carried out with a value of k=3. The confusion matrix obtained is displayed in the following Table 6:

Table 7. Classification results of KNN with ROS on the training data with k=3

	Predicted Negative	Predicted Positive
Actual Negative	451	0
Actual Positive	41	410

The confusion matrix results in Table 6 show an accuracy value of 95.45%. The model can correctly classify current debtor data at a rate of 90% (sensitivity value). The model is also able to correctly classify non-current debtor data. The specificity value obtained has increased by 100%.

Subsequently, the formed model using training data is used to predict the debtor status in the testing data. The confusion matrix results are shown in Table 7 as follows:

Table 8. KNN classification results with ROS on testing data with k=3

	Predicted Negative	Predicted Positive
Actual Negative	1	3
Actual Positive	6	43

Based on Table 7, an accuracy value of 83.02% is obtained, which means that the ability of the KNN model with ROS to classify the creditworthiness of debtors is 83.02%. Based on the sensitivity value, it means that the KNN model with ROS can correctly classify current debtors at a rate of 89.80%. Meanwhile, based on the specificity value, the KNN model without ROS can classify non-current debtors at a rate of 25%. The ROS process increases the specificity value of the model by 25% from the previous 0% model. This means that the KNN model with ROS can classify non-current debtor data, although the obtained value is still quite low.

Comparison of Models at Various Data Split Proportions

The proportion of dividing training and testing data can affect the accuracy results of the obtained model. The data training and testing division process was carried out three times: 70% training data and 30% testing data, 80% training data and 20% testing data, and 90% training data and 10% testing data. The results obtained are shown in the following table:

Table 9. Comparison of the results of the KNN model without ROS and KNN with ROS on testing data

Data Sharing	*K*	Model	Accurancy	Sensitivity	Specificity	Precision
30% testing	3	KNN	93,67%	100%	0%	93,67%
	3	KNN+ROS	84,81%	90%	0%	93%
20% testing	3	KNN	92%	100%	0%	92%
	3	KNN+ROS	84,90%	90%	12%	92,70%
10% testing	3	KNN	92%	100%	0%	92%
	3	KNN+ROS	83,20%	85,75%	25%	93,47%

In Table 8, the values of accuracy, sensitivity, specificity, and precision for each obtained model are shown. It is known from the proportion division of training data and testing data that the KNN model without ROS (Random Over Sampling) cannot correctly classify non-performing debtors (with a specificity value of 0%). Meanwhile, the KNN model with ROS shows satisfactory results on the training data, and on the testing data, the specificity value improves compared to the KNN model without ROS, although still relatively small.

The KNN model without ROS can only effectively classify performing debtors. This is due to the class imbalance between performing and non-performing debtors. The minority class, or non-performing debtors, is overlooked by the model, resulting in a specificity value of 0%.

CONCLUSION

Based on the analysis and discussion of the implementation of random oversampling technique in the K-Nearest Neighbor method for credit feasibility analysis on imbalanced class data, it is concluded that the use of random oversampling technique can increase the specificity value of the classification model by 25% with an accuracy of 84.91%. The KNN method with ROS implementation has better results in each proportion division of training and testing data with a value of k=3. The KNN classification model with ROS can correctly predict data for non-performing debtors. Meanwhile, in the KNN classification model without ROS, the model cannot predict non-performing debtors at all. If the model only predicts performing debtors and cannot predict non-performing debtors, this can pose a risk to the bank in decision-making. Further research is suggested to analyze imbalanced data with other handling methods such as SMOTE or random undersampling (RUS).

REFERENCE

Adi, S., & Winarko, E. (2015). Klasifikasi Data NAP (Nota Analisis Pembiayaan) untuk Prediksi Tingkat Keamanan Pemberian Kredit (Studi Kasus: Bank Syariah Mandiri Cabang Luwuk Sulawesi Tengah). *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, 9(1), 1–12. https://doi.org/10.22146/ijccs.6635

Astuti, F. D., & Lenti, F. N. (2021). Implementasi SMOTE untuk mengatasi Imbalance Class pada Klasifikasi Car Evolution menggunakan K-NN. *Jurnal Jupiter*, 13(1), 89–98.

Bhatia, N. (2010). Survey of nearest neighbor techniques. *ArXiv Preprint ArXiv:1007.0085*, 8(2), 302–305. https://doi.org/10.48550/arXiv.1007.0085 Focus to learn more

Dahlan, I. A. (2022). Klasifikasi Cuaca Provinsi Dki Jakarta Menggunakan Algoritma Random Forest Dengan Teknik Oversampling. *Jurnal Teknoinfo*, 16(1), 87–92. https://doi.org/10.33365/jti.v16i1.1533

Ginting, J. A. (2019). Data Mining untuk Analisa Pengajuan Kredit dengan Menggunakan Metode Logistik Regresi. *Jurnal Algoritma, Logika Dan Komputasi*, 2(2), 164–169. https://doi.org/10.30813/j-alu.v2i2.1845

Hendrian, S. (2018). Algoritma Klasifikasi Data Mining Untuk Memprediksi Siswa Dalam Memperoleh Bantuan Dana Pendidikan. *Faktor Exacta*, 11(3), 266–276. https://doi.org/10.30998/faktorexacta.v11i3.2777

Martha, S., Andani, W., & Rizki, S. W. (2022). Perbandingan Metode k-Nearest Neighbor, Regresi Logistik Biner, dan Pohon Klasifikasi pada Analisis Kelayakan Pemberian Kredit. *Euler: Jurnal Ilmiah Matematika, Sains Dan Teknologi*, 10(2), 262–273. https://doi.org/10.34312/euler.v10i2.16751

Pramadhana, D. (2021). Klasifikasi Penyakit Diabetes Menggunakan Metode CFS Dan ROS dengan Algoritma J48 Berbasis Adaboost. *Edumatic: Jurnal Pendidikan Informatika*, 5(1), 89–98. https://doi.org/10.29408/edumatic.v5i1.3336

Saifudin, A., & Wahono, R. S. (2015). Pendekatan Level Data untuk Menangani Ketidakseimbangan Kelas pada Prediksi Cacat Software. *IlmuKomputer. Com Journal of Software Engineering*, 1(2), 76–85.

Santoso, B., Wijayanto, H., Notodiputro, K. A., & Sartono, B. (2017). Synthetic over sampling methods for handling class imbalanced problems: A review. *IOP Conference Series: Earth and Environmental Science*, 58(1), 12031. https://doi.org/10.1088/1755- 1315/58/1/012031

Wajhillah, R., Ubaidallah, I. H., & Bahri, S. (2019). Analisis Kelayakan Kredit Berbasis Algoritma K-Nearst Neighboar (Studi Kasus: Koperasi AKU). *InfoTekJar (Jurnal Nas. Inform. Dan Teknol. Jaringan)*, 4(1), 121–125. https://doi.org/10.30743/infotekjar.v4i1.1264

Whidhiasih, R. N., Wahanani, N. A., & Supriyanto, S. (2013). Klasifikasi Buah Belimbing Berdasarkan Citra Red-Green-Blue Menggunakan Knn Dan Lda. *PIKSEL: Penelitian Ilmu Komputer Sistem Embedded and Logic*, 1(1), 29–35.

Wicaksono, H. (2017). Penilaian Hasil Kegiatan Belajar Mahasiswa Menggunakan Metode Cluster Non-Hierarki. *Infoman's*, 11(1), 11–21. https://doi.org/10.24076/citec.2019v6i1.178

Wijayanti, N. P. Y. T., Kencana, E. N., & Sumarjaya, I. W. (2021). SMOTE: Potensi dan Kekurangannya pada Survei. *E-Jurnal Matematika*, 10(4), 235–240. https://doi.org/10.24843/MTK.2021.v10.i04.p348

Winata, T. A., Wiryawan, I. W., & Rudy, D. G. (2013). Kendala dalam Penyelesaian Kredit Macet pada PT. Bank Pembangunan Daerah Bali Cabang Denpasar. *Jurnal Ilmu Hukum*, 1(1), 1–9.