Advances in Artificial Intelligence Research (AAIR)
ISSN 2757-7422
www.dergipark.com/aair/

Vol. 4 (No. 2), pp. 87-95, 2024
doi: 10.54569/aair.1585994
Published online: Dec 30, 2024

Prediction of Home Loan Approval with Machine Learning

Gamze Güder [1, *], Utku Köse [1]

[1] Suleyman Demirel University, Faculty of Engineering, Department of Computer Engineering, Isparta, Turkey;

Abstract

With the introduction of computers into our lives, the size and complexity of data have increased. The growing amount of data made manual processing more difficult, and machine learning methods were adopted to minimize human errors. In the banking sector, the increasing volume of data necessitated the use of machine learning techniques. Numerous studies have been conducted in the literature on the banking sector. In this study, machine learning methods, including k-nearest neighbors, random forest algorithm, support vector machines, and logistic regression, were used to predict whether a bank would approve a housing loan or not. Two different datasets were used for the analysis. The results were compared and presented using performance metrics. This study aims to minimize human errors, make the credit approval processes in banks safer, and provide faster results for loan applications.

Keywords: *Knn algorithm; Random Forest algorithm; Support vector machines; Logistic regression.*

1. Introduction

The concept of housing has always maintained its importance throughout history. In ancient times, humans sought shelter primarily for protection from wild animals. However, with the development of technology and the changing needs of society, the concept of the right to housing emerged [1]. As the importance of the right to housing increased, efforts were made to secure it through laws, declarations, and other means. In Turkey, the right to housing has been attempted to be secured through Article 57 of the Constitution [2].

Economic fluctuations and rising inflation have reduced people's purchasing power. Just as in the rest of the world, the decline in purchasing power in Turkey has made access to housing, one of the most basic human rights, more difficult. The increase in rental prices in Turkey has made taking out loans to buy a house more attractive. Loans, one of the main products of banks, are funded by savers who deposit their savings into the bank, and thus, banks aim to achieve the highest profit from loans [3]. Just as important as giving loans, the repayment of those loans is also crucial for banks. For this reason, it has become important for banks to use machine learning techniques to minimize human error and make predictions by utilizing past data during the loan approval process. By using machine learning algorithms, banks are able to conduct faster and more secure processes. In the growing banking sector, where technology plays an increasingly important role, machine learning techniques are frequently employed. The literature contains many studies on loan approval processes, using various machine learning techniques and performance metrics.

In this study, two different datasets were used, which distinguishes it from other studies. The goal was to measure the compatibility and consistency of the results of the algorithms using these two datasets. Additionally, the study observed how the models were affected by different types of data. The k-nearest neighbors algorithm, random forest algorithm, support vector machines, and logistic regression algorithms were used in housing loan approval processes, and the results were presented in a comparative manner using performance metrics. For each algorithm, 90% of the two datasets were used for training data, while 10% was used for testing data.

2. Related Works

Arun et al. [4]: The study focused on predicting the most suitable customer for credit by evaluating customer applications. Decision trees, RF, linear models, neural networks, and AdaBoost algorithms were used in the study. However, the article did not mention the performance evaluation criteria or which algorithm produced the best results.

Gautam et al. [5]: This study focuses on predicting whether a loan will be approved by financial institutions by analyzing the history and reliability of customers applying for credit using machine learning methods. The dataset used is a collection from the banking sector. The results of the study show that the majority of the standard needs of bank employees are met. Additionally, the study emphasizes that the system was trained and tested with current data. The potential for these data to lose their relevance over time was taken into account. Therefore, the importance of integrating artificial intelligence systems with automation systems to incorporate new data into the system was highlighted.

Aphale et al. [6]: This study used a dataset from a real cooperative bank. Two-thirds of the dataset was used

*Corresponding author
E-mail address: gamzeyildirimmm@hotmail.com

Received: 15/Nov/2024; Accepted: 28/Dec/2024.

for training, and one-third for testing. Various algorithms were employed in this study, and the models were compared using performance evaluation criteria. The results showed that the performance evaluation metrics for the Nearest Centroid and Gaussian Naive Bayes algorithms performed reliably well. The algorithms that showed good performance each achieved values ranging from 76% to 80%. Additionally, a predictive model was formulated using linear regression to forecast customers' creditworthiness.

Gupta et al. [7]: This study attempts to predict whether a customer is reliable for a loan by analyzing their previous credit records using machine learning techniques. The dataset used in this study was obtained from Kaggle's open access section. A heatmap was used to show the relationships between the parameters. Supervised learning approaches, specifically Logistic Regression and Random Forest algorithms, were used in the study. Customer information (such as gender, number of dependents, marital status, whether they run their own business, income, etc.) is entered through a user interface and the resulting credit approval status is displayed on the screen.

Ndayisenga et al. [8]: The study aimed to predict whether the repayment of loans in the Rwandan banking sector would occur by assessing an individual's past data, essentially estimating the credit risk and determining whether loan applicants would be approved. The dataset from Kigali Bank, operating in Rwanda, was used. When various models were compared based on performance evaluation metrics, the best results were observed with the Gradient Boosting model. The study concluded that customers with a credit score of B had a low probability of defaulting on their loans.

Fati et al. [9]: In the study, credit status prediction was performed using machine learning algorithms. A dataset from Kaggle's open access, containing 615 rows of training data and 368 rows of test data across 13 columns, was used. The models were compared using performance evaluation criteria, and it was concluded that logistic regression performed better than the others.

Kadam et al. [10]: The study used machine learning algorithms to predict which customers' loans would be approved or rejected by financial institutions. The dataset from Dream Housing Finance Company was used. Support Vector Machines and Naïve Bayes algorithms were applied in this study. The results showed that the Naïve Bayes algorithm had the best performance.

Khan et al. [11]: The study examined and compared different models for predicting loan approval, aiming to identify the model that produced the best results with the least margin of error in determining loan approval. Data mining, statistics, and probability were used to develop the prediction model. Data from various sources were gathered to create a statistical model. As the amount of data increased, the model became more precise, reducing the error rate, thereby lowering the credit risk for banks and saving time for both customers and financial institutions. Accuracy rates were observed as 80.945% for Logistic Regression, 93.648% for Decision Tree, and 83.388% for the Random Forest (RF) algorithm. Following cross-validation, accuracy rates were 80.945% for Logistic Regression, 72.213% for Decision Tree, and 80.130% for the Random Forest algorithm. Based on these results, it was concluded that the Random Forest algorithm provided the best outcome with the least error.

Udhbav et al. [12]: In this study, a predictive modeling system was developed to help mortgage finance companies assess consumers' eligibility for home loans by evaluating eligibility criteria. Two classification models, Logistic Regression and Gradient Boosting, were used. To obtain better accuracy and performance, the Gradient Boosting model was applied to eliminate the errors that occurred in the results of Logistic Regression. In the results section of the study, when comparing Gradient Boosting and Logistic Regression, it was concluded that Gradient Boosting provided better accuracy and precision. As a result of this study, the credit approval process became faster for both consumers and mortgage finance companies.

Tütüncü et al. [13]: This study aims to identify the algorithm that best predicts whether a loan will be repaid by calculating credit risks. In this context, various algorithms were used to build models, which were then compared using different performance evaluation criteria. The dataset was obtained from Home Credit, available in Kaggle's open access. The dataset was divided into 60% for training, 20% for validation, and 20% for testing. In this study, while the KNN algorithm produced less successful results for customers who defaulted and those who did not, it was concluded that the Gradient Boosting algorithm showed the best classification performance. Additionally, when examining the ROC curve results, it was observed that the KNN algorithm had a lower success rate compared to the other algorithms.

Oral et al. [14]: In this study, the Madrid Real Estate Market dataset, available in Kaggle's open access, was used as the dataset. 25% of the data was used for testing, and 75% for training. Machine learning algorithms included in the MATLAB program were applied to the dataset. The top 5 models with the best R^2 values were included in the study. The results showed that the methods with the best performance, in order, were Bagged Trees Ensemble, Fine Tree, Exponential GPR, Wide Neural Network, and Quadratic SVM.

Anand et al. [15]: This study aims to determine the algorithm that best predicts whether a loan will be repaid by calculating the criteria banks use to decide on granting loans, in other words, the credit behaviors of banks. The datasets are composed of consumer credit application requests collected from various websites, including Kaggle's open access. A total of 15 classification algorithms were used. The results of the top five algorithms

that performed the best are provided. The study found that Extra Trees Classifier and Random Forest models produced the most accurate results, and through the comparisons, it was determined that the most effective criteria were income, work experience, and debt status. The significant impact of the credit score on loan approval decisions by banks was emphasized. Taking these factors into account, risky and problematic customers were quickly identified.

Tumuluru et al. [16]: In the study, the dataset was obtained from Kaggle's open-access resources, with 70% allocated to training and 30% to testing. A comparison of algorithm accuracy showed that Random Forest performed best with an accuracy of 81%.

Viswanatha et al. [17]: This study focuses on predicting whether individual loan applications will be approved by financial institutions using machine learning methods. The dataset used is a credit dataset available in the open access section of Kaggle. The libraries used for data processing and model building are scikit-learn, pandas, and numpy. It was observed that the best accuracy result, 83.73%, was achieved by the Naive Bayes algorithm.

Uddin et al. [18]: This study focuses on developing a community-based machine learning credit prediction system to identify reliable customers who are likely to repay loans issued by financial institutions. The dataset used in this study was obtained from Kaggle's open access section. SMOTE (Synthetic Minority Over-sampling Technique) was applied for data preprocessing and balancing the dataset. Among the nine models used, the top three machine learning models were applied for the task. To compare the machine learning models with deep learning models, a Dense Neural Network, Long Short-Term Memory (LSTM), and Recurrent Neural Network (RNN) were employed. Performance evaluation criteria were used to assess the models. Before applying community machine learning, the best-performing models were observed to be Extra Trees, Random Forest, and KNN, in that order. The accuracy of Extra Trees was 86.64%. After applying a voting ensemble to the top three models, the accuracy increased to 87.26%. Additionally, a desktop application was developed for financial institutions, enabling them to check the credit eligibility of their customers. This interface facilitates faster credit approval processes and improves operational efficiency.

Çelik et al. [19]: This study focuses on predicting bank loans using different algorithms. The aim of this study was to make accurate predictions regarding credit risk and credit assessments. Various algorithms were compared based on different dataset sizes and features in terms of classification performance. The dataset used in this study was obtained from Kaggle's open access section. 80% of the dataset was used for training, and 20% was used for testing. According to performance evaluation metrics and the ROC curve, the best-performing algorithm was the CatBoost algorithm. Additionally, it was observed that ensemble learning algorithms performed better in predicting bank loans.

Prasad et al. [20]: The dataset used consists of historical data from the credit market. The libraries used for data processing and model building are scikit-learn, matplotlib, pandas, NumPy, and PyTorch. Various machine learning methods were used and compared using performance evaluation metrics. The best results were observed with the RF and KNN algorithms.

3. Methodology

3.1. K-Nearest Neighbor Algorithm (KNN)

The k-nearest neighbor algorithm (KNN) is a supervised learning algorithm. It is a simple algorithm that performs classification by using similarities within the dataset. KNN classifiers, while classifying, use a predetermined number, k, to check the data points that are closest to the new data based on a distance function. The most commonly used distance functions are Euclidean, Manhattan, and Minkowski. While classifying, the algorithm checks the distance data to determine which group has the most similar data, and the new data is assigned to that group. For most datasets, the optimal value for k is chosen to be between 3 and 10. In some studies, the value of k is selected as the square root of the number of rows in the dataset [8][21].

3.2. Random Forest Algorithm

RF algorithm is an ensemble learning method. It is also used for classification and regression operations [22].

Working Principle of the Random Forest Algorithm:

The data subsets used in the Random Forest algorithm are created by randomly selecting from the dataset. These subsets have the same number of rows as the original dataset and randomly selected columns [23]. Bootstrapping is the process of creating a new dataset using the original dataset. In the Random Forest algorithm, during the bootstrapping process, the same data is used for each decision tree [24]. Decision trees are then trained on these new datasets [25]. A new data row is tested on the trained decision trees, and the resulting outcomes are recorded. The aggregation of these results refers to the process where the most frequently occurring result is selected [22].

3.3. Logistic Regression

Logistic regression, one of the classification algorithms, is used to estimate the probability of a dependent variable taking two values such as 1 or 0, true or false. [26]. Logistic regression is expressed by the formula $f(z) = 1/ 1+e^{-z}$. If the value of z is $-\infty$ the function's value is 0, and if z is $+\infty$ the function's value is 1. Regardless of the value of z, the result of logistic regression is always between 0 and 1 [27].

3.4. Support Vector Machines (SVM)

In Support Vector Machines, which is a supervised learning algorithm, data is considered as points with coordinates in n-dimensional space. SVM performs classification by drawing a hyperplane. SVM algorithms set the distance between the points in different categories and closest to each other to be maximum and this maximum distance is called margin. The points that we call support vectors are the points that touch the margin [8].

4. Results and Discussion

In this study, the development environment used was www.kaggle.com. For each algorithm, 90% of the dataset was allocated for training and 10% for testing.

The value of k can be determined by taking the square root of the number of rows in the dataset or by selecting k between 3 and 10 [21]. In this study, knn algorithm was used and the value of k was chosen over a wide range to train the model. To observe the results over a wider range, 29 different values for the k parameter (values between 5 and 35 were tested one by one) were applied, and the accuracy rates for both datasets are shown in Figure 1 and Figure 2. As shown in the graph for the first dataset in Figure 1, the highest accuracy rate of 87.096 was achieved when k=11; for the second dataset, as shown in the graph in Figure 2, the highest accuracy rate of 92.974 was obtained when k=16.

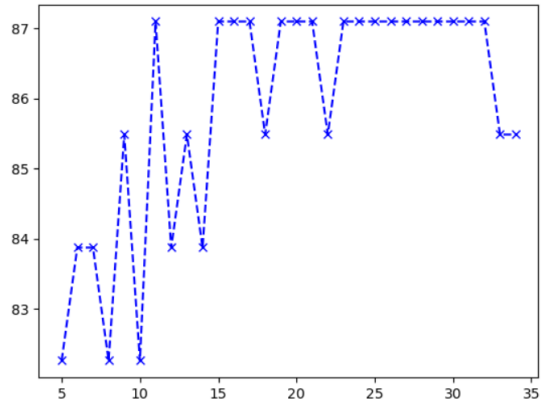

Figure 1. *Accuracy rate chart for the first dataset according to the value of k*

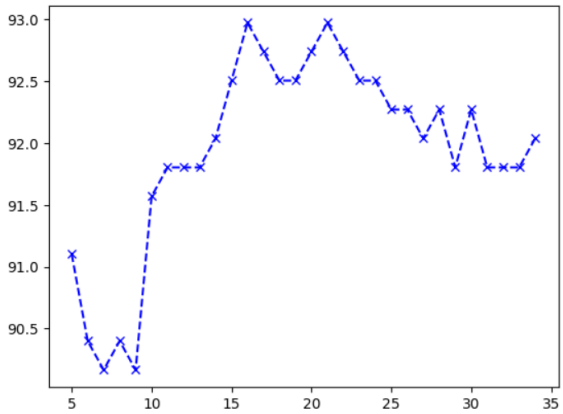

Figure 2. *Accuracy rate chart for the second dataset according to the value of k*

This study used the RF algorithm, where the number of trees was selected between 10 and 101, and the accuracy rates were calculated based on this range. As shown in Figures 3 and 4, the accuracy rates for both datasets are displayed according to the number of trees selected within the given range. For the Dataset_1, the highest accuracy rate was 87.09% when the number of trees was 10; for the Dataset_2, the highest accuracy

rate was 98.82% when the number of trees was 28. The fact that the Dataset_1 reaches the highest accuracy with 10 trees indicates that the Dataset_1 is less complex compared to the Dataset_2. On the other hand, the higher accuracy of the Dataset_2 compared to the first suggests that the model has better learned the complexity of the data and made more accurate decisions with more trees.

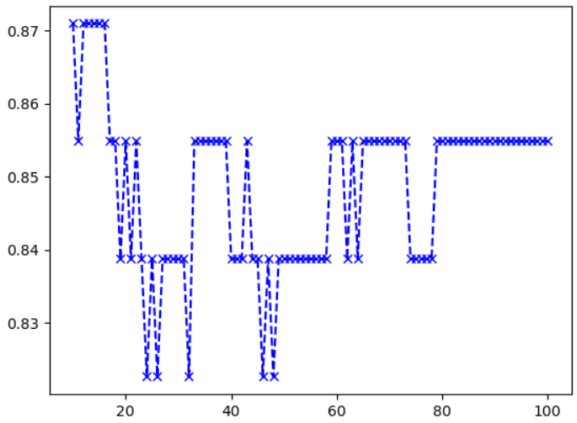

Figure 3. *Accuracy rates for the* Dataset_1 *according to the number of trees*

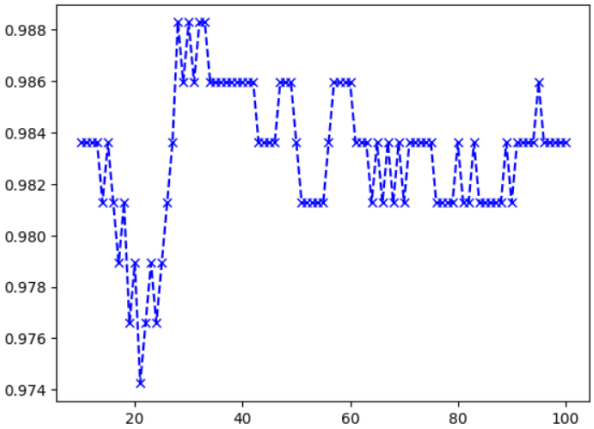

Figure 4. *Accuracy rates for the* Dataset_2 *according to the number of trees*

For both datasets, the columns are visualized in Figures 1.5 and 1.6 in order of importance. When examining the Random Forest Algorithm graphs, it is observed that the factor most influencing the credit approval process for the first dataset is credit history, while for the second dataset, it is credit score. The factor least affecting the credit approval process in both datasets is education status. The feature with the lowest importance, which ranks last, was removed for both datasets. The model was then retrained using only the features with higher importance. For the first dataset, the accuracy rate was 0.870, but after retraining the model, the accuracy rate decreased to 0.8226. For the second dataset, the accuracy rate was 0.9882, and after retraining the model, the accuracy rate increased to 0.9906.

The decrease in accuracy for the first dataset indicates that the education and gender columns play a role in the model's decision-making process. It also suggests that these columns may be interrelated. For example, higher education levels generally increase the likelihood of working in higher-paying jobs and finding employment. Having a higher income is also a factor that increases the likelihood of repaying a loan. The reason gender affects the accuracy rate may be because men typically have longer periods of uninterrupted work in the labor force compared to women, who may have breaks in their careers due to cultural or biological reasons, such as childbirth and childcare. This can represent a higher risk for the lending institution, which is an undesirable situation for them.

In the second dataset, the columns for self-employment and education level are of less importance compared to other columns. Lending institutions tend to prefer customers with stable incomes. A self-employed individual may not always provide reliable information about income stability, as self-employed individuals may have either high or low incomes, and their income level can fluctuate throughout the year. Credit score, however, holds more weight in the second dataset compared to other columns. A good education level does not necessarily indicate a high income. For someone with a good education, factors such as not being able to find employment due to national conditions could also affect their credit score. Assets owned by the customer,

the loan term and amount, and the credit score have a greater impact on the loan approval process, whereas education level and self-employment status, due to their lower importance, have a lesser effect on the credit decision process.

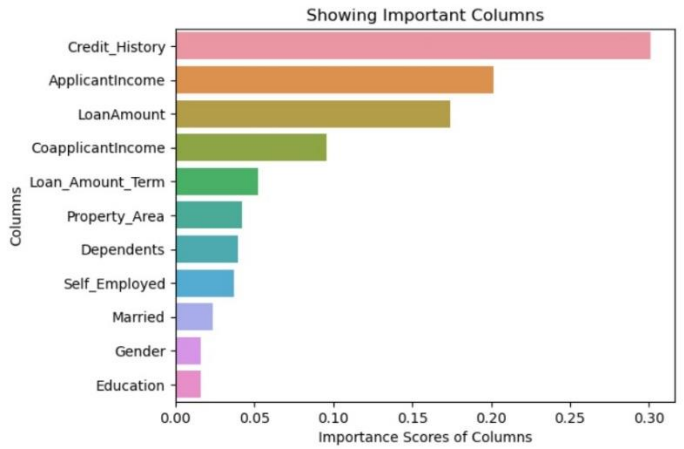

Figure 5. *Column importance and percentage in the RF algorithm for the first dataset*

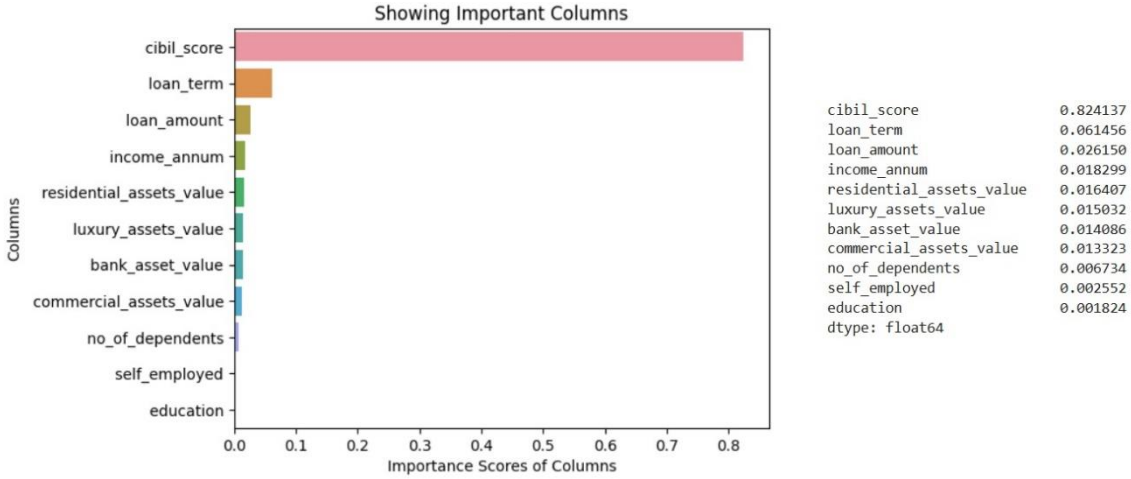

Figure 6*. Column importance and percentage in the RF algorithm for the second dataset*

In this study, using the Logistic Regression algorithm, the accuracy rate was found to be 87.09% for the Dataset_1 and 90.86% for the Dataset_2. Achieving an accuracy rate of 87.09% on the first dataset indicates that the model performs adequately on a small and sparse dataset. In the second dataset, the accuracy rate of 90.86% can be explained by the data being more homogeneous and within a narrower range. The second dataset, with a smaller standard deviation, allowed the model to generalize better. The organization of the data in this way helped logistic regression produce more stable and consistent results, which in turn led to an increase in accuracy.

These results reveal that one of the key factors affecting the performance of the logistic regression model is the structure of the dataset. More homogeneous and lower-variance datasets lead to higher accuracy and better results. This highlights the importance of considering the statistical characteristics of the dataset when analyzing model performance.

In this study, using the Support Vector Machine algorithm, the accuracy rate was found to be 88.7% for the Dataset_1 and 93.9% for the Dataset_2. The difference in accuracy rates is related to the fact that the two datasets have different characteristics. The size and distribution of the dataset are important factors that affect the model's learning ability and accuracy.

4.1. Performance Evaluation Criteria

Accuracy is a classification metric, but it is not sufficient on its own for evaluating classification results [28]. Therefore, in this study, in addition to the accuracy rate, other classification metrics such as precision, recall, F1-score, specificity, negative predictive value, and false discovery rate have been used.

A confusion matrix is the tabulation of actual and predicted values. Performance evaluations are made using the data in the matrix [29]. To determine the performance evaluation metrics of the classification algorithms, the confusion matrices for the algorithms on both datasets are shown in Table 1. Upon examining the table, it is observed that in Dataset_1, the highest number of TN (credit rejection) was correctly predicted by the KNN algorithm, while the highest number of TP was correctly predicted by the support vector machines. In Dataset_2, both the highest number of TN and TP were correctly predicted by the random forest algorithm.

Table 1. *Confusion matrices obtained from classification results*

Confusion matrices			Dataset_1		Dataset_2	
			Prediction			
			N	P	N	P
KNN	Actual values	N	9	6	149	10
		P	2	45	20	248
LR		N	8	7	140	19
		P	1	46	20	248
RF		N	9	6	156	3
		P	2	45	2	266
SVM		N	8	7	149	10
		P	0	47	16	252

Performance evaluations results of the classification algorithm for both datasets are shown in Table 2.

Table 2. *Performance evaluations of the classification algorithm*

Metric	Dataset_1				Dataset_2			
	KNN	RF	LR	SVM	KNN	RF	LR	SVM
Accuracy	0.870	0.870	0.870	0.887	0.929	0.988	0.908	0.939
Precision	0.882	0.882	0.868	0.870	0.961	0.989	0.929	0.962
Recall	0.957	0.957	0.979	1.000	0.925	0.993	0.925	0.940
F1 Score	0.918	0.918	0.920	0.931	0.943	0.991	0.927	0.951
Specificity	0.600	0.600	0.533	0.533	0.937	0.981	0.881	0.937
Negative Predictive Value (NPV)	0.818	0.818	0.889	1.000	0.882	0.987	0.875	0.903
False Discovery Rate (FDR)	0.118	0.118	0.132	0.130	0.039	0.011	0.071	0.038

The accuracy rate is highest in Dataset_1 with 88.70% using the SVM algorithm, and in Dataset_2 with 98.8% using the Random Forest (RF) algorithm. The accuracy rate in dataset_2 is higher compared to dataset_1. This indicates that the dataset_2 model generalizes better.

In Dataset_1, the highest precision metric is 0.882 for the Random Forest algorithm, while the highest recall metric is 1.000 for the SVM algorithm. In Dataset_2, the highest precision is 0.989 and the highest recall is 0.993 for the RF algorithm. The high precision metric for both datasets means that most of the positive predictions are correct. The precision metric reached its highest values for both datasets using the Random Forest algorithm. A high precision metric reduces credit risk. The high recall metric for both datasets means that most of the actual positive cases were correctly predicted. A high recall helps financial institutions accurately identify customers who are likely to repay their loans, thereby contributing to an increase in profitability.

The F1 score in Dataset_1 is highest at 0.931 for the SVM algorithm, and in Dataset_2, it is highest at 0.991 for the RF algorithm. A high F1 score indicates that the test performs well overall, meaning it has both a high number of true positive results and low numbers of FP and FN.

In Dataset_1, the highest specificity metric is 0.600 for the Random Forest algorithm, and in Dataset_2, it is 0.981 for the RF algorithm. The low specificity in the first dataset indicates that there is a high likelihood of

customers who were not granted credit being incorrectly predicted as positive (approved for credit). The high specificity in the second dataset, on the other hand, suggests that customers with high risk can be rejected, helping to prevent financial losses. It also contributes to a more accurate assessment of customers during the credit approval process and increases trust in decision-making by financial institutions.

In Dataset_1, the highest Negative Predictive Value is 1.000 for the SVM algorithm, and in Dataset_2, it is 0.987 for the RF algorithm. The high negative prediction values in both datasets help financial institutions avoid economic losses by accurately identifying customers who should not be granted credit.

In Dataset_1, the highest False Discovery Rate is 0.132 for the Logistic regression algorithm, and in Dataset_2, it is 0.071 for the Logistic Regression (LR) algorithm. Although the rates are low in both datasets, the lower false discovery rate in the second dataset indicates that the model produces fewer false positives and provides more reliable results.

5. Conclusion

In this study, various machine learning methods were employed to predict whether a bank would approve a home loan. Two different datasets were used in the study. The reason for using two datasets was to compare the compatibility and results of the algorithms across multiple datasets and to evaluate the consistency of the results. The two datasets were not considered as conflicting or competing elements in this study, but rather as control groups to validate the work. Additionally, the goal was to observe how the varying sizes of the two datasets would affect the results. In the classification phase of the study, KNN, RF, SVM, and Logistic Regression algorithms were compared. In the study conducted on the two separate datasets, the highest accuracy rates for Dataset_1, excluding precision and specificity, were obtained with the SVM algorithm (Accuracy: 88.7% (SVM), Precision: 88.2% (RF), Recall: 100% (SVM), F1: 93.1% (SVM), Specificity: 60% (RF), NPV: 100% (SVM), FDR: 13.2% (LR)). In Dataset_2, the highest accuracy rates across all performance metrics were observed for the Random Forest and Logistic Regression algorithms (Accuracy: 98.8% (RF), Precision: 98.9% (RF), Recall: 99.3% (RF), F1: 99.1% (RF), Specificity: 98.1% (RF), NPV: 98.7% (RF), FDR: 0.071% (LR)). The results obtained from dataset_1 show that the small and sparse nature of the dataset limits the model's performance. Although SVM demonstrated high success, particularly in metrics like sensitivity and F1 score, it fell behind the RF algorithm in terms of precision and specificity. The low performance in specificity, especially, indicates that the model struggled to correctly predict the negative class. In dataset_2, however, the RF algorithm performed overall better, thanks to the larger amount of data and narrower distribution. The grouping of data points in a closer range and more homogeneous manner in this dataset improved the model's accuracy. The high precision, recall, and F1 score of the RF algorithm indicate that the model accurately classifies both classes, while the specificity and NPV rates further reinforce its overall success. These results demonstrate that larger and more homogeneous datasets increase the model's generalization capacity and allow for more accurate predictions.

The results obtained help financial institutions predict customer behavior and make quicker decisions about whether to approve a loan or not. Additionally, financial institutions can perform credit risk assessments by considering both the advantages and disadvantages of the algorithms used. Furthermore, with these results, financial institutions can increase the potential pool of customers eligible for credit.

Acknowledgement

This article is derived from the Master's thesis titled “Prediction of Home Loan Approval with Machine Learning”.

References

[1] Kösedağ, E. (2023). Türkiye’de Konut Hakkı ve Bu Hakkın Kullanılmasında Ortaya Çıkan Sorunlara Yönelik Değerlendirme. Kent Akademisi, 16(1), 595-611. https://doi.org/10.35674/kent.1220084

[2] Mevzuat Bilgi Sistemi, 1982. Türkiye Cumhuriyeti Anayasası Erişim Tarihi: 08.01.2024. https://www.mevzuat.gov.tr/mevzuat?MevzuatNo=2709&MevzuatTur=1&MevzuatTertip=5

[3] Koyuncu, C., & Berrin, S. (2011). Takipteki Kredilerin Özel Sektöre Verilen Krediler Ve Yatırımlar Üzerindeki Etkisi. Dumlupınar Üniversitesi Sosyal Bilimler Dergisi, (31).

[4] Arun, K., Ishan, G., & Sanmeet, K. (2016). Loan approval prediction based on machine learning approach. IOSR J. Comput. Eng, 18(3), 18-21.

[5] Gautam, K., Singh, A. P., Tyagi, K., & Kumar, M. S. (2020). Loan Prediction using Decision Tree and Random Forest. International Research Journal of Engineering and Technology (IRJET), 7(08), 853-856.

[6] Aphale, A. S., & Shinde, S. R. (2020). Predict loan approval in banking system machine learning approach for cooperative banks loan approval. International Journal of Engineering Trends and Applications (IJETA), 9(8).

[7] Gupta, A., Pant, V., Kumar, S., & Bansal, P. K. (2020, December). Bank loan prediction system using machine learning. In 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART) (pp. 423-426). IEEE.

[8] Ndayisenga, T. (2021). Bank loan approval prediction using machine learning techniques (Doctoral dissertation).
[9] Fati, S. M. (2021). Machine learning-based prediction model for loan status approval. Journal of Hunan University Natural Sciences, 48(10).
[10] Kadam, A. S., Nikam, S. R., Aher, A. A., Shelke, G. V., & Chandgude, A. S. (2021).Prediction for loan approval using machine learning algorithm. International Research Journal of Engineering and Technology (IRJET), 8(04).
[11] Khan, A., Bhadola, E., Kumar, A., & Singh, N. (2021). Loan approval prediction model a comparative analysis. Advances and Applications in Mathematical Sciences, 20(3), 427-435.
[12] Udhbav, M., Kumar, R., Kumar, N., Kumar, R., Vijarania, D., & Gupta, S. (2022). Prediction of Home Loan Status Eligibility using Machine Learning. Swati, Prediction of Home Loan Status Eligibility using Machine Learning (May 27, 2022).
[13] Tütüncü, T. E. (2022). Makine öğrenmesi algoritmaları ile kredi temerrüt riskini tahmin etme (Master's thesis, Bursa Uludağ Üniversitesi).
[14] Oral, M., Okatan, E., & Kırbaş, İ. (2021). Makine öğrenme yöntemleri kullanarak konut fiyat tahmini üzerine bir çalışma: Madrid örneği. Uluslararası Genç Araştırmacılar Öğrenci Kongresi, Burdur, Turkey.
[15] Anand, M., Velu, A., & Whig, P. (2022). Prediction of loan behaviour with machine learning models for secure banking. Journal of Computer Science and Engineering (JCSE), 3(1), 1-13.
[16] Tumuluru, P., Burra, L. R., Loukya, M., Bhavana, S., CSaiBaba, H. M. H., & Sunanda, N. (2022, February). Comparative Analysis of Customer Loan Approval Prediction using Machine Learning Algorithms. In 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS) (pp. 349-353). IEEE.
[17] Viswanatha, V., Ramachandra, A. C., Vishwas, K. N., & Adithya, G. (2023). Prediction of Loan Approval in Banks Using Machine Learning Approach. International Journal of Engineering and Management Research, 13(4), 7-19.
[18] Uddin, N., Ahamed, M. K. U., Uddin, M. A., Islam, M. M., Talukder, M. A., & Aryal, S. (2023). An ensemble machine learning based bank loan approval predictions system with a smart application. International Journal of Cognitive Computing in Engineering, 4, 327-339.
[19] Çelik, E., & Gür, Ö. Banka kredisi tahmini için makine öğrenmesi algoritmalarının performans analizi: Topluluk öğrenmesi algoritmalarının üstünlüğü. Artıbilim: Adana Alparslan Türkeş Bilim ve Teknoloji Üniversitesi Fen Bilimleri Dergisi, 7(1), 1-20.
[20] Prasad, P V V S V, Nageswara Rao, P.V (2024). Loan Approval Prediction System Using Machina Learning. International Journal of Innovative Science and Research Technology, 9(4), 278-281
[21] Sendel, E. (2023). Skin lesion classification with machine learning, Makine öğrenmesi ile cilt lezyonu sınıflandırması.
[22] Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.
[23] Peker, Özkaraca, O., & Kesimal, B. (2017). Enerji tasarruflu bina tasarımı için isıtma ve soğutma yüklerini regresyon tabanlı makine öğrenmesi algoritmaları ilemodelleme. Bilişim Teknolojileri Dergisi, 10(4), 443-449.
[24] Sarica, A., Cerasa, A., & Quattrone, A. (2017). Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: a systematic review. Frontiers in aging neuroscience, 9, 329.
[25] Ekelik, H., & ALTAŞ, D. (2019). Dijital Reklam Verilerinden Yararlanarak Potansiyel Konut Alıcılarının Rastgele Orman Yöntemiyle Sınıflandırılması. Journal Of Research İn Economics, 3(1), 28-45.
[26] Kazan, S., & Karakoca, H. (2019). Makine öğrenmesi ile ürün kategorisi sınıflandırma. Sakarya University Journal of Computer and Information Sciences, 2(1), 18-27.
[27] Kleinbaum, D. G., Klein, M. (2010). Logistic regression: a self-learning text (Third Edition). New York: springer.
[28] Zheng, A. (2015). Evaluating Machine Learning Models, Farnham, U.K.:O'Reilly Media, Inc.
[29] Santra, A. K., & Christy, C. J. (2012). Genetic algorithm and confusion matrix for document clustering. International Journal of Computer Science Issues (IJCSI), 9(1), 322.