International Journal of Electrical and Computer Engineering (IJECE)
Vol. 13, No. 2, April 2023, pp. 1837~1847
ISSN: 2088-8708, DOI: 10.11591/ijece.v13i2.pp1837-1847

Supervised and unsupervised data mining approaches in loan default prediction

Jovanne C. Alejandrino[1], Jovito Jr. P. Bolacoy[1], John Vianne B. Murcia[2]
[1]Information Systems Program, Professional Schools, University of Mindanao, Davao City, Philippines
[2]College of Business Administration Education, University of Mindanao, Davao City, Philippines

Article Info

Article history:

Received Oct 20, 2021
Revised Sep 21, 2022
Accepted Oct 18, 2022

Keywords:

Data mining
Decision tree
k-nearest neighbor
Loan default prediction
Logistic
Naïve Bayes
Weka

ABSTRACT

Given the paramount importance of data mining in organizations and the possible contribution of a data-driven customer classification recommender systems for loan-extending financial institutions, the study applied supervised and supervised data mining approaches to derive the best classifier of loan default. A total of 900 instances with determined attributes and class labels were used for the training and cross-validation processes while prediction used 100 new instances without class labels. In the training phase, J48 with confidence factor of 50% attained the highest classification accuracy (76.85%), *k*-nearest neighbors (*k*-NN) 3 the highest (78.38%) in IBk variants, naïve Bayes has a classification accuracy of 76.65%, and logistic has 77.31% classification accuracy. *k*-NN 3 and logistic have the highest classification accuracy, F-measures, and kappa statistics. Implementation of these algorithms to the test set yielded 48 non-defaulters and 52 defaulters for *k*-NN 3 while 44 non-defaulters and 56 defaulters under logistic. Implications were discussed in the paper.

Corresponding Author:

John Vianne B. Murcia
Business Analytics Department, College of Business Administration Education, University of Mindanao
Davao City, Philippines
Email: jv_murcia@umindanao.edu.ph

1. INTRODUCTION

Technology is rapidly changing, and many organizations are adapting to such changes, including bank institutions. In today's business industry, a vast number of customers are engaged in banks and financial companies [1]. Today, financial institutions are progressively competitive [2]. Because technology advancements are one of the factors of economic success, it the reason why these organizations spend more on information processing and technology to gain more profits and to attract numerous clients [3]. Many bank services such as credit cards, automated teller machines (ATMs), e-banking, debit cards, and run through several platforms. These channels show a competitive advantage and an opportunity for making themselves ahead of their competitors and lead to an increase in the quality of various bank services [4].

Many companies today are engaged in data mining, even banks. It is a tool for utilizing big data sets by exploring patterns and structures and is anchored to data analytic disciplines like machine learning and pattern recognition [5]. Financial sectors use data mining for profitability, customer segmentation, tracing fraudulent transactions, checking high-risk loan applications, and the like [6]. Because of the large number of data available in every bank, customer relationship management (CRM) systems are established to manage and utilize these data from the customers to improve service functions, reorganize workflows, and update the institution's overall optimization [7], [8]. This marketing concept is widely used by most banks nowadays. Since the data are available in the CRMs, data mining allows extracting information from the available data

and predict the results of different scenarios. These results help top-level management to provide business decisions and increase customer intimacy and satisfaction [9].

Extending loans, on the other hand, is one of the important services of every bank that includes a high risk of clients' non-payment [10]. The bank will have no profit if there is no loan repayment that will deem a loan default. Financial, operational, legal, and credit risk immensely increase the borrowers' possibility of loan defaults because of poor credit risk management [11], [12]. Loan defaults drop the customer's credit score, increase the interest rates for future loans, and, worse, seize the client's assets as collateral of previous loans. The need for checking the customers' financial history and background is a crucial factor that every bank institution should consider before granting loans and credits to customers to reduce this risk [13]–[15]. Hence, credit risk management is one of the essential responsibilities for the banking sector's financial stability and liquidity [16]. Credit risk decisions are key factors for bank institutions' success since many losses result from wrong decisions and wrong credit loan approval [17].

This study aims to employ various useful data mining techniques for loan default prediction. Several algorithms are used to classify the datasets. The succeeding part of this paper provides a summary of related studies and definitions of the data mining techniques utilized, describes the methodology followed, shows the datasets used for the research and discusses the result and conclusion of the study.

2. RELATED WORKS

Data consists of raw, unrefined, and commonly unfiltered information. Data alone is useless until it will evolve into information [18], [19]. The data needs to be filtered, refined, and processed to make it more useful for some form of analysis [18], [20]. On the other hand, knowledge is the application of information that enables individuals and organizations to create valuable decisions and insights for decision support [20], [21]. The formation of the knowledge base and its application for businesses' benefits is becoming a strategic tool to compete and gain competitive advantages over their competitors [22]. Many enterprises, including the governments, corporations, and individuals, have collected a vast amount of data. However, the transformation of data into useful and valuable information is needed to discover insights and create knowledge and information-based decision-making [23], [24]. Knowledge discovery from data, or described as data mining, is the method of finding and extracting important information from the data gathered [25].

Since the 1990s, the need for modernization in terms of the information and technology infrastructure has been realized by the banking industry. They have made efforts to deliver their services more technically firm and customer oriented. They have shifted to centralized databases and made banking more accessible through ATMs and online transactions [22]. Data mining is one of the important techniques banks used to discover knowledge from databases. Business intelligence tools have played a vital role in supporting the banking sector, such as increasing customer retention, profitability, market penetration, and efficiency. Data mining, as defined by Gartner Group, is "the process of discovering significant new patterns, correlations, and trends by filtering through large amounts of data stored in repositories and by the application of pattern recognition technologies as well as statistical and mathematical techniques" [26]. In most circumstances, these insights are driven by the study and analyses of historical data [27].

Many types of research have been conducted in the financial and banking sector using data mining. Sudhakar *et al*. [28] have shown a two-step loan credibility prediction system that aids the bank and other financial institutions decide whether to approve or reject the borrowers' loan request. The decision tree induction data mining algorithm is applied to predict the attributes relevant to customer credibility.

Hamid and Ahmed [6] presented a new model for classifying the risks of loan in the banking sector employing data mining. The model aims to predict the standing of loans from the banking sector. The proposed model made use of the J48, Bayes Net, and naive Bayes algorithms. The study found out the J48 algorithm has a higher accuracy among all three algorithms.

Zurada and Zurada [29] used data mining techniques, namely decision trees, logit regression, neural networks, and ensemble models, to predict creditors to default or pay off their loans using their financial attributes. In all data mining algorithms used in the study, the ensemble model and neural network yielded the best in classifying loans into either a good or bad loan. Islam *et al*. [30] constructed a model using the multilayer perceptron algorithm of neural network to predict the probability that a given creditor will default a credit or loan. The predictive model yielded an 83.86% success rate than the other techniques like logistic regression and discriminant analysis, which both gained a 76.4% success rate. Another study conducted to a rural bank in Indonesia [31] used data mining to suggest a model that will improve the financial institution's credit risk assessment and lessen non-performing loans to not more than 5%. The study made use of the decision tree C5 algorithm to classify performing or non-performing loan risks.

The study by Lahsasna *et al*. [17] about predicting loan default introduced a loan default prediction model based on the random forest algorithm. The study's experimental result shows that the random forest

algorithm has a higher prediction with 98% accuracy than the decision tree, support vector machine (SVM), and logistic regression algorithms, which only gained 95%, 75% 73% accuracy, respectively Vimala and Sharmili [32] proposed using data mining to predict the status of loans. The study used naïve Bayes and SVM algorithms using the UCI germen credit dataset. The study's comparative result shows that SVM has higher accuracy compared to the naïve Bayes with 79% and 77% accuracy, respectively.

Five data mining algorithms are used [33] for loan default risk analysis that enables banks to reduce loan defaults. The Bayesian algorithm, decision tree, boosting, bagging, and random forest algorithms are applied to different data sets of various sizes. The random forest algorithm is the most consistent and has the highest accuracy compared to the others.

Hassan and Abraham [34] produced a prediction model to predict load default using three training algorithms. The algorithms used are scaled conjugate gradient (SCG) backpropagation, one-step secant (OSS) backpropagation, and Levenberg-Marquardt (LM) algorithm. The result of the study indicated that there is a significant improvement in the training algorithm design of the loan prediction model. The study also found out that the ensemble models also work better than the other models.

3. PROBLEM DEFINITION

The lack of clearcut basis for financial institutions in the Philippines to predict loan performance remains a challenge for these institutions in terms of collection efficiency and profitability. With high loan default incidences leading to low efficiency in loan collection, financial institutions have realized that these could have been anticipated by using recommender systems based on available relevant information within their database. However, the lack of these systems is common in the Philippines, as financial institutions are challenged with strict laws like Republic Act No. 3765 (otherwise known as Truth in Lending Act) and Republic Act No. 10173.

Despite these challenges, management of financial institutions believed that loan defaults could be predicted using recommender systems. Such innovations could be driven by machine learning approaches, but these are uncommon in the country. Having these approaches at hand, this would entail the management to use crucial customer data attributes could contribute to loan default likelihood. Having such systems could be helpful, as they would provide preliminary outcomes of a future event based on available historical client data. With systems like this, investors and clients alike are assured of security and stability of the institution's safeguarding policies leading to profit maximization. For these reasons, this study proposes solutions that aim of helping loan-extending institutions.

4. RESEARCH METHOD

4.1. Dataset

Data on loan default was provided with consent by a loan-extending agency (a multipurpose cooperative) located in Davao City, Philippines. As shown in Table 1, the loan default dataset contained 29 attributes, which included 27 explanatory attributes, 1 class attribute, and 1 attribute for ID. As agreed with the dataset provider, 1,000 instances must be extracted from the dataset for the analysis, wherein 900 were used for training and cross-validation while 100 were used for prediction as a test set. The selected test set has its actual class labels removed.

4.2. Data preparation

Visual inspection of the dataset further revealed that missing values are found both in the training set (first 900 instances) and test set (last 100 instances). To address this, an unsupervised instance filter called replace missing values was utilized, replacing missing instances with mean for numeric attributes and mode for nominal attributes. This step is performed once the data types have been identified for each of the attributes. The *ignoreClass* parameter was set at the default (*False*) so that the 100 missing class labels set in the test set will not be imputed with values. Imputation of missing entries among instances was done after the proper assignment of data type for each of the attributes in the dataset. The unsupervised instance filter replaces missing values will replace missing instances under numeric attributes with mean and mode for those nominal attributes.

Irrelevant attributes can greatly affect the performance of most classifiers. An attribute with a lot of missing values, or those attributes with only one distinct value, can be considered irrelevant, as they provide no variations towards the target attribute (i.e., class) to be explained or classified [35]. In this case, there are two attributes which qualified as useless attributes (coded A12 and A13). A13 has 1,000 instances with one distinct response (F) while A12 has 999 instances with one distinct response (T). They must be removed as there are no variations among the instances that could account for a better explanation or attribution to the class attributes. Also, while it is understood that instances under ID are merely used to serve as markers or

cases for instances of much important attributes, this was also removed so that it could be excluded by mistake as an explanatory attribute in order not to meet any problems in the classification stage. This leaves us having 25 explanatory attributes and 1 class attribute.

Table 1. List of attributes of the loan prediction dataset

Attribute	Type	Description	Transformation Required
A1	Numeric	Amount of loan applied in pesos.	Yes
A2	Numeric	Number of months the loan will be payable, i.e., 12, 24 or 36 months	No
A3	Numeric	Interest rate for the loan in percent	No
A4	Nominal	Source of income categories, possible values range from 0 to 5.	Yes
A5	Nominal	Industry or nature of business where the loan applicant works or engages with, which includes a list of industries, coded from 1 to 36, and 99 (others).	Yes
A6	Nominal	*Barangay* (village) where the loan applicant resides, coded from 1 to 182.	Yes
A7	Nominal	*Barangay* (village) where the loan applicant does business, coded from 1 to 182.	Yes
A8	Numeric	Monthly income earned from employment or engagement with business/trade in pesos	Yes
A9	Numeric	Estimated monthly expenditures of the loan applicant in pesos.	Yes
A10	Nominal	If the loan application was applied for the first time (Y) or not (N).	No
A11	Nominal	If collateral is required for the loan, coded Y (Yes) or N (No).	No
A12	Nominal	If the loan application has been subjected to background investigation by a credited officer– marked T if tagged (completed), and F for further referral.	No
A13	Nominal	If all necessary client information are filled out – marked F if full, and M if some are missing.	No
A14	Numeric	Current amount of fixed capital share or deposits in pesos.	Yes
A15	Numeric	Number of years where loan applicant's membership was approved.	No
A16	Nominal	If the loan has another co-maker (Y/N).	No
A17	Nominal	*Barangay* (village) where the co-maker resides, coded from 1 to 182.	Yes
A18	Nominal	*Barangay* (village) where the co-maker does business, coded from 1 to 182.	Yes
A19	Nominal	Purpose of the loan, coded from 1 to 12 for common reasons, and 99 (others).	No
A20	Nominal	If source of income of loan applicant has attached documents or is verified during background investigation (Y/N).	No
A21	Nominal	If attachments were pre-checked (Y/N).	No
A22	Nominal	If the loan applicant has another residence outside the city (Y/N).	No
A23	Nominal	If late payment fees/penalties were paid and filed (Y/N).	No
A24	Nominal	If loan amount needs endorsement from several members of the Board, coded from 1 to 5.	No
A25	Numeric	Number of days since loan was filed.	No
A26	Nominal	If loan applicant has defaulted or incurred late payment prior to this application (Y/N).	No
A27	Nominal	If any of the co-makers/guarantors are coop members (Y/N).	No
A28	Nominal	If any of the co-makers/guarantors have defaulted in the past (Y/N).	No
A29	Nominal	Initial decision on the loan application by the coop manager.	No
Class	Nominal	Classification whether the loan application has defaulted (1) or otherwise (0).	No

To convert numeric attributes to nominal attributes, an unsupervised attribute filter called numeric to nominal was used. In the *attributeIndices* parameter, the attribute numbers of numeric attributes corresponding to A4, A5, A6, A7, and Class were identified before their conversion to nominal data type. After conversion, class attribute, now assuming nominal data type, changed its colors from black and white to blue (class 1 with 0 as label) and red (class 2 with 1 as label). All in all, there are 18 nominal attributes including the class attribute, and 8 numeric attributes in the dataset.

4.3. Data normality

Several classifier algorithms in Weka assume that numeric attributes have instances following Gaussian distribution [36]. With this, an unsupervised instance filter Normalize was implemented, which rescales the numeric attributes to values between 0 and 1. Figure 1 shows the descriptive statistics for attribute A10 before and after the normalization process. As seen in the first picture, the mean value was 3,271.3 (*SD*=2,822.7). The standard deviation is too large, which means that the values of the instances are spread too much, which may pose less reliable prediction performance [37]. Normalizing the same now produces a different result, having the instances with values from 0 to 1. The mean value of the normalized dataset for this attribute is 0.166 (*SD*=0.155). The same is done for the remaining seven numeric attributes.

4.4. Feature selection

To ensure that relevant attributes are included prior to the classification procedure, feature selection was carried out with the use of three most prominent feature selection algorithms in Weka [38]. Following this suggestion, the algorithms selected are correlation-based feature selection (*CorrelationAttributeEval*), information gain-based or entropy-based feature selection (*InfoGainAttributeEval*), and learner-based feature selection (*WrapperSubsetEval*).

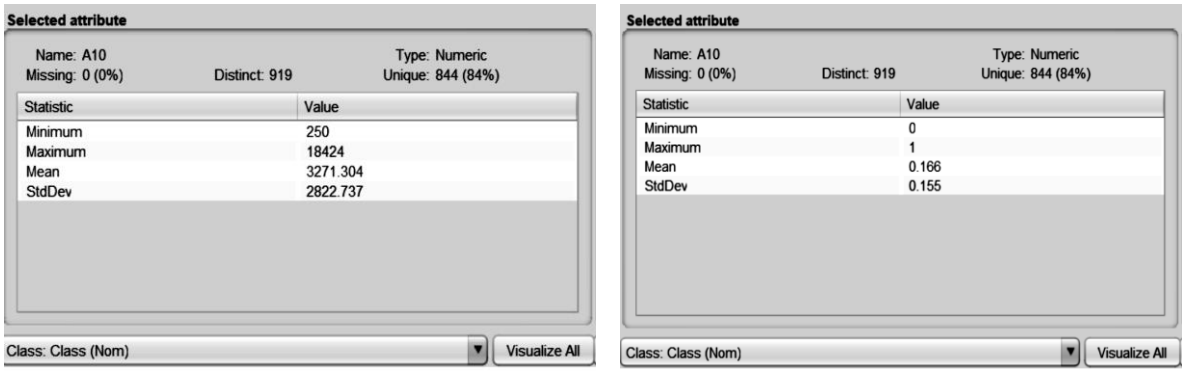

Figure 1. A10 attribute before and after applying normalize algorithm

Results of the correlation-based feature selection revealed that the five most-correlated attributes to the class attribute are A21 (r=0.219), A8 (r=0.217), A10 (r=0.169), A24 (r=0.118), and A30 (r=0.117), while the five least-correlated are A19 (r=0.008), A18 (r=0.009), A7 (r=0.0098), A5 (r=0.010) and A27 (r=0.014). It was suggested that attributes with r-values of 0.20 above can be retained to have an optimal classification performance. However, only two attributes (A21 and A8) passed this criterion.

With regards to information gain (entropy)-based feature selection, the five highest-ranked attributes are A21 (r=0.082), A22 (r=0.046), A8 (r=0.037), A23 (r=0.023) and A10 (r=0.023), while the attributes in the bottom five are A20, A16, A19, A17 and A18, each garnering 0 as entropy values. It was suggested that an arbitrary entropy/information gain cut-off should be 0.05, yet only one attribute (A21) fulfilled this criterion.

Results of the learner-based feature selection which utilizes the *WrapperSubsetEval* attribute evaluator with *BestFit* search method. This has the capability to select the attributes that are automatically selected, likened to suppression of variables of stepwise regression procedure. The results revealed that of the 25 attributes to be used to classify the class attribute, only four are likely usable (A9, A10, A22, and A24) at 73.8% merit of the 197 subsets found in the training set. Adjusting the search method parameter to *GreedyStepwise* further revealed the same four usable attributes likely to be included in the classification procedure. Based on the three feature selection methods undertaken, two attributes are eliminated–coded A19 and A18–as both appeared in the least-correlated and least-ranked entropy-based features, both of which are not included also in the learner-based feature selection.

4.5. Data imbalance

The training set appears to have an imbalance of class attributes. There were 250 instances of class label 0 while there are 650 instances of class label 1. Clearly, class label 0 is the minority class. The danger of class imbalance is that conventionally, algorithms tend to become biased by predicting the overall accuracy towards the class with bigger observations [39]. To address this, we utilized a supervised instance filter called synthetic minority oversampling technique (SMOTE). The idea of SMOTE is to synthesize new data from available instances to boost classification/prediction accuracy from a limited data volume [40]. The oversampling was implemented, having in mind the need to match the number of instances for both classes. Hence, the need to increase the 250 zero-labeled class to 650 requires the addition of 400 instances to be at par with the one-labeled class, a 160% increase of instances in the minority. Upon the implementation of SMOTE, synthetic instances alongside the needed 400 instances for the zero-labeled class were inserted at the end of the dataset. To this effect, there are a total of 1,300 instances for the training set. This number is a candidate for 13 folds in the subsequent cross-validation process, where each fold requires 100 instances to develop prediction models for the training set.

The SMOTE filter produced 650 instances each for the two classes. However, to achieve a better classification performance, it is suggested that each of the folds shall have an equal number of instances from both classes, in this case, each fold should have 50 instances of all attributes from the zero-labeled class and 50 instances of the same attributes from the one-labeled class. This is to ensure better performance of the classifiers and to have a much more reliable result in training the dataset. To do this, an unsupervised attribute filter called randomize can be useful, such that it randomly reorders the training data which contains the original as well as synthetic instances generated in the previous SMOTE oversampling procedure. The idea of a randomized filter is to ensure that each of the folds in the cross-validation receives a well-represented number of classes to avoid overfitting during classification [41].

4.6. Data classification and cross-validation

Three classifiers were selected to perform the classification of the training dataset as well as the prediction of class label in the test set. We employed the most common classifiers: *k*-NN, naïve Bayes, and decision trees (J48). Also, since the class labels to be predicted come in binary (0 and 1), a fourth classifier, logistic, was used. Every classifier typically has parameters to tune. In this phase, parameter tunings were implemented in the number of *k* in *k*-NN and on the confidence factor of the decision tree (J48). In *k*-NN, the higher the value of k, the lesser the chance of error [42], while confidence factor in J48 is the parameter altered to test the effectiveness of post-pruning [43]. In *k*-NN, classifications were done in *k*-NN 3, *k*-NN 5, *k*-NN 7, *k*-NN 9, *k*-NN 11, and *k*-NN 13. We decided to exclude the default *k*-NN 1 (or simple *k*-NN) because test examples tend to be given a similar label as the closest example in the training set [44]. The *k* was set to start at *k*=3 so that three closest classes will be checked, and the most common label is assigned; same is true to higher *k*. On the other hand, confidence factor in J48 classifier was set at an increment of 0.25 from the default value of 0.25. Four classification procedures were done (C.0.25, C.0.5, C.0.75 and C.1.0).

An issue of overfitting in the performance of cross-validation of the training set can greatly affect the performance of the classifiers, which may even affect the trustworthiness of the prediction phase. While there are two usual methods of cross-validation, we preferred oversampling and randomization of the instances instead of splitting the original training set to two and making the second the validation set. As stated in the processes we did in addressing data imbalance and data instances issues, class imbalance was addressed through oversampling filter SMOTE enabling the class minority to synthetically growth the number of instances by 160% to match with the one-labeled class. To ensure that both classes are represented in each fold in cross-validation, the unsupervised filter randomize was used to randomly reorder the original 900 training data and the 400 new synthetically created data. For 1,300 instances with 50 zero-labeled and 50 one-labeled classes in each fold, cross-validation was carried out in 13 folds.

5. RESULTS AND DISCUSSION

5.1. Classification accuracy

There are 11 cross-validations conducted using the four classifiers: four using J48, five using IBk (*k*-NN), one for naïve Bayes and one for logistic. Table 2 shows the percentage of correctly classified instances of the 12 cross-validation runs. Based on the cross-validations runs on the training dataset, the J48 with confidence factor of 50% attained the highest classification accuracy (76.85%) among its J48 variants, and *k*-NN 3 was the highest (78.38%) among the IBk variants. Naïve Bayes has a classification accuracy of 76.65% while logistic has a 77.31% classification accuracy.

Table 2. Classification accuracy of the classifiers and their variants on training dataset

Classifier	Variants	Correctly Classified Instances
Decision Tree (J48)	C.0.25	991 (76.23%)
	C.0.50	999 (76.85%)
	C.0.75	994 (76.46%)
	C.1.0	994 (78.38%)
lBk (k-NN)	3	1019 (78.38%)
	5	985 (75.77%)
	7	986 (75.85%)
	9	988 (76%)
	11	978 (75.23%)
Naïve Bayes		995 (76.54%)
Logistic		1005 (77.31%)

As an addendum to Table 1, Figure 2 provides information on the confusion matrices of all 11 cross-validation runs revealed the distribution of correctly and incorrectly classified class labels. It can be remembered that there are 650 instances with class labels of 0 and 650 instances with 1 as class label. Taking the J48 cross-validation runs, it can be seen in J48 (0.25) that of the 650 instances, 495 zero-labeled instances were correctly classified under class label 0 while its 155 instances were predicted to fall under class label 1. Moreover, of its 650 one-labeled class instances, 496 were correctly classified to fall under class label 1, while 154 class 1 instances were classified under class label 0. In J48 (0.5), 612 out of 650 zero-labeled class instances were correctly classified while 613 out of 650 one-labeled class instances were correctly classified. As for J48 (0.75) and J48 (1.0), there are 503 zero-labeled and 491 one-labeled instances correctly classified, respectively. Also, Kappa statistics and mean absolute error are shown in Figure 3. Based on the

classification runs, the J48 classifier set at 0.50 confidence factor has the highest Kappa statistic while in terms of mean absolute error, J48 C.0.75 and C.1.0 have the least mean absolute error.

J48 (0.25)

a	b	<-- classified as
495	155	a = 0
154	496	b = 1

J48 (0.5)

a	b	<-- classified as
612	38	a = 0
37	613	b = 1

J48 (0.75)

a	b	<-- classified as
503	147	a = 0
159	491	b = 1

J48 (1.0)

a	b	<-- classified as
503	147	a = 0
159	491	b = 1

***k*-NN 3**

a	b	<-- classified as
600	50	a = 0
231	419	b = 1

***k*-NN 5**

a	b	<-- classified as
588	62	a = 0
253	397	b = 1

***k*-NN 7**

a	b	<-- classified as
591	59	a = 0
255	395	b = 1

***k*-NN 9**

a	b	<-- classified as
593	57	a = 0
255	395	b = 1

***k*-NN 11**

a	b	<-- classified as
597	53	a = 0
269	381	b = 1

NaiveBayes

a	b	<-- classified as
494	156	a = 0
149	501	b = 1

Logistic

a	b	<-- classified as
499	151	a = 0
144	506	b = 1

Figure 2. Confusion matrices for different algorithms used in the study

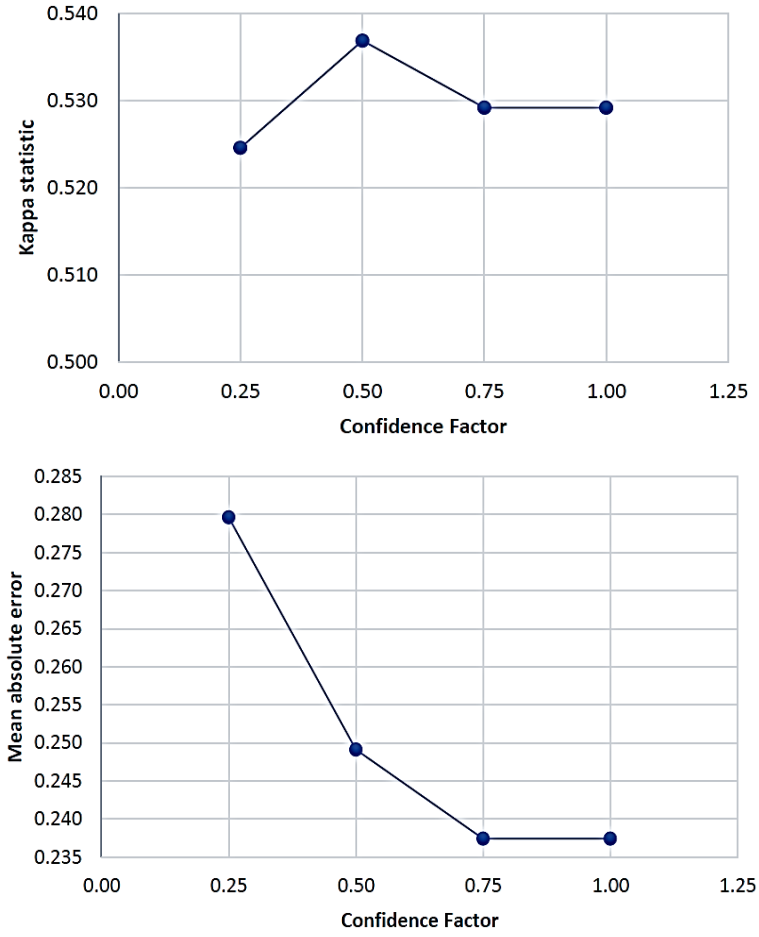

Figure 3. Kappa statistic and mean absolute errors for J48 classifiers

As for IBk (*k*-NN) cross-validations runs, the following are the results: at *k*=3, 600 zero-labeled and 419 one-labeled instances were correctly classified; at *k*=5, 588 zero-labeled and 397 one-labeled instances correctly classified; at *k*=7, 591 zero-labeled and 395 one-labeled instances correctly classified; at *k*=9, 593 zero-labeled and 395 one-labeled instances correctly classified; and at *k*=11, 597 zero-labeled and 381 one-labeled instances correctly classified. Also, average F-measure and mean absolute error are shown in Figure 4. Based on the classification runs, *k*-nearest neighbor of 3 has the highest average F-measure statistic and the least mean absolute error.

Lastly, using naïve Bayes, there are 494 zero-labeled and 501 one-labeled instances correctly classified while in logistic, there are 499 zero-labeled and 506 one-labeled instances correctly classified. With this, it is safe to assume that under J48 classifier, the one with 0.50 confidence factor has the best classification accuracy. For IBk, the classifier with the best classification accuracy is *k* nearest neighbor of 3.

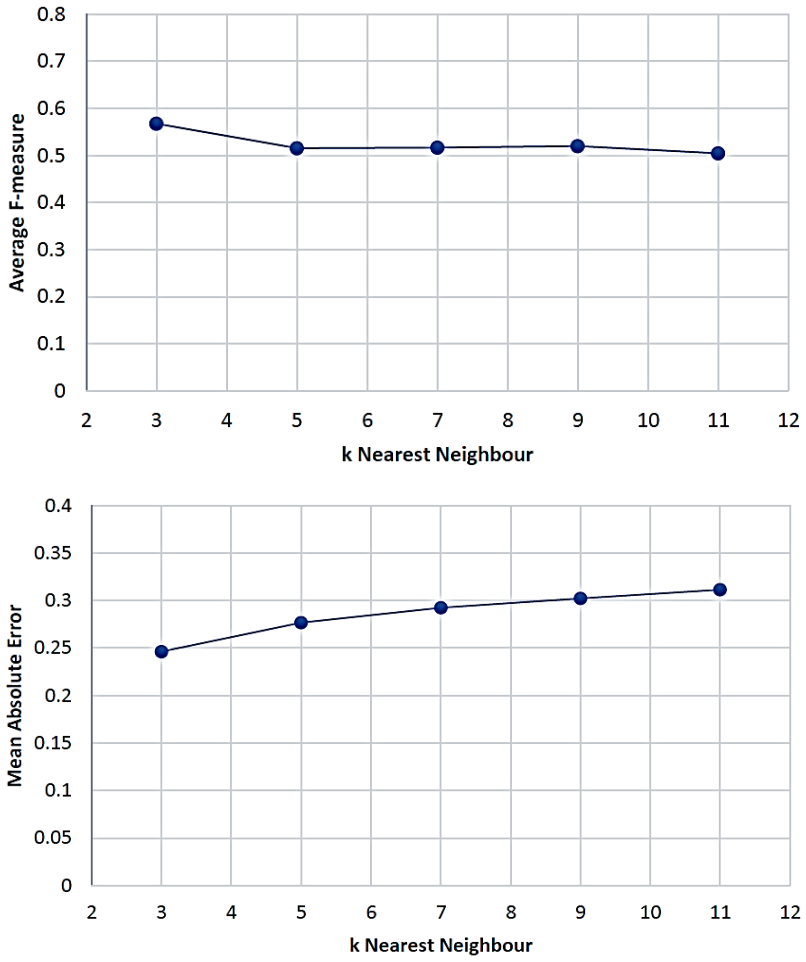

Figure 4. Average F-measure and mean absolute errors for lBk classifiers

5.2. Classifier comparison

Three factors were considered in assessment whether cross-validation of the training set are effective in each of the classifiers as well as their parameter-tuned sub-analyses: average F-measure, number of correctly classified instances, and the value of kappa statistic. The four classifiers were assessed including the variation of the runs when there were parameter tunings made on *k* nearest neighbor (IBk) and decision trees (J48). Table 3 shows the summary of the classification performance results. As previously stated, the J48 with confidence factor of 0.50 was found to have the greatest number of correctly classified instances, the highest kappa statistic as well as the highest F-measure among the other J48 variants. As for IBk (*k* nearest neighbors), *k*-NN 3 has the greatest number of correctly classified instances, the highest F-measure as well as kappa statistic. Increasing the number of k noticeably decreased the F-measure, number of correctly classified instances and kappa statistic, although it increased at *k*-NN 9.

Among the 11 classifications, naïve Bayes has an F-measure of 0.765, has 995 correctly classified instances and kappa statistic value of 0.5308, while logistic has a F-measure of 0.773, has 1,005 correctly classified instances and kappa statistic value of 0.5462. Based on the comparison, it appears that *k*-NN 3 and logistic have the highest number of correctly classified instances which instantaneously means the highest classification accuracy, the highest F-measure, and the highest kappa statistics. This means that the two can now be used to predict the test set's class labels (1 for default and 0 otherwise).

Table 3. Classification performance of the utilized classifiers and their variants on training dataset (*N*=1300)

Classifier	Variants	F-measure	Correctly Classified Instances (*f*)	κ
Decision Tree (J48)	C.0.25	0.762	991	0.5246
	C.0.50	0.768	999	0.5369
	C.0.75	0.765	994	0.5292
	C.1.0	0.765	994	0.5292
lBk (k-NN)	3	0.780	1,019	0.5677
	5	0.752	985	0.5154
	7	0.753	986	0.5169
	9	0.754	988	0.5200
	11	0.745	978	0.5046
Naïve Bayes	-	0.765	995	0.5308
Logistic	-	0.773	1,005	0.5462

5.3. Prediction results

Using the developed models under *k* nearest neighbor of 3 and logistic in the previous section, the algorithms were used to assess the model in the test set, which was the last 100 instances of the original supplied **.csv* file with unlabeled classes. The guiding contention to ensure that the best classifier will be chosen is that the number of predicted classes should be close to 50 zero-labeled and 50 one-labeled classes. Hence, whichever has the best approximation will be selected in addition to the previous selection criteria in the previous section.

Upon the implementation of the algorithms, *k*-NN was able to predict 48 instances with 0 as class label and 52 with 1 as class label. On the other hand, logistic was able to predict 44 instances with 0 as class label and 56 with 1 as class label. This implies that with respect to the criteria in the previous selection, a better classification accuracy would result in better prediction accuracy.

6. CONCLUSION

The study was able to implement different supervised and unsupervised data mining algorithms to identify the best classifier of a given loan default dataset. Results revealed that J48 with 0.50 confidence factor has the best classification accuracy among its variants; for IBk, the classifier with the best classification accuracy is k nearest neighbor of 3. *k*-NN 3 and logistic have the highest number of correctly-classified instances which instantaneously means the highest classification accuracy, the highest F-measure, and the highest kappa statistics. Implementation of these algorithms to the loan default test dataset yielded 48 zero-labeled and 52 one-labeled class instances for k-NN 3 while 44 zero-labeled class instances and 56 one-labeled class instances under logistic. For future works, it is recommended that the implemented classifiers will be applied to bigger datasets to further validate their accuracy. Recommender systems can be developed coming from this trajectory.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude to the University of Mindanao Research and Publication for the research grant used in the conduct and publication of this study.

REFERENCES

[1] D. A. Kumar and V. Ravi, "Predicting credit card customer churn in banks using data mining," *International Journal of Data Analysis Techniques and Strategies*, vol. 1, no. 1, 2008, doi: 10.1504/IJDATS.2008.020020.

[2] A. Cockrill, M. M. H. Goode, and A. Beetles, "The critical role of perceived risk and trust in determining customer satisfaction with automated banking channels," *Services Marketing Quarterly*, vol. 30, no. 2, pp. 174–193, Mar. 2009, doi: 10.1080/15332960802619231.

[3] A. Z. Siam, "Role of the electronic banking services on the profits of Jordanian banks," *American Journal of Applied Sciences*, vol. 3, no. 9, pp. 1999–2004, Sep. 2006, doi: 10.3844/ajassp.2006.1999.2004.

[4] J. Thornton and L. White, "Customer orientations and usage of financial distribution channels," *Journal of Services Marketing*, vol. 15, no. 3, pp. 168–185, Jun. 2001, doi: 10.1108/08876040110392461.

[5] D. J. Hand and N. M. Adams, "Data mining," in *Wiley StatsRef: Statistics Reference Online*, Wiley, 2015, pp. 1–7.
[6] A. J. Hamid and T. M. Ahmed, "Developing prediction model of loan risk in banks using data mining," *Machine Learning and Applications: An International Journal*, vol. 3, no. 1, pp. 1–9, Mar. 2016, doi: 10.5121/mlaij.2016.3101.
[7] B. Fang and S. Ma, "Data mining technology and its application in CRM of commercial banks," in *2009 First International Workshop on Database Technology and Applications*, Apr. 2009, pp. 243–246, doi: 10.1109/DBTA.2009.128.
[8] K. Chitra and B. Subashini, "Data mining techniques and its applications in banking sector," *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, no. 8, pp. 219–226, 2013.
[9] K. I. Moin and Q. B. Ahmed, "Use of data mining in banking," *International Journal of Engineering Research and Applications (IJERA)*, vol. 2, no. 2, pp. 738–742, 2012.
[10] S. Moradi and F. Mokhatab Rafiei, "A dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks," *Financial Innovation*, vol. 5, no. 1, Dec. 2019, doi: 10.1186/s40854-019-0121-9.
[11] L. Zhu, D. Qiu, D. Ergu, C. Ying, and K. Liu, "A study on predicting loan default based on the random forest algorithm," *Procedia Computer Science*, vol. 162, pp. 503–513, 2019, doi: 10.1016/j.procs.2019.12.017.
[12] E. Richard, M. Chijoriga, E. Kaijage, C. Peterson, and H. Bohman, "Credit risk management system of a commercial bank in Tanzania," *International Journal of Emerging Markets*, vol. 3, no. 3, pp. 323–332, Jul. 2008, doi: 10.1108/17468800810883729.
[13] H. A. Bekhet and S. F. K. Eletter, "Credit risk assessment model for Jordanian commercial banks: Neural scoring approach," *Review of Development Finance*, vol. 4, no. 1, pp. 20–28, Jan. 2014, doi: 10.1016/j.rdf.2014.03.002.
[14] M. Tiwari, "Data mining: a competitive tool in retail industries," *Global Journal of Enterprise Information System*, vol. 2, no. 2, pp. 78–88, 2010, doi: 10.18311/gjeis/2010/3020.
[15] Z. Alomari and D. Fingerman, "Loan default prediction and identification of interesting relations between attributes of peer-to-peer loan applications," *New Zealand Journal of Computer-Human Interaction*, vol. 2, no. 2, pp. 1–21, 2017.
[16] Z. Ereiz, "Predicting default loans using machine learning (OptiML)," in *2019 27th Telecommunications Forum (TELFOR)*, Nov. 2019, pp. 1–4, doi: 10.1109/TELFOR48224.2019.8971110.
[17] A. Lahsasna, R. N. Ainon, and T. Y. Wah, "Credit scoring models using soft computing methods: A survey," *The International Arab Journal of Information Technology*, vol. 7, no. 2, pp. 115–117, 2010.
[18] R. A. Evans, "Information vs. data," *IEEE Transactions on Reliability*, vol. R-30, no. 1, Apr. 1981, doi: 10.1109/TR.1981.5220933.
[19] J. Harris, "Data is useless without the skills to analyze it," Harvard Business Review Home, 2012.
[20] A. Targowski, "From data to wisdom," *Dialogue and Universalism*, vol. 15, no. 5, pp. 55–71, 2005, doi: 10.5840/du2005155/629.
[21] I. Monarch and J. Matthews, "Perspectival review," *Management Learning*, vol. 32, no. 4, pp. 518–524, Dec. 2001, doi: 10.1177/1350507601324008.
[22] M. Gobindgarh, "Application of data mining in banking sector," *International Journal of Computer Science and Technology*, vol. 4333, pp. 199–202, 2011.
[23] V. Grover, R. H. L. Chiang, T.-P. Liang, and D. Zhang, "Creating strategic business value from big data analytics: A research framework," *Journal of Management Information Systems*, vol. 35, no. 2, pp. 388–423, Apr. 2018, doi: 10.1080/07421222.2018.1451951.
[24] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing," *ACM Computing Surveys*, vol. 42, no. 4, pp. 1–53, Jun. 2010, doi: 10.1145/1749603.1749605.
[25] C. E. Lopez Guarin, E. L. Guzman, and F. A. Gonzalez, "A model to predict low academic performance at a specific enrollment using data mining," *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, vol. 10, no. 3, pp. 119–125, Aug. 2015, doi: 10.1109/RITA.2015.2452632.
[26] Y. Shi, L. Zhang, Y. Tian, and X. Li, "Data mining and knowledge management," in *Proceedings of Air Forum*, 2015, pp. 1–11, doi: 10.1007/978-3-662-46193-8_1.
[27] M. B. Hammaw, "Data mining for banking and finance," *Oriental Journal of Computer Science and Technology*, vol. 4, no. 2, pp. 273–280, 2011.
[28] M. Sudhakar and C. V. K. Reddy, "Two step credit risk assesment model for retail bank loan applications using decision tree data mining technique," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, vol. 5, no. 3, pp. 705–718, 2016.
[29] J. Zurada and M. Zurada, "How secure are good loans: Validating loan-granting decisions and predicting default rates on consumer loans," *Review of Business Information Systems (RBIS)*, vol. 6, no. 3, pp. 65–84, Jul. 2002, doi: 10.19030/rbis.v6i3.4563.
[30] S. Islam, L. Zhou, and F. Li, "Application of artificial intelligence (artificial neural network) to assess credit risk: a predictive model for credit card scoring," 2009.
[31] I. G. N. N. Mandala, C. B. Nawangpalupi, and F. R. Praktikto, "Assessing credit risk: an application of data mining in a Rural Bank," *Procedia Economics and Finance*, vol. 4, pp. 406–412, 2012, doi: 10.1016/S2212-5671(12)00355-3.
[32] S. Vimala and K. C. Sharmili, "Prediction of loan risk using Naive Bayes and support vector machine," *International Conference on Advancements in Computing Technologies*, vol. 4, no. 2, pp. 110–113, 2018.
[33] A. A. Sawant and P. M. Chawan, "Comparison of data mining techniques used for financial data analysis," *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, no. 6, pp. 112–116, 2013.
[34] A. K. I. Hassan and A. Abraham, "Modeling consumer loan default prediction using ensemble neural networks," in *2013 International Conference on Computing, Electrical and Electronic Engineering (ICCEEE)*, Aug. 2013, pp. 719–724, doi: 10.1109/ICCEEE.2013.6634029.
[35] M. Saar-Tsechansky and F. Provost, "Handling missing values when applying classification model," *Journal of Machine Learning Research*, vol. 8, pp. 1625–1657, 2007.
[36] S. Singhal and M. Jena, "A study on WEKA tool for data preprocessing, classification and clustering," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 2, no. 6, pp. 250–253, 2013.
[37] G. B. Demisse, T. Tadesse, and Y. Bayissa, "Data mining attribute selection approach for drought modelling: A case study for greater horn of Africa," *International Journal of Data Mining & Knowledge Management Process*, vol. 7, no. 4, pp. 1–16, Jul. 2017, doi: 10.5121/ijdkp.2017.7401.
[38] S. Gnanambal, M. Thangaraj, V. T. Meenatchi, and V. Gayathri, "Classification algorithms with attribute selection: an evaluation study using WEKA," *International Journal of Advanced Networking and Applications (IJANA)*, vol. 9, no. 6, pp. 3640–3644, 2018.
[39] M. Anis and M. Ali, "Investigating the performance of smote for class imbalanced learning: a case study of credit scoring datasets," *European Scientific Journal, ESJ*, vol. 13, no. 33, pp. 340–353, Nov. 2017, doi: 10.19044/esj.2017.v13n33p340.

[40] R. Pears, J. Finlay, and A. M. Connor, "Synthetic minority over-sampling technique (SMOTE) for predicting software build outcomes," *arXiv:*1407.2330.
[41] P. Domingos, "A few useful things to know about machine learning," *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, Oct. 2012, doi: 10.1145/2347736.2347755.
[42] S. A. Dudani, "The distance-weighted k-nearest-neighbor rule," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 6, no. 4, pp. 325–327, Apr. 1976, doi: 10.1109/TSMC.1976.5408784.
[43] Rajesh P and K. M, "A comparative study of data mining algorithms for decision tree approaches using WEKA tool," *Advances in Natural and Applied Sciences*, vol. 11, no. 9, pp. 230–241, 2014.
[44] G. G. Enas and S. C. Choi, "Choice of the smoothing parameter and efficiency of k-nearest neighbor classification," in *Statistical Methods of Discrimination and Classification*, Elsevier, 1986, pp. 235–244.

BIOGRAPHIES OF AUTHORS

Jovanne C. Alejandrino is currently studying Master of Information Systems in the University of Mindanao Professional Schools. He completed his bachelor's degree in Information Technology and earned professional education units at the University of Southeastern Philippines, Tagum City. He is a licensed professional teacher and currently connected in Davao del Norte State College as an IT instructor under the Institute of Computing. His research interest is in data mining and information systems. He can be contacted via jovanne.alejandrino@dnsc.edu.ph.

Jovito Jr. P. Bolacoy is currently pursuing his Master of Information Systems in the University of Mindanao Professional Schools. He works as a Lecturer under the Institute of Computing of Davao del Norte State College, Panabo City, Philippines, where he handles computer programming, and application development and emerging technologies courses. He obtained his bachelor's degree in Information Technology from the University of Southeastern Philippines in Tagum City, Philippines, *cumlaude*. His line of expertise focuses on web and desktop-based application development, information systems analysis, and design. He is now exploring the application of data mining techniques to business organizations for his future research endeavors. He can be contacted at j.bolacoy.483347@umindanao.edu.ph.

John Vianne B. Murcia is currently the university statistician and concurrent director of the Institute of Economy and Enterprise Studies of the University of Mindanao. He also chairs the Business Analytics Program of its College of Business Administration. He received his bachelor's degree in Marketing from University of Mindanao (Digos), obtained two master's degrees – MBA from University of Southeastern Philippines and Master of Predictive Analytics from Curtin University in Western Australia and finished his Ph.D. specializing in development research and administration from University of Southeastern Philippines. His expertise focuses on the applications of advanced, multivariate statistics in marketing and business decisions as well as in organizational analytics. He can be contacted at jv_murcia@umindanao.edu.ph.