Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2026. Vol. 5. No. 2; e-ISSN: 2808-4519

Loan Default Risk Prediction System in Online Loan Services using Machine Learning and Streamline

Yosi[1], Septian Jose[2*], Celvin Andrean[3], Sarmila[4], Weiskhy Steven Dharmawan[5]

[1,2,3,4]*Program Studi Informatika, Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika*
[5]*Program Studi Sistem Informasi Akuntansi, Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika*
cresstayosi@gmail.com [1], septianjose5@gmail.com[2*], andreancelvin405@gmail.com[3], mila29918@gmail.com[4], weiskhy.wvn@bsi.ac.id[5]

Abstract

The rapid development of information technology has driven innovation in the financial sector, particularly in the field of credit lending services. However, the increasing number of credit lending services also leads to a higher risk of default, which can lead to financial losses for lenders. This study aims to develop a loan default risk prediction system using machine learning algorithms, namely Naïve Bayes and K-Nearest Neighbor (KNN), implemented through the Streamlit framework. This study applies a quantitative method with a data mining approach based on the CRISP-DM framework, utilizing the German Credit dataset consisting of variables such as age, occupation, housing, savings account, loan amount, and purpose. The models were evaluated using a confusion matrix to measure accuracy. The results show that the Naïve Bayes algorithm achieved the highest accuracy (86.4%) in predicting loan decisions, followed by KNN (60%). The developed Streamlit-based application provides interactive visualizations, training models, and prediction features, enabling users to assess credit risk efficiently. This system is expected to help financial institutions identify potential defaulters more accurately and improve the overall performance of credit lending services.

Keywords: *Loan default prediction, Machine Learning, Naïve Bayes, K-Nearest Neighbor, Streamlit*

1. Introduction

With the advancement of information technology, innovations in the financial sector have emerged, such as financial technology services, which facilitate access to online transactions (Rahmahafida, 2020). In the Financial Services Sector Consumer Protection regulations stipulated in POJK Number 1/POJK.07/2013, it has not been able to reach the peer-to-peer lending market because there are no regulations stating that peer-to-peer lending is included in the consumer protection regulations for the financial services sector [2]. The target of this online money lending service facility is quite broad, Both high-skilled and low-skilled individuals. Online money lending services have grown significantly in recent years due to the widespread use of e-commerce, which is inextricably linked to the internet. The ease and speed of access to online loan applications has attracted the younger generation to apply for loans. Despite challenges related to financial risk analysis and a limited understanding of basic accounting, the ease of the risk assessment process within online loan applications can expedite loan application results for customers. In addition, the loan platform size and term offered in the application are automatically presented based on the calculation results of the data inputted by the customer during initial registration.

By connecting with each other in real time, the business world process in the field of lending and borrowing money is carried out by creditors as lenders to debtors as loan recipients. Technology in adults is now increasingly providing a very significant impact, especially on human life, which we can all know in information technology which is considered to have very fast and even rapid development, a number of advantages related to the development of information technology can be promoted and we can see that there is convenience in carrying out a number of social lives [3]. To address the increasing risk of default in online lending services, a system capable of accurately and efficiently predicting potential default is needed. The approach used is the Naive Bayes algorithm, K-Nearest Neighbor (KNN). The Naive Bayes method is one method that can be used in decision making to obtain better results in a classification problem [4] . This algorithm uses Bayes' Theorem, the main assumption, the features are independent of each other, although not completely fulfilled in real data Naïve Bayes still gives quite good results in many cases[5]. The Naïve Bayes algorithm is a simple but effective method for predicting a possibility from historical data, using Bayes' theorem and the assumption of independence between predictor variables, which is able to produce fast and accurate decisions. So it is expected to help financial institutions in identifying potential borrowers who have the potential to experience more precise defaults, reducing the level of credit congestion, and increasing security and trust in the digital financial ecosystem. K-Nearest Neighbor (KNN) is a simple non-parametric method used for classification and regression based on proximity to training samples [6]. The KNN method algorithm is very simple, working based on the shortest distance from the test sample to the training sample to determine the KNN.

2. Research Methodology

2.1 Types and Approaches to Research

Table 1: Data description

Column	Description
Unnamed	Index number data
Age	Age/request
Sex	Gender
Job	Occupation type (0–3)
Housing	Residence status
Saving accounts	Savings amount
Checking account	Checking account balance
Credit amount	Amount of loan offered
Duration	Loan term (months)

This study uses a quantitative method with a data mining approach to build a prediction model for the risk of default on online loan services. Data mining is the process of discovering hidden patterns and knowledge from data sets [7]. The type of research used is computational experimental research. Computational experiments are experiments that use computer simulations or algorithms to analyze models and compare the results with empirical data [8]. The focus of the research is to produce a classification model that is able to predict whether a lender has the potential to be current or default based on historical patterns in the dataset.

2.2 Research Data

The data used in this study is secondary data, in the form of customer loan histories obtained from a historical financial technology dataset containing attributes such as Age, Gender, Occupation, Housing, Savings Account, Checking Account, Loan Amount, Duration, Purpose, and Status. The data was then processed into two classes: default (bad) and non-default (good). The data collection process was carried out using a documentation study method, namely by compiling the secondary dataset German Credit Data.

2.3 Research Stage

The research stage follows the CRISP-DM (Cross Industry Standard Process for Data Mining) framework, an international standard framework used to systematically execute data mining projects from business understanding to implementation models [9]. This consists of several main stages.

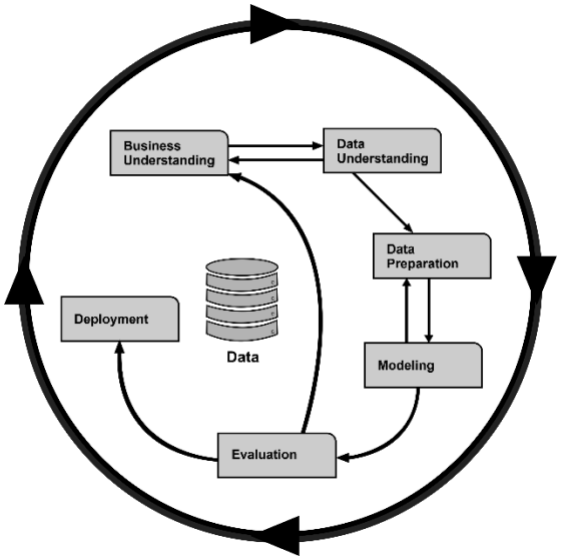

Fig 1 : CRISP-DM Image

The business understanding stage identifies the objective, which is to minimize the risk of payment failure through machine learning-based predictions. The data understanding stage analyzes patterns, class distributions, and variable characteristics. The data preparation stage includes data cleaning, attribute reduction, normalization, and dividing the dataset into training and test data. The modeling stage is carried out by applying two algorithms:
Naïve Bayes and K-Nearest Neighbor (KNN). Next, in the evaluation stage, the model is tested using a confusion matrix. A confusion matrix is a table that describes the performance of a classification model on a set of test data whose true values are finally known [10]. In the deployment stage, the model is ready to be implemented as a component of a credit scoring system. Naive Bayes uses Bayes' theorem to calculate the probability of each class and classify the data into the class with the highest probability.
The basic formula used is:

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (1)$$

Where P(H|X) is the probability of hypothesis H after seeing data X, P(X|H) is the probability of data X appearing if hypothesis H is true, P(H) is the initial probability of the hypothesis, and P(X) is the probability of data appearing. For numerical data, the Gaussian distribution is used, while for categorical data, the frequency of attribute appearance is used.
The K-Nearest Neighbor (KNN) algorithm works by classifying test data based on its closest distance to a set of training data. The distance between two data points is calculated using the Euclidean Distance formula as follows:

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

Model results were evaluated using a confusion matrix, assessing True Positive, True Negative, False Positive, and False Negative values to measure classification performance. From this matrix, an accuracy metric was calculated using the following formula:

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

Accuracy is a measure that shows the proportion of correct classification results against all tested data [11]. Accuracy is used to show the percentage of success of the model in predicting loan status. A model is said to be good if it produces high accuracy and has low classification errors.

3. Result and Discussion

This system is implemented using the Python programming language with the Streamlit framework as the user interface. Streamlit is an open-source Python framework used to quickly and easily build interactive web applications, especially for machine learning and data science purposes. The system is designed with three main pages: View Data, Train Model, and Prediction Page.

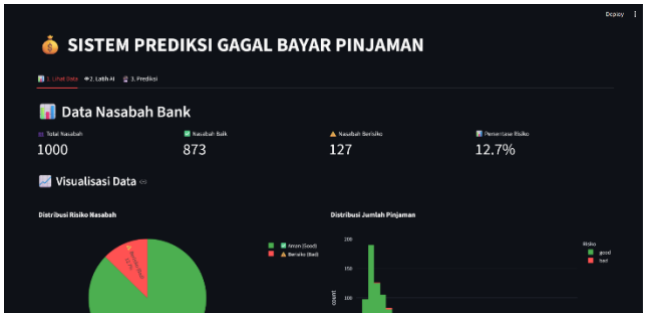

Fig 3 : View Data Menu Display (continued)

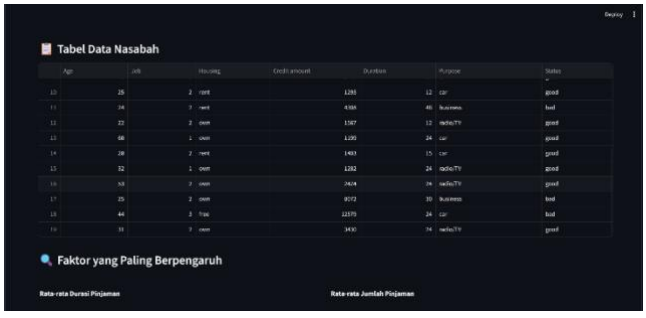

Fig 2 : View Data Menu Display

In the View Data menu, users can view the entire customer dataset used in the training model. User data is displayed in tabular form and complemented by interactive graphs and histograms that show general characteristics of the customer data.

On the "Train Model" page, users can select one or more algorithms from three options: Naive Bayes and K-Nearest Neighbor and specify

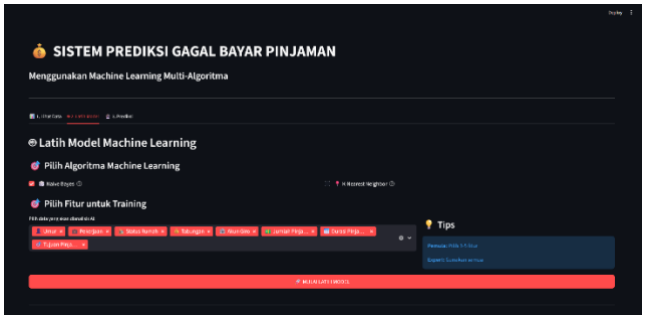

Fig 4 : Train Model Menu View

the features to be used for training with an intuitive tag system after which the system will start the training process with the "START TRAINING MODEL" button. After the training process is complete, the system displays a comprehensive compilation of the three algorithms in the form of tables and graphical visualizations.

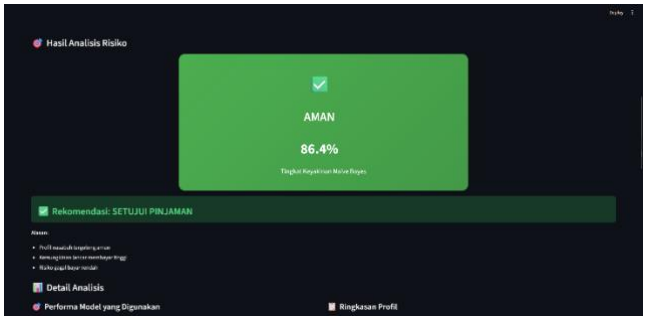

Fig 6 : Naive Bayes Algorithm Prediction Results

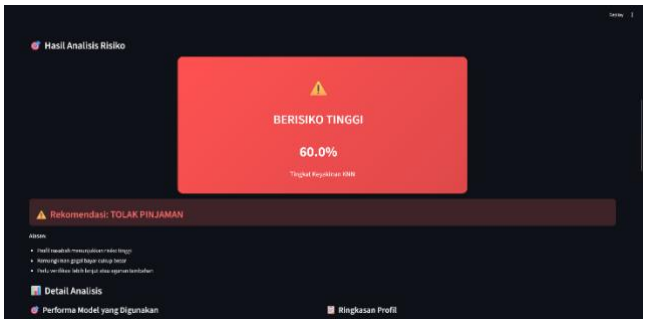

Fig 5 : K-Nearest Neighbors Prediction Results

Naive Bayes recommends a loan agreement with a high confidence level of 86.4% because the customer profile is considered safe and the risk of default is low, while KNN (K-Nearest Neighbors) recommends rejecting the loan on the grounds of high risk with a prediction result of 60.0%.

4. Conslusion

This research successfully developed a machine learning-based online loan default risk prediction system implemented using Streamlit. The two algorithms used, Naïve Bayes and K-Nearest Neighbor, showed varying results in terms of accuracy and confidence. Test results showed that the Naïve Bayes algorithm provided the best results with the highest confidence level. The developed application is capable of displaying customer data, training models, and making predictions in an interactive and user-friendly manner. This system can help financial institutions analyze potential customer defaults more quickly, efficiently, and accurately, thereby minimizing the risk of credit default and increasing security and trust in online lending services.

5. References

[1] N. I. Rahmahafida, "Perlindungan Hukum Pihak Pemberi Pinjaman pada Layanan Pinjaman Pendidikan Berbasis Teknologi Informasi terhadap Risiko Gagal Bayar," *Jurist-Diction*, vol. 3, no. 2, 2020, doi: 10.20473/jd.v3i2.18203.

[2] D. Wati and T. Syahfitri, "DAMPAK PINJAMAN ONLINE BAGI MASYARAKAT," *Community Development Journal : Jurnal Pengabdian Masyarakat*, vol. 2, no. 3, 2022, doi: 10.31004/cdj.v2i3.2950.

[3] A. M. A. Mentari, "Analisis Faktor-Faktor Keputusan Pemberian Kredit Pinjaman Online (Studi Kasus PT. Cicil Solusi Mitra Teknologi)," *Jurnal Ilmiah Mahasiswa FEB*, vol. 9, no. 2, 2021.

[4] D. Alita, I. Sari, A. R. Isnain, and S. Styawati, "PENERAPAN NAÏVE BAYES CLASSIFIER UNTUK PENDUKUNG KEPUTUSAN PENERIMA BEASISWA," *Jurnal Data Mining dan Sistem Informasi*, vol. 2, no. 1, 2021, doi: 10.33365/jdmsi.v2i1.1028.

[5] N. Umar and M. A. Nur, "Application of Naïve Bayes Algorithm Variations On Indonesian General Analysis Dataset for Sentiment Analysis," *Jurnal RESTI*, vol. 6, no. 4, 2022, doi: 10.29207/resti.v6i4.4179.
[6] P. Rani, "A Review of various KNN Techniques," *Int J Res Appl Sci Eng Technol*, vol. V, no. VIII, 2017, doi: 10.22214/ijraset.2017.8166.
[7] J. Ha, M. Kambe, and J. Pe, *Data Mining: Concepts and Techniques*. 2011. doi: 10.1016/C2009-0-61819-5.
[8] R. Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*. 2014. doi: 10.4135/9781473909472.
[9] R. Wirth, "CRISP-DM : Towards a Standard Process Model for Data Mining," *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, no. 24959, 2000.
[10] D. M. W. POWERS, "Estimation of high affinity estradiol binding sites in human breast cancerEVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION," *Journal of Machine Learning Technologies*, vol. 2, no. 1, 2011.
[11] S. R. Durugkar, R. Raja, K. K. Nagwanshi, and S. Kumar, "Introduction to data mining," 2022. doi: 10.1002/9781119792529.ch1.