Emerging Information Science and Technology
Vol. 3, No. 1, (2022), pp. 13-18

# Prediction of Student Decisions in Choosing the Type of Bank Using Support Vector Machine (SVM)

Muhammad Habil*, Slamet Riyadi, Asroni

*Universitas Muhammadiyah Yogyakarta, Jln.Brawijaya, Tamantirto, Kasihan, Bantul, Yogyakarta 55183, Indonesia*
**Corresponding author: muhammad.habil.2016@ft.umy.ac.id*

***Abstract***

*A bank is an intermediate financial institution authorized to take deposits, lend money, and issue promissory notes or banknotes. In the present day, every adult must have at least one bank account. Additionally, bank services range from regular and hajj savings to large-scale loans. Students, one of the bank's customers, usually utilize services confined to savings to preserve pocket money received from their parents and ordinary transactions like transfers and payments. Several factors, including the atmosphere, administrative fees, and the accessibility of ATMs and bank branch offices, impact students' decisions about where to save money. It prevents the bank from predicting which services must be enhanced to encourage customers, particularly students, to select the bank. Therefore, prediction is required to ascertain the students' choice of bank. This study employed data mining and the Support Vector Machine (LibSVM) algorithm. The quantity of data impacted the outcomes of the SVM classification. In addition, kernel types, k-fold values, and sampling techniques also influenced classification accuracy. LibSVM with a kernel type of RBF, a k-fold of 8, and shuffled sampling classified 200 data with an accuracy of 68.40%.*

***Keywords:*** *Decisions in Choosing a Bank, Classification, Multiclass Label, Support Vector Machine (LibSVM)*

## 1. Introduction

Every adult today must have at least one bank account. According to [1]. Banks are one of the veins of a country's economy. Besides that, banks too is a trusted institution to help smooth the payment system, and Equally important are the institutions that are the means of implementation government policy, namely monetary policy.People often utilize two sorts of banks: conventional and Islamic. Traditional banks run business conventionally, whereas Islamic banks conduct business following Sharia or Islamic law principles. The bank offers various services, including ordinary savings, Hajj savings, and large-scale loans.

Students are one of the bank's customers. Typically, the services utilized by students are confined to savings to conserve pocket money and common transactions such as transfers and payments. Nevertheless, some students utilize loan services and deposits, and others employ savings services. Students select a bank for various reasons, including the accessibility of an ATM and branch office, mutual family or friends being the bank's customers, and administrative fees. It hampers the bank from predicting what services must be improved to attract customers, particularly students, to its institution.

Therefore, predictions are necessary to determine student decisions to become bank customers. These prediction results will enable banks (conventional and Sharia) to enhance their services and offerings.

The author will describe some of the research that carried out with the concept of data mining and related to banking. In research conducted regarding prediction

reduction in the number of commercial bank customers, the external situation has changed in operations, competition among commercial bank customers is increasingly fierce [2]. Analysis Customer attrition has become a significant research topic for commercial banks. Other research conducted regarding predictions successful marketing of banking service products that contain, Level Predictions Bank Telemarketing Success using Data Mining Classification Algorithm can be applied to prospective customer data contacted to predict marketing success at the bank [3]. The next research was conducted to handle class imbalance in customer segmentation in banking. The validation process uses tenfold-cross validation, while testing model using confusion matrix and ROC curve [4]. The research results show SVM accuracy increased from 74.6% to 76.50% when combined with PSO [5].

This study measures the performance and compare the results of measuring the accuracy of the algorithm. The four algorithms produce different model accuracies for the same dataset. algorithm that Different can provide different accuracy. The higher the accuracy value produced, the more accurate the algorithm is used for prediction next customer financing. Using cross validation, NBC Algorithm gives an average level of accuracy of 66.95%, SVM of 63.71%, C 4.5, of 66.74% and Nearest Neighbor of 63.03%. Meanwhile with bootstrap validation, the NBC algorithm gives an average accuracy rate of 64.79%, SVM of 61.25%, C 4.5, of 65.91% and Nearest Neighbor of 62.26% [6]. In order to increase the role of intermediary institutions, banks compete for withdraw funds from the community and channel it to the community through improving services to customers, both individual and institutional customers. Bank vying to increase the types of products/services provided (such as various types of credit cards and debit cards), the use of various facilities with the latest technology modern services (such as online services), increasing the ability of personnel in providing services, providing convenience in applying for credit, granting rates Attractive deposit interest rates and relatively low loan interest rates low [7].Based on test results which has been carried out using Weka tools whose performance is evaluated using a confusion matrix. Data Mining is a new technology that is very useful to help companies find critical information from available data warehouses. In this case The company in question is a bank [8].

## Theoretical basis

### Bank

Bank is a financial intermediary or so-called financial intermediaries. That is, a bank institution is an institution whose activities are related with money problems. Therefore, bank business will always be associated with money problems which is the main means of facilitating the occurrence of trade [9].

### Datamining

Data mining is the process of getting useful information from a database repository big data. Data mining can also be interpreted as extracting new information taken from big chunks of data which helps in decision making. data terms. Mining is sometimes also called knowledge discovery [10].

### Support Vector Machine (SVM)

The definition of Support Vector Machine (SVM) is a learning system that uses hypothetical space in the form of linear functions in a high-dimensional feature and trained using learning algorithms based on optimization theory. SVM was first introduced in 1992 by Vapnik as a series of several superior concepts in the field of pattern recognition [11].

**LIBSVM**

Using LIBSVM usually involves two steps: first, training the data set to get a model, use the model to predict the information from the set test data. For SVC and SVR, LIBSVM can also generate probability estimates. many LIBSVM extensions are available in libsvmtools.

## 2. Method

### 2.1 Literature Review

A literature review was performed on research using data mining to discover evidence to back up the theories applied in this study. Exploring relevant literature helped this study locate useful resources and methods.

### 2.2 Data Collection

To collect the primary data on what influences students' bank selection, a Google Form questionnaire was administered to a sample of currently enrolled students at Universitas Muhammadiyah Yogyakarta (UMY). Approximately two months were spent disseminating the questionnaire.

### 2.3 Data Preprocessing

After completing data collection, a database for test materials was created. However, not all data in the database were feasible or useable, requiring data preprocessing. The initial step was data selection, focusing on selecting data suitable for test materials.

The subsequent step was data cleaning, concentrating on existing attributes. This method involved removing redundant attributes, such as those with just one class, invalid data, noise, and those that did not significantly impact the classification.

The cleaned data subsequently underwent data transformation under the requirements of the study. Not only was the modified data an attribute, but it was also the class or content of the existing data. Data transformation aimed to reduce the number of classes within each attribute to ease data classification.

### 2.4 Algorithm Implementation

Employing the stated algorithm to classify data is known as algorithm implementation. This phase is critical since it is where data mining is applied. The training data collected from data processing were classified using the Support Vector Machine (SVM) method and RapidMiner software. This study utilized the attribute of a bank of choice as a label, with three classes: conventional banks, Islamic banks, and choosing both. Age, gender, reason, number of relatives using the bank, pocket money, employment status, and parents' salary were the factors in choosing a bank.

At this point, RapidMiner's cross-validation was applied to facilitate the collection of training and testing data sets. This approach was generally employed to produce model predictions and measure the accuracy of the predictive model. In a prediction problem, a model is often provided with a known dataset to use for training (training data), as well as an unknown dataset (or data first observed) to test against the model under test (testing data). The data set consisted of 200 preprocessed data.

RapidMiner contained many SVM algorithms. However, because this study applied a multiclass label scenario, LibSVM was selected. Sampling techniques were also employed, encompassing linear, shuffled, and stratified. The selection of sampling techniques was followed by the determination of k-fold values: 3, 6, and 8. The cross-validation operator specified the sampling techniques and the k-fold values. In addition to

employing several sampling techniques and k-fold values, the kernel types were modified by the LibSVM operator. This study utilized Poly, RBF, and Sigmoid kernels.

Moreover, the gamma was adjusted to 0.8 in the kernel configuration settings. If gamma is set to its default value (gamma = 0.0), class precision and class recall will be 0.00% for one or two classes in the label. The gamma setting is required to bring the effect of 1 training data closer to influencing the classification. The gamma value of 0.8 was determined after trying all the available gamma values and discovering that it had the highest accuracy for all kernel types and could provide class precision and class recall values for all labeled classes.

## 3. Analysis Results

### 3.1 Testing Results

Compared to the Poly and Sigmoid types, the RBF kernel obtained the highest accuracy for each k-fold value, as displayed in Table 3.1. According to prior testing, a k-fold value of 8 and a kernel type of RBF yielded a maximum accuracy of 68.40%. With a k-fold of 3 and a kernel type of RBF, the second-highest accuracy was 68.33%, followed by an accuracy of 67.84% acquired by a k-fold of 6 and a kernel type of RBF.

Table 1. Comparison of Accuracy with Kernel Types and K-Fold Values

| k-fold | Accuracy | | |
|---|---|---|---|
| | RBF | Poly | Sigmoid |
| k = 3 | 68.33% | 63.81% | 60.76% |
| k = 6 | 67.84% | 61.79% | 64.26% |
| k = 8 | **68.40%** | 63.79% | 62.33% |

As illustrated in Figure 1, the run button on the toolbar was pressed when all operators had been appropriately linked to their respective ports.

Figure 1 Run Button in RapidMiner

In addition, the same test was conducted with LibSVM, a kernel type of RBF, and a k-fold value of 8 but with shuffled, stratified, and linear sampling techniques. Table 2 exhibits the results.

Table 2 Comparison of Accuracy with Sampling Techniques

| Sampling Technique | Accuracy | | |
|---|---|---|---|
| | RBF | Poly | Sigmoid |
| Shuffled | **68.40%** | 63.79% | 62.33% |
| Stratified | 67.83% | 65.83% | 58.79% |
| Linear | 65.31% | 64.27% | 63.75% |

### 3.2 Analysis and Test Results

Training and testing data were shared and automatically decided throughout data mining modeling. Cross-validation divided data into k subgroups of equal-sized data sets. K-fold cross-validation was employed to reduce classification bias. There were k training and examination sessions. Suppose the k-fold is 8; the total amount of data to be classified is divided into eight folds. In the first data mining, the testing data were the first fold data. For the second process, the fold second data were utilized; and so on, until the process was repeated eight times.

The study results implied that the selection of kernel types and k-fold values influenced the degree of accuracy. Each kernel type and k-fold value combination had different accuracy results. In addition to the kernel type and k-fold value selection, the quantity of data also affected the accuracy. The greater the quantity of data utilized in the classification, the more precise the outcomes. Inversely, the smaller the data utilized, the less accurate the findings.

In other words, RBF, a k-fold of 8, and shuffled aside were the kernel type settings with the highest accuracy compared to those with other kernel types, k-fold values, and sampling techniques.

## 4. Conclusion

The following conclusions were derived from examining the data mining model for testing and classification analysis predicting student decisions in selecting the type of bank. The LibSVM classification algorithm could predict student bank choice depending on age, gender, reason, number of relatives using Islamic banks, pocket money, employment status, and parents' salary. The amount of data significantly impacted the accuracy level. The accuracy attained for each classification using different kernel types and k-fold values was relatively low, ranging from 60.76% to 68.40% due to the small sample size of 200. The kernel type of RBF, a k-fold of 8, and shuffled sampling yielded the highest accuracy of 68.40%.

### References

[1] Umardani, D., & Muchlish, A. (2017). "Comparative Analysis of the Financial Performance of Islamic Banks and Conventional Banks in Indonesia" (In Indonesian). Jurnal Manajemen Dan Pemasaran Jasa, 9(1), 129.

[2] He, B., Shi, Y., Wan, Q., & Zhao, X. (2014). "Prediction of customer attrition of commercial banks based on SVM model". Procedia Computer Science, 31, 423–430.

[3] Dewi, S. (2016). "Comparison of 5 Data Mining Classification Algorithm Methods on Predicting the Success of Banking Service Product Marketing" (In Indonesian). None, 13(1), 60–66.

[4] Hairani, Setiawan, N. A., & Adji, T. B. (2016). "Data Mining Classification Method and SMOTE Sampling Technique Handle Class Imbalance for Customer Segmentation in the Banking Industry". (In Indonesian). Seminar Nasional Sains Dan Teknologi, 168–172.

[5] Muhamad, H., Prasojo, C. A., Sugianto, N. A., Surtiningsih, L., & Cholissodin, I. (2017). "Optimization of Naïve Bayes Classifier Using Particle Swarm Optimization on Iris Data" (In Indonesian). Jurnal Teknologi Informasi Dan Ilmu Komputer. https://doi.org/10.25126/jtiik.201743251

[6] abidarin rosidi,heri sismoro,emha taufiq luthfi,hanif al fatta,hartatik,hastari utama. (2017). "Data Management and Information Technology". (In Indonesian). Jurnal Ilmiah Dasi, 1411–3201.

[7] Kuswanto, A. (2009). "The Effect of Service Quality on Customer Satisfaction Levels". (In Indonesian). Jurnal Ilmiah Ekonomi Bisnis, 14(2), 5873.

[8] Chang, C. C., & Lin, C. J. (2011). LIBSVM: A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology, 2(3), 1–39.

[9] Wilarjo, S. B. (2014). "Definition, Role, and Development of Islamic Banks in Indonesia". (In Indonesian). Igarss 2014, 2(1), 1–5.

[10] Haryati, S., Sudarsono, A., & Suryana, E. (2015). "Implementation of Data Mining to Predict Student Study Period Using the C4.5 Algorithm (Case Study: Dehasen University Bengkulu)". (In Indonesian). Jurnal Media Infotama, 11(2), 130–138.

[11] Puspitasari, A. M., Ratnawati, D. E., & Widodo, A. W. (2018). "Classification of Dental and Oral Diseases Using the Support Vector Machine Method". (In Indonesian). Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer, 2(2), 802–810.