Gazi Üniversitesi
Fen Bilimleri Dergisi
PART C: TASARIM VE TEKNOLOJİ

Gazi University
Journal of Science
PART C: DESIGN AND TECHNOLOGY

GU J Sci, Part C, 12(2): 644-663 (2024)

A Comparative Study of Loan Approval Prediction Using Machine Learning Methods

Vahid SİNAP[1*]

[1]*Ufuk University, Faculty of Economics and Administrative Sciences, Department of Management Information Systems, Ankara, Türkiye*

Article Info

Research article
Received: 20/03/2024
Revision: 29/04/2024
Accepted: 30/04/2024

Keywords

Data Mining
Loan Prediction
Machine Learning
Banking Sector
Feature Selection
Cross-validation

Makale Bilgisi

Araştırma makalesi
Başvuru: 20/03/2024
Düzeltme: 29/04/2024
Kabul: 30/04/2024

Anahtar Kelimeler

Veri Madenciliği
Kredi Tahmini
Makine Öğrenmesi
Bankacılık Sektörü
Özellik Seçimi
Çapraz Doğrulama

Graphical/Tabular Abstract (Grafik Özet)

In this study, models were developed using machine learning algorithms for loan approval prediction and the effects of various feature selection methods on the models were investigated. / Bu araştırmada kredi onayı tahmini için makine öğrenmesi algoritmaları kullanılarak modeller geliştirilmiş ve çeşitli özellik seçim yöntemlerinin modeller üzerindeki etkileri incelenmiştir.

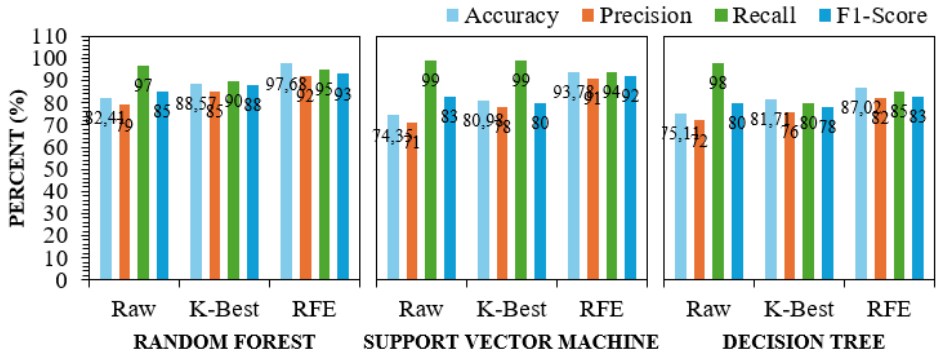

Figure A: *Impact of feature selection methods on the performance of models /* ***Şekil A:*** *Özellik seçim yöntemlerinin modellerin performansına etkisi*

Highlights (Önemli noktalar)

- *The Random Forest (RF) algorithm showed the highest performance with an accuracy rate of 97.71% for loan approval predictions. / RF algoritması, kredi onayı tahminlerinde doğruluk oranı %97.71 ile en yüksek performansı göstermiştir.*
- *The RFE method significantly improved model performance. Models built with RFE with selected features achieved higher accuracy, recall, precision, and F1-Score values than models built with K-Best method or with all features. / RFE yöntemi, model performansını önemli ölçüde artırmıştır. RFE ile seçilen özelliklerle oluşturulan modeller, K-Best yöntemiyle veya tüm özelliklerle oluşturulan modellere göre daha yüksek doğruluk, duyarlılık, kesinlik ve F1-Skor değerleri elde etmiştir.*

Aim (Amaç): *The aim of this study is to evaluate the effects of feature selection methods, K-Best and RFE, on model performance in loan approval prediction. / Bu çalışmanın amacı, kredi onayı tahminlemede özellik seçimi yöntemleri olan K-Best ve RFE yöntemlerinin model performansları üzerindeki etkisini değerlendirmektir.*

Originality (Özgünlük): *With the models developed in the study, significantly higher accuracy rates were obtained in loan approval prediction than similar studies in the literature. / Araştırmada oluşturulan modeller ile kredi onay tahminlemede literatürdeki benzer çalışmalardan önemli ölçüde yüksek doğruluk oranları elde edilmiştir.*

Results (Bulgular): *It was found that model performance was significantly improved using the RFE method, the RF algorithm achieved the highest accuracy rate, and the cross-validation method provided more consistent results in measuring model performance compared to the Training, Testing and Validation technique. / Çalışmada, RFE yöntemi kullanılarak model performansının belirgin şekilde iyileştiği, RF algoritmasının en yüksek doğruluk oranına ulaştığı, çapraz doğrulama yöntemi, model performansını ölçmede Eğitim, Test ve Doğrulama tekniğine kıyasla daha tutarlı sonuçlar sağladığı tespit edilmiştir.*

Conclusion (Sonuç): *Feature selection methods can improve model performance and redundant features can negatively affect model performance. / Özellik seçimi yöntemleri model performansını iyileştirebilmekte ve gereksiz özellikler model performansını olumsuz etkileyebilmektedir.*

*Corresponding author, e-mail: vahidsinap@gmail.com
DOI: 10.29109/gujsc.1455978

GU J Sci, Part C, 12(2): 644-663 (2024)

Gazi Üniversitesi
Fen Bilimleri Dergisi
PART C: TASARIM VE TEKNOLOJİ

Gazi University
Journal of Science
PART C: DESIGN AND TECHNOLOGY

http://dergipark.gov.tr/gujsc

A Comparative Study of Loan Approval Prediction Using Machine Learning Methods

Vahid SİNAP[1*]

[1]*Ufuk University, Faculty of Economics and Administrative Sciences, Department of Management Information Systems, Ankara, Türkiye*

Article Info

Research article
Received: 20/03/2024
Revision: 29/04/2024
Accepted: 30/04/2024

Keywords

Data Mining
Loan Prediction
Machine Learning
Banking Sector
Feature Selection
Cross-validation

Abstract

Loan prediction plays an important role in the process of evaluating loan applications by financial institutions. Machine learning models can automate this process and make the lending process faster and more efficient. In this context, the main objective of this research is to develop models for loan approval prediction using machine learning algorithms such as Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Decision Tree, and Random Forest and to compare their performances. In addition, determining the effect of K-Best and Recursive Feature Elimination feature selection methods on model performances is another important objective of the research. Furthermore, the evaluation of the effectiveness of techniques such as cross-validation (K-Fold) and Train, Test and Validation in measuring the performance of models is also among the objectives of the research. The findings revealed that married individuals are more likely to be approved for loans than single individuals, high income individuals more likely than low-income individuals, males more likely than females, and university graduates more likely than non-university graduates. According to the performance measures, Random Forest was the most successful algorithm with an accuracy rate of 97.71% in loan approval prediction. To achieve this accuracy rate, feature selection was performed with the Recursive Feature Elimination method and the measurement was made with the cross-validation method. It was found that the feature selection methods have a significant impact on the model performances and the Recursive Feature Elimination method was the most successful method. Moreover, the highest accuracy rate achieved by the Random Forest algorithm, which showed the highest performance in all cases, was measured by cross-validation.

Makine Öğrenmesi Yöntemleri Kullanılarak Kredi Onay Tahmini Üzerine Karşılaştırmalı Bir Çalışma

Makale Bilgisi

Araştırma makalesi
Başvuru: 20/03/2024
Düzeltme: 29/04/2024
Kabul: 30/04/2024

Anahtar Kelimeler

Veri Madenciliği
Kredi Tahmini
Makine Öğrenmesi
Bankacılık Sektörü
Özellik Seçimi
Çapraz Doğrulama

Öz

Kredi tahmini, finans kuruluşlarının kredi başvurularını değerlendirme sürecinde önemli bir rol oynamaktadır. Makine öğrenmesi modelleri bu süreci otomatikleştirebilmekte ve kredi onay sürecini daha hızlı ve verimli hale getirebilmektedir. Bu bağlamda, bu araştırmanın temel amacı Lojistik Regresyon, K-En Yakın Komşu, Destek Vektör Makinesi, Karar Ağacı ve Rastgele Orman gibi makine öğrenmesi algoritmalarını kullanarak kredi onay tahmini için modeller geliştirmek ve performanslarını karşılaştırmaktır. Ayrıca, K-Best ve Yinelemeli Özellik Eleme (Recursive Feature Elimination) özellik seçim yöntemlerinin model performansları üzerindeki etkisinin belirlenmesi de araştırmanın bir diğer önemli amacıdır. Buna ek olarak, çapraz doğrulama (K-Fold) ve Eğit, Test Et ve Doğrula gibi tekniklerin modellerin performansını ölçmedeki etkinliğinin değerlendirilmesi de araştırmanın amaçları arasındadır. Bulgular, evli bireylerin bekar bireylere, yüksek gelirli bireylerin düşük gelirli bireylere, erkeklerin kadınlara ve üniversite mezunlarının üniversite mezunu olmayanlara kıyasla kredilerinin onaylanma olasılığının daha yüksek olduğunu ortaya koymuştur. Performans ölçütlerine göre, Rastgele Orman kredi onayı tahmininde %97,71 doğruluk oranıyla en başarılı algoritma olmuştur. Bu doğruluk oranına ulaşmak için özellik seçimi Yinelemeli Özellik Eleme yöntemi ile gerçekleştirilmiş ve ölçüm çapraz doğrulama yöntemi ile yapılmıştır. Özellik seçim yöntemlerinin model performansları üzerinde önemli bir etkiye sahip olduğu ve Özyinelemeli Özellik Eleme yönteminin en başarılı yöntem olduğu görülmüştür. Ayrıca, tüm durumlarda en yüksek performansı gösteren Rastgele Orman algoritmasının elde ettiği en yüksek doğruluk oranı çapraz doğrulama ile ölçülmüştür.

*Corresponding author, e-mail: vahidsinap@gmail.com
DOI: 10.29109/gujsc.1455978

1. INTRODUCTION (GİRİŞ)

The loan approval process is a critical step in a financial institution's lending decisions to customers. A properly structured loan approval process improves the lender's profitability by lending to customers who are unlikely to be insolvent. This process also improves risk management and avoids potential losses by avoiding lending to high-risk customers [1]. Moreover, the right loan approval process increases customer satisfaction, strengthens customer loyalty, and expands the customer base [2]. This has the potential to provide more loans and increase the lender's revenue. In addition, the proper execution of the approval process has a direct impact on the financial well-being of applicants. A proper loan approval process determines the loan terms, considering the financial situation of the applicants. In this way, applicants can receive loans on favorable terms and build a more financially sound foundation [3]. In other words, the right loan terms can reduce the financial burden of applicants by offering payment plans and interest rates that suit their financial situation [4]. This makes the loan repayment process more manageable for applicants and helps them build a more financially sound foundation. Furthermore, providing loans with favorable loan terms can make it easier for applicants to achieve their financial goals and increase their financial well-being. Thus, identifying the right loan terms can improve the financial situation of applicants and be an important step towards a more financially secure future [5].

The traditional loan approval process used to be manually assessed on certain parameters such as the applicant's credit history, income level, employment status and similar financial metrics. Credit history shows a person's past loan and debt repayments, and payment history is an important factor indicating a borrower's eligibility for a loan [6]. The income level determines whether a person has the financial strength to afford the loan payments, while the employment status reflects the person's ability to earn a regular income [7]. In addition, financial metrics such as the debt-to-income ratio are among other important parameters considered during the loan approval process. These parameters help the lender assess the applicant's eligibility for the loan and set the loan terms. However, since this process is manual, it is time-consuming and the risk of making mistakes is high.

Recently, loan usage and applications have increased significantly in Türkiye. Especially with the various loan products and campaigns offered by banks, there has been a significant increase in loan requests. In the July-September 2023 period, a total of TRY 277 billion worth of loans were extended to approximately 6.5 million people. These figures represent a 13% increase in the number of borrowers and a 61% growth in the amount of loans disbursed compared to the same period of the previous year [8]. This increase in demand has made manual processes in the loan approval process even more challenging. The intensity of manual processes may prevent the rapid evaluation and finalization of loan applications, thus increasing the risk of errors. In addition, efficiency issues stand out among the disadvantages that manual loan approval processes face with increasing loan applications. During peak application periods, manual processes are inefficient and may hinder the rapid processing of loan applications. This leads to long periods of waiting for answers and dissatisfaction on the part of applicants [9]. Furthermore, the time-consuming nature of manual processes can incur additional costs for organizations. Factors such as hiring additional staff and repeating processes are factors that increase operating costs. On the other hand, the human factor increases the possibility of making mistakes in manual processes. Staff working under intensity and stress may enter data incorrectly or make erroneous decisions [10]. Long loan approval processes and erroneous decisions can negatively affect organizations' reputation and customer loyalty [11]. Finally, the weight of manual processes in the rapidly digitalizing financial sector can reduce competitiveness. Rival organizations with faster and more efficient processes may be preferred by customers [12]. For these reasons, the need for automated and data-driven loan approval systems is becoming more and more important. These systems evaluate loan applications faster and more efficiently, reducing the workload of banks and providing a better service to customers.

Automated and data-driven loan approval systems enable faster evaluation of loan applications, reducing the workload of banks and providing better service to customers. Applicants' financial history, income level and other important factors are analyzed using technologies such as big data analytics, artificial intelligence and machine learning, and decisions such as loan approval or rejection are made automatically [13]. This ensures that applications are finalized quickly and at the same time reduces the risk of making mistakes. In addition to reducing the workload of banks, automated loan approval systems provide a better service to customers. Rapid decision-making has the potential to increase customer satisfaction, while at the same time increasing the competitiveness of

organizations. These systems also ensure that loan approval processes are more transparent and fairer. Data-driven decisions are made based on objective criteria and the impact of human error is minimized [14]. This results in a positive outcome for both banks and customers.

Along with the advantages of automated loan approval systems, there are also some disadvantages [15]. For instance, incorrect decisions can be made if these systems are programmed incorrectly or if data are misinterpreted. In addition, the sensitivity level of these systems needs to be adjusted and continuously updated. If these systems are misused or abused by malicious people, customer privacy may be at risk and unfair practices may arise. Therefore, the security and accuracy of these systems is of paramount importance and should be continuously reviewed [16]. In addition, the fact that these systems eliminate the human factor may, in some cases, reduce the importance of human observation and assessment, leading to potential errors. For these reasons, automated loan approval systems need to be properly programmed, regularly updated, and secured to be used effectively [17].

The selection of machine learning models is an important step for making the right decisions in loan approval systems. If the right model is not selected, the performance of the system may decrease, and wrong decisions may be made. In the literature, there are various techniques used in loan application approval prediction. These techniques include machine learning algorithms and statistical methods. Some of these techniques are as follows:

- Logistic Regression (LR): Provides a simple and interpretable model. However, it may be limited in problems with complex relationships, such as loan application approval prediction. It may struggle to express complex interactions between income, credit history and other factors. Therefore, it can be combined with other methods or replaced with more flexible models to model more complex relationships.

- Decision Trees (DT): Effective for modeling complex decision structures. It can be used to explain complex decision processes such as loan application approval prediction. However, they can be prone to overfitting, meaning that they may overfit the training data and lose the ability to generalize. To avoid this, it is important to control the depth of the trees. Deeper trees are generally more prone to overfitting, while shallower trees can produce more generalizable models. Therefore, it is important to determine the optimal depth of the decision tree model [18].

- Random Forest (RF): RF is another method used for credit approval prediction. This method is an ensemble model that is built by combining multiple decision trees. RF is more resistant to overfitting because it uses a common decision algorithm that is built by combining many different trees [19]. This means that the ensemble model can often produce more generalizable results, even if each individual tree is prone to overfitting. RF generally provides high accuracy and is capable of modeling complex relationships. However, this complexity can reduce the interpretability of the model. That is, it can be difficult to understand why the model makes a particular prediction or which features are important. Therefore, when using complex models such as RF, it is important to carefully evaluate the model's performance as well as the model's results so that they can be interpreted correctly.

- K-Nearest Neighbors (KNN): This algorithm is known as a simple and interpretable classification method. To classify an instance, KNN looks at the labels of its nearest neighbors and takes a majority decision and assigns that instance to that class [20]. Therefore, it is basically simple and easy to understand. However, KNN also has some disadvantages. For large datasets, the computational cost can be high because for each prediction, it may be necessary to calculate the distance to all other instances in the dataset. It can also be sensitive to noise in the dataset, meaning that small random changes in the dataset can significantly affect the model's predictions.

- Support Vector Machine (SVM): SVM is resistant to overfitting, meaning that it does not overfit the training data and retains the ability to generalize. It can also perform well on high-dimensional datasets. However, the computational cost of SVM when working on large datasets can be high because SVM classifies each instance in the dataset by comparing it with support vectors. Therefore, it should be considered that the computational cost and time may increase when using SVM on large datasets [21]. Also, due to the complexity of SVM, it can be difficult to interpret the results of the model, especially when working on multidimensional datasets. For these reasons, SVM can often be a good option for medium-sized datasets, while for large datasets

it is a method that should be used with more caution. However, when properly applied, SVM can provide high accuracy and can be an important tool with the ability to model complex relationships.

Determining the hyperparameters of the models is an important step to avoid the disadvantages of the techniques. Incorrect specification of hyperparameters can affect the accuracy of the model and reduce its performance [22]. For instance, overfitting or underfitting levels for a model can negatively affect the model's performance. These problems can be avoided by making the right hyperparameter adjustments. Feature selection is also an important process that affects the performance of the models in predicting the loan approval procedure. Selecting the wrong or unnecessary features can reduce the performance of the model [23]. For instance, ignoring an important feature such as income level may prevent accurate loan decisions from being made. If all these factors are not handled correctly, the performance of loan approval systems can be degraded, and incorrect decisions can be made. This can result in losses for both lenders and customers. Therefore, attention should be paid to the selection of machine learning models, determination of hyperparameters and feature selection.

This study deals with predicting the loan approval status of a bank using customer data. Within the scope of the research, the models are trained with LR, KNN, SVM, DT and RF machine learning algorithms and the performance of the models is evaluated by comparing the test data with actual loan approval results. The main objective of the study is to evaluate the effect of techniques such as K-Best and Recursive Feature Elimination (RFE) used in feature selection on model performance. Another important objective of the study is to evaluate the effectiveness of cross-validation (K-Fold) and Train, Test and Validation techniques in measuring the performance of the models.

2. RELATED RESEARCH (İLGİLİ ARAŞTIRMALAR)

Predicting loan approval is a topic of great importance for financial institutions because accurate predictions help them minimize financial risks and increase profits. Research on this topic has evaluated the usability and effectiveness of various machine learning and statistical methods.

The study by Kadam et al. [24] emphasizes the importance of forecasting loan defaults in banking systems. A large portion of banks' revenue comes from loan interest and therefore, predicting loan defaults significantly affects banks' profitability. The aim of the study is to examine and compare different machine learning methods used to predict loan defaults. The study finds that the Naïve Bayes model outperforms the SVM model for predicting loan defaults.

Kadam et al. [25] aim to develop a web-based application for banks to evaluate loan applications more efficiently. The web application developed in their study provides instant loan approval predictions to users. The application uses LR to predict the probability of loan approval and calculates a credit score called CIBIL score. The developed model provides an efficient performance for accurately evaluating loan applications and calculating the credit score.

In the study conducted by Saini et al. [26], RF, KNN, SVM, and LR models were used to predict customers' loan approval outcomes, and their performances were compared. According to the results of the study, the RF algorithm was the most successful algorithm with an accuracy rate of 98.04%.

The aim of the study by Singh et al. [27] is to use machine learning models to predict whether loan applications will be approved in the banking sector and to determine the most successful algorithm. For this purpose, various classification algorithms such as LR, RF classifier, SVM classifier were used. As a result of the experimental studies, it was determined that the best performance was obtained with the RF classifier.

With the increasing demand for loans, banks are forced to lend despite their limited resources. This creates the need to reduce risks so that banks can make safer choices when lending. Diwate et al. [28] examined the use of artificial intelligence models to predict the safety of loan applications by data mining on data from banks' previous lending experiences. In this way, it is aimed to contribute to the safe lending process by saving banks' efforts and resources. SVM algorithm was used in the research and an accuracy rate of 81% was obtained.

The aim of the study by Alaradi and Hilal [29] is to develop a high-performance forecasting model for loan approval prediction using decision trees. They experimented with different tree methods starting from the most simplified and comprehensible decision tree to the most complex random forests. The results showed poor performance over

simplified decision trees because due to the highlighted correlated and complex feature space, most critical parameters affecting loan approval are not reflected, resulting in an oversimplified tree that is impractical to implement. However, in terms of performance, relevance and interpretability, the DT algorithm stood out. The accuracy on the test dataset was 97.25%. Therefore, the DT-based prediction model is proposed to facilitate the decision-making process regarding the eligibility of loan application based on the applicants' characteristics.

Kumar et al. [30] analyzed bank loan data using machine learning methods. The study aimed to identify the features that are important for accurately predicting the loan value of customers. The analysis shows that the identified important features are effective in accurately determining the loan value of customers. Naive Bayes, DT and LR algorithms were used in the study and the most successful algorithm in predicting the loan value of customers was Naive Bayes with an accuracy rate of 80%.

Uddin et al. [31] discuss the challenges faced by banks in the process of evaluating loan applications. In order to overcome these difficulties, a system that enables automatic evaluation of loan applications using machine learning methods has been developed. The study includes the use of LR, DT, RF, Extra Trees (ET), SVM, KNN, Gaussian Naive Bayes, AdaBoost, Gradient Boosting, Dense Neural Network, Long Short-Term Memory, and Recurrent Neural Network algorithms and evaluating the performance of these algorithms. The experiments show that the ensemble model performs better than the individual models. In this context, ET algorithm was the most successful algorithm with 86.64% accuracy.

Tejaswini et al. [32] mentioned in their study that forecasting loan borrowers is a difficult task for the banking sector. Loan recovery is an important parameter in a bank's financial statements. Predicting a customer's probability of loan repayment is a significant challenge. The researchers mentioned that machine learning techniques can be useful in such tasks. In their study, LR, DT, and RF machine learning algorithms were used to predict customer loan approval. The experimental results show that the accuracy of the DT machine learning algorithm is better than LR and RF.

The study by Ramachandra et al. [33] aims to deploy the model on cloud-based platforms using machine learning algorithms and concepts to identify and understand the working method of loan systems for loan prediction. The main objective of the project is to predict which of the customers will or will not pay their loans, using leading algorithms such as DT, LR and RF. The LR algorithm achieved 86% accuracy with minimal error.

While Meshref [34] notes that the Bank Marketing dataset on Kaggle is often used to predict long-term deposit subscription, he thinks that this dataset can also be used to predict whether loan applications will be approved or not. The research builds a loan approval prediction model using ensemble machine learning techniques such as Bagging and Boosting. The results showed that the AdaBoost model had an accuracy rate of 83.98%.

Gupta et al. [35] points out that with the advancement of technology, there have been many developments in the banking sector and the number of applications for loan approval increases every day. There are certain policies that banks need to consider when selecting an applicant for loan approval. Based on certain parameters, the bank needs to decide which applicant it finds most suitable for approval. It is difficult and risky to manually check each individual and then recommend them for loan approval. In their work, they developed a web-based application that utilizes LR and RF algorithms to predict creditworthy borrowers based on the borrower's past records.

Sheikh et al. [36] stated that there are many products such as loans in the banking sector and the main source of income of banks comes from these products. It is stated that by predicting loan approval results in advance, banks can reduce the Non-Performing Assets (NPA) problem. In the study conducted with the data collected from Kaggle, the LR model was used to predict loan approvals. In the study, which emphasizes the importance of attribute selection in terms of the performance of the model, an accuracy rate of 81% was obtained in determining the selection of customers eligible for loan approval.

Tumuluru et al. [37] noted that in today's increasingly competitive market, estimating the risk involved in a loan application is one of the most important challenges to the survival and profitability of banks. The study mentioned that most banks use credit scoring and risk assessment procedures to review loan applications and make loan approval decisions, yet every year many people fail to repay their loans or default on their loans. This causes financial institutions to lose significant amounts of money. In the study, Machine Learning

algorithms were used to extract patterns from a common loan dataset and predict future loan defaulters. In the analysis, customer data such as age, income, loan amount and tenure were used. RF, SVM, KNN and LR algorithms were evaluated and compared with standard metrics. Among the algorithms, RF achieved a better accuracy of 81%.

This study differs from other studies in the literature by focusing on the impact of advanced techniques such as cross-validation, feature selection, hyperparameter optimization on models while evaluating the effectiveness of machine learning methods in predicting loan approval in the banking industry. The main objective of the study is to provide a comprehensive analysis to determine the most appropriate model by comparing various machine learning algorithms. In this analysis, the effectiveness of machine learning algorithms such as LR, KNN, SVM, DT and RF will be examined, and it will be determined which algorithm predicts loan applications more accurately.

K-Best and RFE methods, which are feature selection methods, will be discussed in this study. It will be examined whether and how these methods can be used to determine which features are the most important in evaluating loan applications and which method provides the best performance. In addition, this study will evaluate K-Fold cross-validation and Train, Test, and Validation techniques. It will be investigated which method provides the best results and how these techniques can be used to accurately evaluate the model performance. As a result of these analyses, an approach that allows banks to evaluate loan applications more effectively will be proposed. This proposal will offer a new perspective to improve the effectiveness of machine learning methods in loan approval prediction.

3. MATERIAL AND METHODS (MATERYAL VE METOD)

In this section, descriptions of the machine learning algorithms used in the study, performance criteria used in comparing the algorithms, characteristics of the dataset, and data preparation process are provided. The aim is to establish the methodological and analytical foundations of the research, enhance its scientific contribution, and ensure reproducibility.

3.1. Classification Algorithms (Sınıflandırma Algoritmaları)

In this study, five supervised classification algorithms are used for loan approval prediction. These algorithms include LR, KNN, SVM, DT, and RF. The algorithms used in the study are briefly explained in subheadings.

3.1.1. Logistic regression (Lojistik regresyon)

LR is a classification algorithm often used in machine learning. This algorithm attempts to predict the probability of a dependent variable by taking a set of linear combinations of independent variables. Typically, LR is used if the dependent variable is divided into two classes (binary classification), such as predicting whether loan applications will be approved. These predictions are then classified at a threshold (usually set at 0.5) to determine the class of the dependent variable [38]. The basic formula for LR is given in Equation 1.

$$P(Y=1) = \frac{1}{1+e^{-(b_0+b_1X_1+b_2X_2+\ldots+b_kX_k)}} \quad (1)$$

When the equation is analyzed, $P(Y=1)$ signifies the likelihood of the dependent variable being 1. The letter e represents Euler's number, and the coefficients b_0, b_1, b_2, ... b_k are the model's estimated parameters. X_1, X_2, ... X_k stand for the independent variables.

3.1.2. K-nearest neighbors (K-en yakın komşu)

KNN is a simple yet effective classification and regression algorithm. The core idea is to classify or evaluate a new data point based on the classes or values of its nearest neighbors [39]. The working principle of the algorithm is itemized below:

1. In KNN, each data point is represented by an inter-axis distance calculation. Euclidean Distance is usually used for classification. This distance measure calculates the direct distance between two data points.

2. For the algorithm to work, a value K is set. This represents the number of neighbors. For instance, if K=3, then for each new data point, the 3 closest neighbors are looked at.

3. For the given value of K, the K closest neighbors to the new data point are determined.

4. For classification, a majority vote is taken between the classes of these K neighbors. That is, the new data point is assigned to the class of the majority of its nearest neighbors.

5. For regression, an average or weighted mean is calculated between the values of these K

neighbors and the new data point is assigned an approximation to this value.

KNN classification is formally expressed as in Equation 2:

$$\acute{Y} = \text{argmax}_{y_i} \left(\sum_{i=1}^{K} I(y_i = y) \right) \quad (2)$$

In Equation 2, Ý is the predicted class of the new data point. y_i, is the i-th class of K neighbors. $I()$, is an indicator function and checks whether y_i is equal to y.

3.1.3. Support vector machine (Destek vektör makinesi)

The aim of the SVM is to find a hyperplane that best discriminates data points for classification. The working principle of SVM includes the following steps:

1. The dataset consists of labeled samples separated into two or more classes. Each sample is represented by a feature vector.

2. SVM tries to create a hyperplane between classes using feature vectors. This hyperplane is determined to best separate the classes.

3. The main goal of SVM is to maximize the distance between the closest examples of two classes, called margin. This allows for better discrimination between classes.

4. In some cases, the dataset cannot be linearly separated. In this case, SVM makes the data linearly separable by transforming it into high-dimensional space using a method called kernel trick [40].

3.1.4. Decision tree (Karar ağacı)

DT algorithm is a machine learning technique used in classification and regression problems. DT classifies data by creating simple decision rules from features in the dataset [41]. The DT algorithm includes the following steps:

1. A node is created to represent each instance in the dataset. These nodes are separated according to the values of the features in the dataset.

2. The DT aims to divide the data at each node into homogeneous subsets (branches). This splitting process involves determining the feature and threshold value that will best classify the dataset.

3. By dividing (splitting) the dataset, the DT creates a tree structure that will best classify the entire dataset as it branches.

5. When a new data point arrives, the DT classifies it using decision rules, starting from the root node and moving downwards (towards the branches).

3.1.5. Random forest (Rastgele orman)

RF is a model created by combining decision trees, an ensemble learning algorithm. This algorithm aims to obtain a more powerful and balanced model by creating multiple decision trees and combining the result of each tree. The RF algorithm creates a training dataset for each tree by randomly selecting a subset of the dataset. These subsets include random selection of features and data samples. Each tree tries to learn the relationship between the inputs and outputs of the instance. Once the decision trees are created, each tree is used to make predictions [42]. In classification problems, voting is used to determine the final prediction by taking the majority of the classes predicted by each tree. In regression problems, the final prediction is made by averaging the predicted values of each tree. The RF algorithm is resistant to overfitting and generally provides high accuracy. It can also be used to determine the order of importance of different features.

3.2. Performance Metrics (Performans Metrikleri)

Performance metrics play a crucial role in evaluating the effectiveness and efficiency of machine learning models. These metrics provide quantitative measures to assess how well a model is performing, allowing researchers and practitioners to compare different models and select the most suitable one for a particular task. In machine learning, performance metrics are used to evaluate various aspects of a model's performance, such as its accuracy, precision, recall, and F1 score. In addition to these basic metrics, other performance metrics such as receiver operating characteristic curve (ROC) and area under the ROC curve (AUC) provide further insights into the model's performance.

Performance metrics are calculated based on the values of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). These values are typically organized into a confusion matrix, which provides a tabular

representation of a model's predictions against the actual values in the dataset. TP are the cases where the model correctly predicts the positive class (e.g., approved loans). FP are the cases where the model incorrectly predicts the positive class. TN are the cases where the model correctly predicts the negative class (e.g., denied loans). FN are the cases where the model incorrectly predicts the negative class.

3.2.1. Accuracy (Doğruluk)

Accuracy refers to the proportion of instances that a classification model predicts correctly. It is calculated by the formula in Equation 3.

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

3.2.2. Precision (Kesinlik)

Precision is a performance metric that measures how many of the instances that a classification model predicts as positive are actually positive. Precision is considered an important performance metric, especially in imbalanced classification problems, that is, when the number of instances between classes is very different. It is calculated by the formula in Equation 4.

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

3.2.3. Recall (Duyarlılık)

Recall is a performance metric that measures how many TP a classification model correctly identifies. Recall is considered an important performance metric, especially when FN have a high cost. In such cases, it is important not to miss TP. It is calculated by the formula in Equation 5.

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

3.2.4. F1-score (F1-skor)

The F1-Score is the harmonic mean of the precision and recall metrics of a classification model. The F1-Score provides a balance by considering the effects of both FP and FN. F1-Score takes values between 0 and 1, with 1 representing the best performance and 0 representing the worst performance. F1-Score will have a high value when precision and recall are balanced. It is calculated with the formula in Equation 6.

$$F1 = 2 * \frac{(precision * recall)}{(precision + recall)} \quad (6)$$

3.2.5. Receiver operating characteristic curve (Alıcı işletim karakteristik eğrisi)

ROC Curve is a graphical method used to evaluate the performance of classification models. The ROC curve shows the relationship between sensitivity and specificity of a model. The ROC curve allows to visually assess the performance of the model at different classification thresholds. A model's ROC curve shows the relationship between the model's TP rate and FP rate at each threshold value. For an ideal classifier, the ROC curve approaches a diagonal line at a 45-degree angle starting from the upper left corner. Mathematically, the ROC curve is calculated with the formulas in Equation 7 and Equation 8.

$$TPR\ (True\ Positive\ Rate) = \frac{TP}{TP + FN} \quad (7)$$

$$FPR\ (False\ Positive\ Rate) = \frac{FP}{FP + TN} \quad (8)$$

3.2.6. Area under the ROC curve (ROC eğrisi altında kalan alan)

AUC is a measure of the classification performance of the model. The AUC value is between 0 and 1 and the closer it is to 1, the better the performance of the model. If the AUC value is 0.5, the model's performance is indistinguishable from random guessing. The ROC curve and AUC help to evaluate the performance of the model at different classification thresholds and provide a more comprehensive understanding of the overall performance of the model.

3.3. Dataset (Veri Seti)

"The Loan Status Prediction" dataset contains information on applicants who have previously applied for loans secured by property. Banks use various factors such as Applicant Income, Loan Amount, previous Credit History, Co-applicant Income, among others, to determine whether to approve or reject a loan application. The purpose of this dataset is to test the development of machine learning models that can predict whether a loan application will be approved or rejected for an applicant. The dataset was taken from a Hackathon on Kaggle, a platform for those interested in data science and machine learning [43]. The dataset contains 13 features and 381 records. The features and their descriptions are given in Table 1. The

correlation matrix and heat map of the dataset are given in Figure 1.

Table 1. Dataset features and descriptions (Veri seti özellikleri ve açıklamaları)

Feature	Description
Loan_ID	A unique loan ID.
Gender	Gender of the applicant (1: Male, 0: Female).
Married	Marital status of the applicant (1: Married, 0: Not Married).
Dependents	Number of dependents on the applicant.
Education	Education level of the applicant (1: Graduate, 0: Not Graduate).
Self_Employed	Whether the applicant is self-employed (1: Yes, 0: No).
ApplicantIncome	Income of the applicant.
CoapplicantIncome	Income of the co-applicant.
LoanAmount	Loan amount in thousands.
Loan_Amount_Term	Term of the loan in months.
Credit_History	Whether the applicant's credit history meets guidelines (1: Yes, 0: No).
Property_Area	Area where the applicant lives (1: Urban, 2: Semi-Urban, 3: Rural).
Loan_Status	Whether the loan was approved (1: Approved, 0: Not Approved).

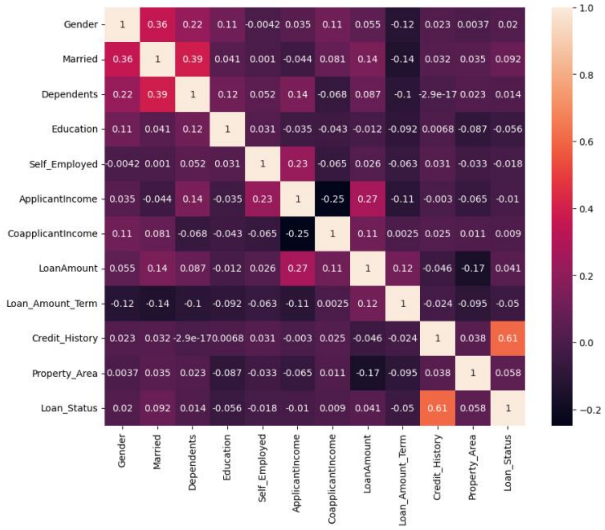

Figure 1. Heatmap of the dataset (Veri setinin ısı haritası)

3.4. Data Preparation (Veri Hazırlama)

Data Preparation is one of the most fundamental steps in the data analysis process and an important step for the success of data science projects. Data preparation is the process of making the dataset suitable for analysis and modeling. This process includes correcting missing or erroneous data in the dataset, removing redundant or repetitive data, and transforming data to improve the understandability and processability of the dataset. Proper data preparation is important to achieve more accurate results and improve model performance. Therefore, the data preparation process should be carried out rigorously.

During the data preparation phase, several crucial steps were taken to ensure the dataset was suitable for analysis and modeling. Firstly, missing values in columns such as Gender, Dependents, Self_Employed, Loan_Amount_Term, and Credit_History were addressed by filling them with the mode value, which represents the most frequently occurring value in each column. This step helped maintain the integrity of the dataset and ensured that all necessary information was available for analysis. Secondly, categorical features like Gender, Married, Education, Self-employed, and Loan status were converted into binary values. This conversion simplified the representation of these features, making them more suitable for use in machine learning algorithms. Another important transformation concerns the Loan_Amount_Term column, where a significant majority of values (around 84%) have a value of 360, indicating a long-term loan. To capture this distinction, the column was transformed such that values greater than or equal to 360 were encoded as 1, while values less than 360 were encoded as 0. Additionally, the representation of loan amounts in the Loan Amount column was adjusted to be in thousands. This adjustment was made by multiplying all values in this column by 1000, ensuring consistency in the representation of loan amounts throughout the dataset. Furthermore, for better clarity and understanding, the Education column was renamed to Graduated, and the Loan_Amount_Term column was renamed to Long_term. Lastly, the Loan_ID column, which did not provide relevant information for the analysis and modeling process, was dropped from the dataset.

In the dataset used in the study, there was an imbalance between the number of approved and unapproved loan applications, with approved loans significantly outnumbering unapproved ones (Figure 2a). To tackle this imbalance, a technique called resampling was employed for the minority class (unapproved loan applications). In the first step, the number of observations for the minority class was determined, and then random samples were taken until this number matched the number of observations for the majority class (approved loan applications). These samples were added back to the dataset to increase the number of observations for the minority class. As a result of this process, the class imbalance in the dataset was mitigated, and the model was trained on a more balanced dataset (Figure 2b).

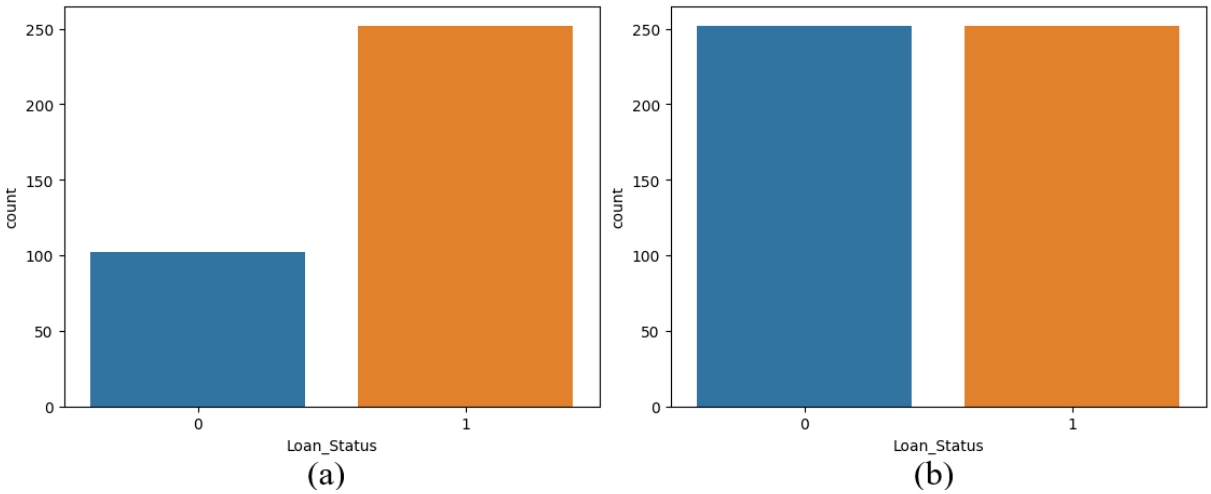

Figure 2. Class distributions: a) before resampling technique; b) after resampling technique (Sınıf dağılımları: a) yeniden örnekleme tekniğinden önce; b) yeniden örnekleme tekniğinden sonra)

3.5. Feature Selection (Özellik Seçimi)

Feature selection is a process used to determine the importance of features in a dataset and select the most appropriate ones. This process prevents overfitting by reducing the complexity of the model and increases the generalization ability of the model. It also reduces computational time by removing unnecessary features and provides better interpretability. One of the important sub-objectives of the research is to examine the effect of using various feature selection methods on the performance of the algorithms. In this context, without using feature selection, K-Best and RFE methods were used to measure the performance of the algorithms and comparisons were made.

In the K-Best method, the relationship of each feature in the dataset with the target variable is evaluated and the most important features are identified. The "K" value determines the number of features to be retained [44]. For this study, K is set as 8, which has the highest accuracy and precision rates according to the experiments. In the RFE method, a model with all features is initially created and then the model is re-evaluated by removing the least effective features one by one. This process continues until a set number of features (the number at which the model performs best) is reached. The main idea of RFE is that by removing the least influential features, the model becomes simpler and more generalizable [45]. In this study, the number of features where the model performs best was determined as 9. Figure 3 shows the flowchart of RFE.

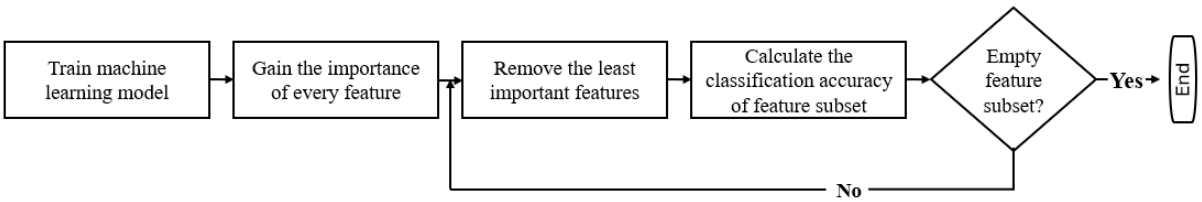

Figure 3. Flowchart of RFE (RFE Akış Şeması)

Table 2 shows the features that both methods find important for predicting the Loan_Status variable. Commonly selected features include Married, Self_Employed, ApplicantIncome, LoanAmount,

Credit_History and Education. However, the RFE method selected additional features such as Dependents and Property_Area, while the K-Best method did not find these features important.

Table 2. Selected features (Seçilen özellikler)

Method	Number of Features	Features
K-Best	8	Married, Self_Employed, ApplicantIncome, CoapplicantIncome, LoanAmount, Credit_History, Education, Loan_Status
RFE	9	Married, Self_Employed, ApplicantIncome, Dependents, LoanAmount, Credit_History, Education, Property_Area, Loan_Status

3.6. Model Setups (Model Ayarları)

In this study, machine learning classification algorithms LR, KNN, SVM, DT and RF algorithms were used. In order to create a model in the dataset, the data was divided into 75% training and 25% testing. Class distribution of training and test sets is given in Figure 4. In all algorithms, the random state was set as 42. The best hyperparameter settings for the models were determined using the GridSearchCV method. This method selects the best performing hyperparameters by trying different combinations within the specified hyperparameter ranges [46]. Table 3 lists the best hyperparameter settings determined using GridSearchCV for the machine learning models used in the study.

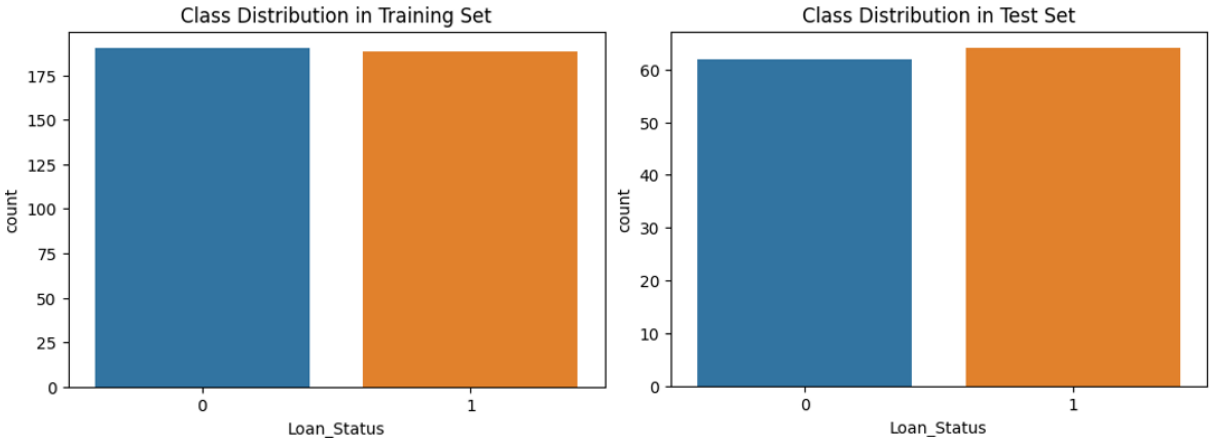

Figure 4. Class distribution in training and test sets (Eğitim ve test setlerindeki sınıf dağılımı)

Table 3. Hyperparameter settings for algorithms (Algoritmalar için hiper parametre ayarları)

Model	Hyperparameters	Settings
LR	C, Class Weight, Max Iter, Penalty, Solver	1, {0: 0.15, 1: 0.85}, 100, l1, liblinear
KNN	Algorithm, Metric, N Neighbors, Weights	auto, euclidean, 20, distance
SVM	C, Class Weight, Gamma, Kernel	1, balanced, 10, rbf
DT	Criterion, Max Depth, Min Samples Split	entropy, 20, 2
RF	Max Depth, Max Features, Min Samples Leaf, Min Samples Split, N Estimators	20, auto, 1, 2, 200

In this study, Python programming language was used for data analysis and model testing. Basic data processing libraries such as pandas and NumPy were used for data analysis, while scikit-learn was preferred for model building and testing. Visualization libraries such as matplotlib and seaborn were used to analyze the results and the findings were presented graphically. All these processes were carried out in the Jupyter Notebook development environment.

4. EXPERIMENTAL STUDY AND FINDINGS (DENEYSEL ÇALIŞMA VE BULGULAR)

In the experimental phase of the research, firstly, Exploratory Data Analysis (EDA) was conducted to examine the dataset in detail. As a result of this analysis, important inferences about the dataset were obtained. Then, to evaluate the performance of the machine learning algorithms used in the research, various measurements were made, and the performances of the algorithms were compared. Figure 5 shows the matrix containing the graphs

showing the loan application success of the applicants according to the features.

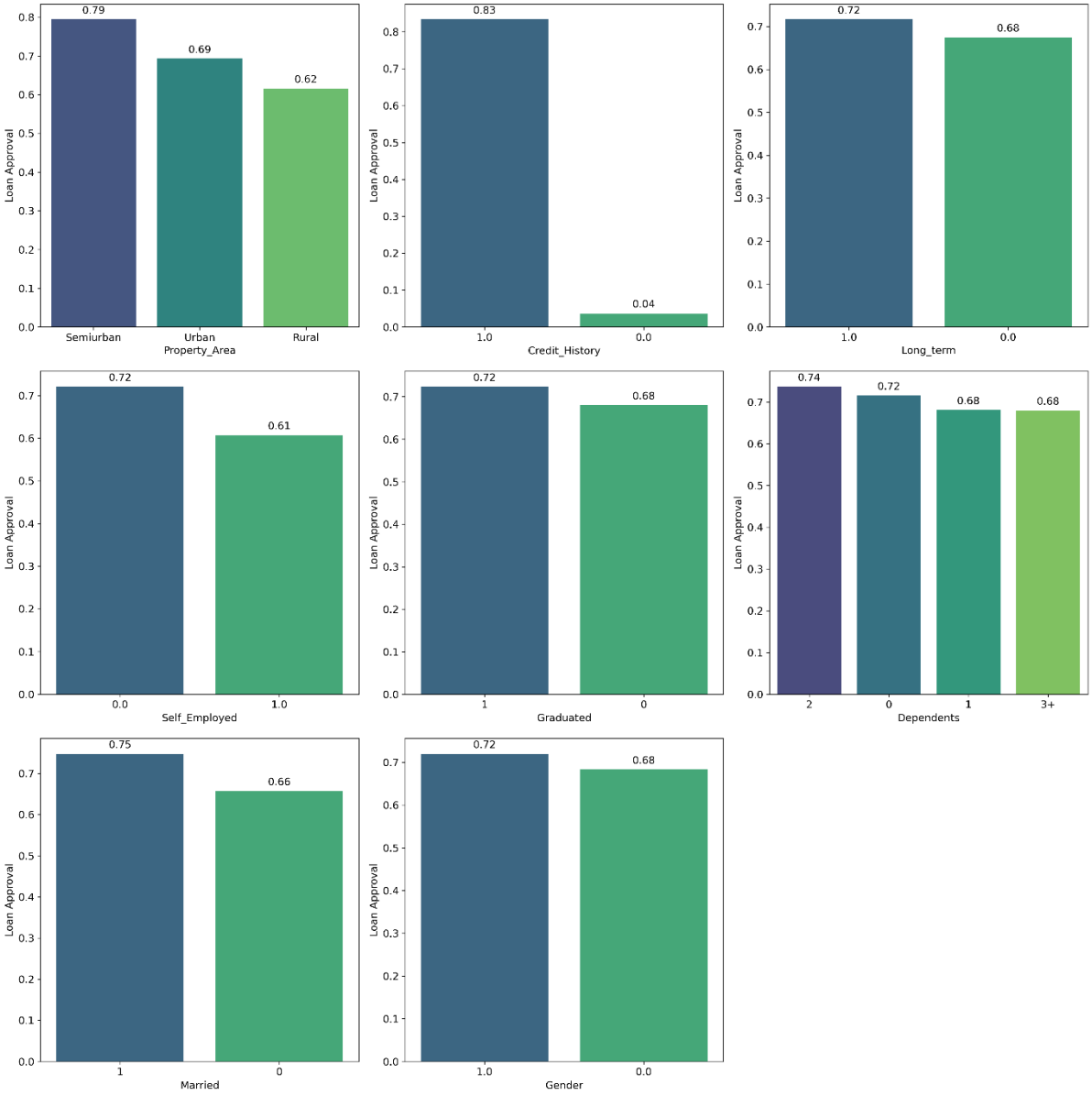

Figure 5. Distribution of features by loan approval status (Özelliklerin kredi onay durumuna göre dağılımı)

In the analysis of loan approval predictors, it was observed that males were more likely to have their loan applications approved compared to females. Similarly, married individuals had a higher likelihood of loan approval than unmarried individuals. Graduates were also more likely to get their loans approved compared to non-graduates. On the other hand, self-employed individuals were less likely to have their loan applications approved. Long-term loans showed a higher probability of approval than short-term loans. Moreover, a strong credit history significantly increased the chances of loan approval compared to a weak credit history. Additionally, residents in semiurban areas had a higher likelihood of loan approval than urban residents, while rural residents had the lowest probability. Furthermore, individuals with 2 dependents had a higher probability of loan approval compared to those with 0, 1, or more than 3 dependents. Figure 6's graph illustrates the relationship between credit history and applicant income, and its impact on loan approval.

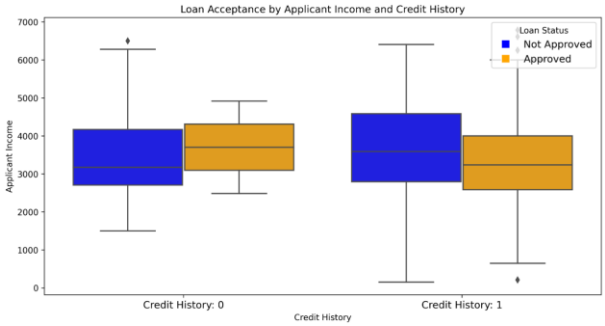

Figure 6. Applicant income and credit history impact on loan approval (Başvuru sahibinin geliri ile kredi geçmişinin kredi onayı üzerindeki etkisi)

Figure 6 shows that the income ranges of applicants with a positive credit history are higher than those with a negative credit history, the applications of applicants with higher income are more likely to be approved, and the average income of applicants are close to each other according to credit history or loan approval status. Figure 7 shows the relationship between the amount of loan applied for and the region of the applicant's residence and its impact on loan approval.

When the relationship between loan amount and loan acceptance rate is analyzed, it is observed that the loan acceptance rate decreases as the loan amount increases. Moreover, there are differences in loan amounts according to the type of region where applicants live. In particular, loan amounts of applicants living in urban areas are higher than those of applicants living in semi-urban and rural areas.

Table 4 shows the model performance values measured without feature selection. Table 5 shows the model performance values obtained using the K-Best method and Table 6 shows the model performance values obtained using the RFE method.

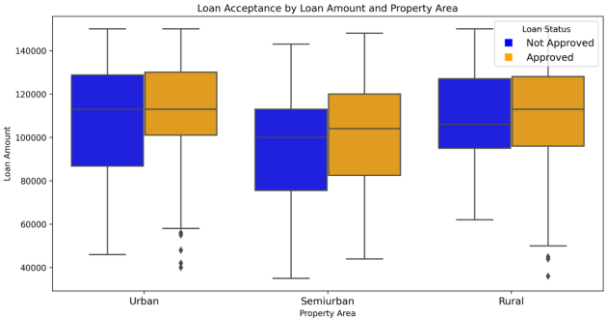

Figure 7. Loan approval by loan amount and property area (Kredi tutarı ve mülk bölgesine göre kredi onayı)

Table 4. Model performance measured without feature selection (Özellik seçimi yapılmadan modellerin performans değerleri)

Model	Accuracy	Precision	Recall	F1-Score
LR	0.7423	0.75	0.98	0.84
KNN	0.7926	0.77	0.96	0.82
SVM	0.7435	0.71	0.99	0.83
DT	0.7511	0.72	0.98	0.80
RF	0.8241	0.79	0.97	0.85

Table 5. Model performance values with K-best method (K-best yöntemi ile modellerin performans değerleri)

Model	Accuracy	Precision	Recall	F1-Score
LR	0.7730	0.79	0.85	0.82
KNN	0.7409	0.82	0.87	0.85
SVM	0.8098	0.78	0.82	0.80
DT	0.8171	0.76	0.80	0.78
RF	0.8857	0.85	0.90	0.88

Table 6. Model performance values with RFE method (RFE yöntemi ile modellerin performans değerleri)

Model	Accuracy	Precision	Recall	F1-Score
LR	0.7950	0.79	0.86	0.83
KNN	0.7841	0.80	0.88	0.86
SVM	0.9378	0.91	0.94	0.92
DT	0.8702	0.82	0.85	0.83
RF	0.9768	0.92	0.95	0.93

Tables 4, 5 and 6 show how the use of different feature selection methods (K-Best and RFE) affects model performance. According to the model performances measured without feature selection (Table 4), the model with the highest accuracy value is the RF algorithm, with an accuracy value of

0.8241 and an F1-Score value of 0.85. The performance of the models built with features selected by the K-Best method (Table 5) resulted in an increase in the accuracy value for the RF model (0.8857, F1-Score: 0.88). The performance of the models with features selected by RFE method (Table 6) showed a more significant increase in accuracy for the RF model (0.9768, F1-Score: 0.93). The RF model shows higher accuracy, precision, recall and F1-Score values compared to the other models in all tables.

It is seen that the features selected with the RFE method significantly improve the model performance. The models created with the features selected with RFE show higher performance than the models created using the K-Best method or all features. This shows that the feature selection method can improve model performance and unnecessary features can negatively affect model performance.

Table 7. Performance results of models with cross-validation (Çapraz doğrulama ile modellerin performans sonuçları)

Model	Cross-Validation (K-fold) Accuracy Rate	Train, Test, and Validation Accuracy Rate	AUC Rate	SD
LR	0.7672	0.7950	0.85	0.0562
KNN	0.7557	0.7841	0.95	0.0643
SVM	0.9441	0.9378	0.94	0.0361
DT	0.8782	0.8702	0.87	0.0316
RF	0.9771	0.9768	0.96	0.0309

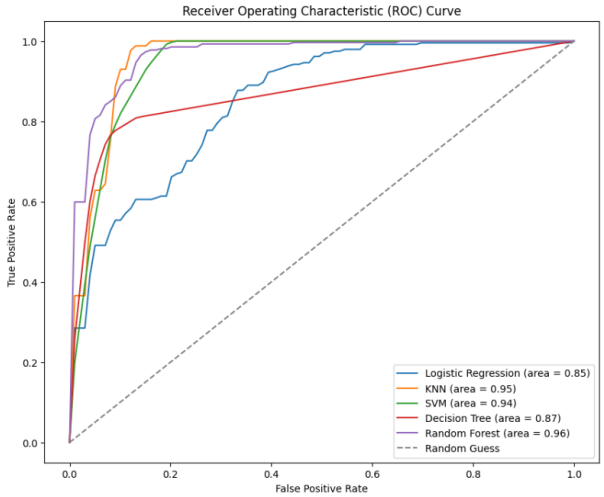

Figure 8. ROC curves (ROC eğrileri)

In this part of the study, the results obtained with cross-validation and Train, Test, and Validation methods are compared. Since the highest model performances were obtained with the RFE method, these analyses were conducted on the features selected with the RFE method. Table 7 shows the accuracy rates, AUC rates and standard deviation (SD) values of the models. Figure 8 shows the ROC curve graph of the models. The K-Fold method was used for the accuracy rate obtained by the cross-validation method. In this method, the dataset is divided into K parts and each part is used as a test set and the remaining part is used as a training set. In this way, K different models are created, and accuracy values are obtained [47]. In this research, the number of folds was set as 10 for all algorithms. Train, Test, and Validation Accuracy Rate is the accuracy rate obtained after training the model by separating the dataset into a single training and test set. In addition, the AUC ratio measures the classification performance of the model. The SD value indicates the variability of the model's performance. According to Table 7, the RF algorithm performs the best in predicting loan approval, with a cross-validation rate of 0.9771, Train, Test, and Validation accuracy rate of 0.9768, AUC rate of 0.96, and SD of 0.0309.

5. DISCUSSION (TARTIŞMA)

In this study, machine learning models for loan approval prediction are developed and their performance is evaluated. In addition, EDA was conducted on the dataset to reveal various information about loan approval. The findings of the paper are discussed and compared with similar studies in the literature.

The impact of demographic factors on loan approval is the starting point of this study and this has been frequently discussed in the literature. Dansana et al. [48] and Stavins [49], in line with the findings of this study, find that married individuals are more likely to be approved for a loan than single individuals. There are several possible reasons for this. First, married individuals may generally have more stable sources of income and may be more likely to plan and manage their income according to the needs of the household. Moreover, married individuals often have joint income with their spouse, which may make their loan repayments stronger and increase the likelihood that banks will approve their loan applications. However, given that married individuals bear more responsibility for their families, their sense of responsibility for

making timely loan payments may be higher. These reasons may increase the likelihood that married individuals are more likely to have their loan applications approved than single individuals.

Similar to the findings of this study, Escalante et al. [50] report that men's loan applications are more likely to be approved than women's. Among social factors, income inequality may lead men to generally have higher incomes [51], which may make loan payments more secure and increase the likelihood of loan applications being approved. Moreover, the fact that men are generally perceived as more competent and trustworthy in financial matters may be associated with traditional gender roles and perceptions [52], which may contribute to favorable evaluation of loan applications.

The effect of education level on loan approval is also a topic examined in the literature. In their study, Bandyopadhyay [53] found that graduates have a higher chance of loan approval than non-graduates. This finding is consistent with the results of this study and shows that education level is an important factor on loan approval. The fact that graduates have a higher chance of loan approval compared to non-graduates can be attributed to several reasons. Education level is often associated with income level [54]. A higher level of education may imply a higher income and financial stability. This may lead to the perception that loan repayments will be more reliable. Moreover, educational attainment is associated with financial literacy and financial planning skills [55]. This can lead to more careful and informed loan applications, which in turn increases the likelihood of loan approval.

The impact of financial factors on loan approval is also an important issue in the literature. For instance, Ravina [56] finds that loan applications of individuals with higher income are more likely to be approved. Similarly, Netzer et al. [57] report that individuals with a strong credit history are more likely to have their loan applications approved. These results are consistent with the findings of this study and suggest that financial factors such as income level and credit history have an impact on loan approval. High income is associated with financial reliability and loan repayment capacity [58]. This assumes that the loan applicant is more likely to repay the loan. Therefore, it is a factor that increases the likelihood of loan approval. Moreover, higher income provides a stronger position when applying for a loan as an indicator of financial stability and security. In addition, the impact of financial history on loan approval is also significant. Having a strong credit history indicates that previous loan payments have been made regularly and on time, which is an important factor in the favorable evaluation of the loan application [59]. In this context, the impact of financial factors such as income level and credit history on loan approval are important criteria evaluated by financial institutions and are scrutinized during the loan application process.

The study reveals a significant finding regarding the relationship between loan amount and loan acceptance rate, indicating a decrease in the acceptance rate as the loan amount increases. This trend suggests that higher loan amounts are perceived as riskier, leading to more thorough scrutiny and a lower probability of approval. This observation aligns with previous studies [60]. Furthermore, the study highlights regional disparities in loan amounts. Specifically, applicants residing in urban areas tend to have higher loan amounts compared to those in semi-urban and rural areas. This discrepancy may be attributed to the higher cost of living in urban areas, resulting in a greater need for loans among urban residents.

For loan approval forecasting, this study provides important findings on how the use of different feature selection methods (K-Best and RFE) affects model performance. According to the model performances measured without feature selection, the RF algorithm has the highest accuracy of 82.4% (F1-Score: 0.85). However, these results reflect the case where all features are used. An increase in the performance of all models created with the features selected with the K-Best method was observed. The most successful algorithm in the K-Best method was RF with an accuracy of 88.5% (F1-Score: 0.88). The highest values were obtained in the performances of the models obtained with the features selected by RFE method. In fact, compared to the model performances measured without feature selection, the performances of the models obtained with the features selected with the RFE method increased up to 26% and the performance of the models created with the features selected with the K-Best method increased up to 19%. The most successful algorithm in predicting loan approval was the RF algorithm with an accuracy of 97.6%, and this was achieved with the features selected with the RFE method (F1-Score: 0.93). These results show that feature selection can improve model performance and redundant features can negatively affect model performance. Similar studies in the literature also emphasize the importance of feature selection. Feature selection can prevent overfitting by reducing the complexity of the model and increase the generalization ability

of the model [61]. Moreover, by removing redundant features, the model can achieve higher performance [62]. Similar to the findings of this study, Meshref [34], in his study on loan approval prediction, found that feature selection improves the performance of machine learning models rather than using all features. Sarizeybek and Sevli [47] achieved an average performance increase of 7% with the K-Best method in their study on customers' propensity to take loans. Similarly, in this study, the performance of the models increased by about 19% when feature selection was made with the K-Best method. Apart from loan approval prediction, various feature selection methods have been found to improve the performance of models in image processing and speech processing [63], disease risk prediction [64], bank marketing and human activity recognition [65]. Many studies from different fields in the literature show that feature selection methods improve the performance of machine learning models. In Table 8, most of the scores obtained from studies on loan approval prediction are achieved without using feature selection methods. It is expected that these studies can achieve higher scores by using feature selection methods. In addition, when the accuracy of the models is calculated using K-Fold, one of the cross-validation methods, the RF algorithm reaches the highest accuracy value with 97.71% (AUC: 0.96). Cross-validation is an important technique used to assess how well the model fits real-world data. Instead of dividing the dataset into training and test sets, this technique allows for a more reliable evaluation of the model's performance by dividing the data into different subsets. For instance, Adagbasa et al. [66] evaluated the performance of a deep learning model using K-Fold cross-validation. In their study, instead of a single training-test partition, they divided the data into 5 different subsets with 5-fold cross-validation. In this way, they analyzed in more detail how the model performed on each subset. Their results show that the model performs well overall but underperforms on some subsets. On the other hand, Valavi et al. [67] evaluated the performance of a classification model using a single training-test partition. Their results showed that the model fits the training data well but the test data poorly. This suggests that the model cannot accurately assess how well it fits real-world data. In addition, in line with the findings of this study, Sarizeybek and Sevli [47] found that the success rates obtained using 10-fold cross-validation were significantly higher than a single training-test partitioning in their study predicting customers' propensity to take loans.

There are some notable studies in the literature that perform loan approval prediction on various datasets. Table 8 shows the comparison of models and accuracy rates of loan approval prediction studies. From the table, the model proposed in this study differs significantly from many studies in the field and achieves a high classification accuracy of 97.71%. Only the study by Saini et al. [26] surpassed this study with an accuracy of 98.04%, but despite the high accuracy rates, the F1-Score of their model is 0.85. A high F1-Score indicates that both the classification accuracy and the FP and FN rates of the model are balanced. In this study, the F1-Score of the RF algorithm was measured as 0.93. This result indicates that the model has a high accuracy rate as well as a good balance between FP and FN predictions. Although Saini et al. [26] achieved 98.04% accuracy in their study, the F1-Score of 0.85 suggests that the model focuses on a certain class and neglects other classes or shows an unbalanced performance. In addition to this information, when the table is analyzed, it is understood that tree-based algorithms achieve a higher success in the loan approval prediction task compared to other algorithms. Especially ensemble methods such as RF allow many decision trees to come together to form a stronger model. This reduces the noise in the dataset and allows for more robust predictions.

Table 8. Comparison of accuracy rates of loan approval prediction models in the field (Alandaki kredi onay tahmin modellerinin doğruluk oranlarının karşılaştırılması)

Reference	Year	Model	Accuracy
Saini et al. [26]	2023	RF	98.04%
Uddin et al. [31]	2023	ET	86.64%
Tumuluru et al. [37]	2022	RF	81.00%
Ramachandra et al. [33]	2021	LR	86.00%
Singh et al. [27]	2021	RF	77.00%
Diwate et al. [28]	2021	SVM	81.00%
Alaradi & Hilal [29]	2020	DT	97.25%
Meshref [34]	2020	AdaBoost	83.98%
Sheikh et al. [36]	2020	LR	81.00%
Current study	-	RF	97.71%

5. CONCLUSION (SONUÇ)

The findings of this research have important implications for the development and performance evaluation of machine learning models for loan approval prediction. The EDA study examined the effects of demographic factors, education level, financial status, and other factors on loan approval. When the impact of demographic factors on loan approval is analyzed, it is found that married individuals and individuals with higher income are more likely to be approved for loans. Similarly, it was found that loan applications of men were more likely to be approved than those of women. Education level was also found to be an important factor in loan approval. It has been determined that loan applications of individuals with a university degree are more likely to be approved. In terms of financial factors, it is observed that loan applications of individuals with high income and a strong credit history are approved more frequently. In addition, in the relationship between loan amount and loan acceptance rate, it was found that the loan acceptance rate decreased as the loan amount increased.

In evaluating the performance of machine learning algorithms, it was observed that the use of feature selection methods (K-Best and RFE) significantly improved model performance. The performance of the models obtained with the features selected with the RFE method reached the highest accuracy, precision, and F1-Score values. RF algorithm showed the highest accuracy, precision, recall and F1-Score values in all model performances measured without and after feature selection. In addition, the accuracy rate obtained with the K-Fold method, which is one of the cross-validation methods in SVM and RF models, is significantly higher than the accuracy rate obtained with the Train, Test, and Validation method. The proposed models can be used by financial institutions and lenders to evaluate loan applications. By providing an automated evaluation process, these models can enable faster review of loan applications and faster feedback to customers. This can increase customer satisfaction. The models can also reduce credit risks, thereby reducing costs and operational risks for institutions. As a result, using the proposed models can both reduce costs and improve customer experience for financial institutions, which can lead to a significant competitive advantage in the sector.

DECLARATION OF ETHICAL STANDARDS (ETİK STANDARTLARIN BEYANI)

The author of this article declares that the materials and methods they use in their work do not require ethical committee approval and/or legal-specific permission.

Bu makalenin yazarı çalışmalarında kullandıkları materyal ve yöntemlerin etik kurul izni ve/veya yasal-özel bir izin gerektirmediğini beyan ederler.

AUTHORS' CONTRIBUTIONS (YAZARLARIN KATKILARI)

***Vahid SİNAP*:** He conducted the experiments, analyzed the results, and performed the writing process.

Deneyleri yapmış, sonuçlarını analiz etmiş ve makalenin yazım işlemini gerçekleştirmiştir.

CONFLICT OF INTEREST (ÇIKAR ÇATIŞMASI)

There is no conflict of interest in this study.

Bu çalışmada herhangi bir çıkar çatışması yoktur.

REFERENCES (KAYNAKLAR)

[1] B. Huang and L. C. Thomas, "Credit card pricing and impact of adverse selection," *J. Oper. Res. Soc.*, vol. 65, no. 8, pp. 1193-1201, 2014.

[2] V. Leninkumar, "The relationship between customer satisfaction and customer trust on customer loyalty," *Int. J. Acad. Res. Bus. Soc. Sci.*, vol. 7, no. 4, pp. 450-465, 2017.

[3] M. Siles, S. D. Hanson, and L. J. Robison, "Socio-economics and the probability of loan approval," *Appl. Econ. Perspect. Policy*, vol. 16, no. 3, pp. 363-372, 1994.

[4] J. E. Stiglitz and A. Weiss, "Incentive effects of terminations: Applications to the credit and labor markets," *Am. Econ. Rev.*, vol. 73, no. 5, pp. 912-927, 1983.

[5] S. T. Bharath, S. Dahiya, A. Saunders, and A. Srinivasan, "Lending relationships and loan contract terms," *Rev. Financial Stud.*, vol. 24, no. 4, pp. 1141-1203, 2011.

[6] S. M. Livingstone and P. K. Lunt, "Predicting personal debt and debt repayment: Psychological, social and economic determinants," *J. Econ. Psychol.*, vol. 13, no. 1, pp. 111-134, 1992.

[7] N. W. Hillman, "College on credit: A multilevel analysis of student loan default," *Rev. High. Educ.*, vol. 37, no. 2, pp. 169-195, 2014.

[8] The Banks Association of Türkiye, "Consumer Loans and Housing Loans," 2023. [Online]. Available: https://www.tbb.org.tr/Content/Upload/istatistikiraporlar/ekler/4227/Tuketici_Kredileri_Raporu-Eylul_2023.pdf

[9] S. Carter, E. Shaw, W. Lam, and F. Wilson, "Gender, entrepreneurship, and bank lending: The criteria and processes used by bank loan officers in assessing applications," *Entrepreneurship Theory and Practice*, vol. 31, no. 3, pp. 427-444, 2007.

[10] C. Parkan and M. L. Wu, "Measurement of the performance of an investment bank using the operational competitiveness rating procedure," *Omega*, vol. 27, no. 2, pp. 201-217, 1999.

[11] J. S. Chiou, "The antecedents of consumers' loyalty toward Internet service providers," *Inf. & Manage.*, vol. 41, no. 6, pp. 685-695, 2004.

[12] R. S. Swift, *Accelerating customer relationships: Using CRM and relationship technologies*. Prentice Hall Professional, 2001.

[13] S. Sachan, J. B. Yang, D. L. Xu, D. E. Benavides, and Y. Li, "An explainable AI decision-support-system to automate loan underwriting," *Expert Syst. Appl.*, vol. 144, p. 113100, 2020.

[14] B. Lepri, N. Oliver, E. Letouzé, A. Pentland, and P. Vinck, "Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges," *Philos. Technol.*, vol. 31, pp. 611-627, 2018.

[15] J. F. Martínez Sánchez and G. Pérez Lechuga, "Assessment of a credit scoring system for popular bank savings and credit," *Contad. y Adm.*, vol. 61, no. 2, pp. 391-417, 2016.

[16] R. Parasuraman, M. Mouloua, R. Molloy, and B. Hilburn, "Monitoring of automated systems," in *Automation and human performance*, CRC Press, 2018, pp. 91-115.

[17] M. McKay, "Best practices in automation security," in *2012 IEEE-IAS/PCA 54th Cement Industry Technical Conference*, May 2012, pp. 1-15.

[18] D. Bertsimas, and J. Dunn, "Optimal classification trees," *Machine Learning*, vol. 106, pp. 1039-1082, 2017.

[19] A. J. Wyner, M. Olson, J. Bleich, and D. Mease, "Explaining the success of adaboost and random forests as interpolating classifiers," *Journal of Machine Learning Research*, vol. 18, no. 48, 1-33, 2017.

[20] Z. G. Liu, Q. Pan, and J. Dezert, "A new belief-based K-nearest neighbor classification method," *Pattern Recognition*, vol. 46, no. 3, pp. 834-844, 2013.

[21] I. W. Tsang, J. T. Kwok, P. M. Cheung, and N. Cristianini, "Core vector machines: Fast SVM training on very large data sets," *Journal of Machine Learning Research*, vol. 6, no. 4, 2005.

[22] J. Wu, X. Y. Chen, H. Zhang, L. D. Xiong, H. Lei, and S. H. Deng, "Hyperparameter optimization for machine learning models based on Bayesian optimization," *Journal of Electronic Science and Technology*, vol. 17, no. 1, pp. 26-40, 2019.

[23] A. Janecek, W. Gansterer, M. Demel, and G. Ecker, "On the relationship between feature selection and classification accuracy," in *New challenges for feature selection in data mining and knowledge discovery*, PMLR, 2008, pp. 90-105.

[24] E. Kadam, A. Gupta, S. Jagtap, I. Dubey, and G. Tawde, "Loan approval prediction system using logistic regression and CIBIL score," in *2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Jul. 2023, pp. 1317-1321.

[25] A. S. Kadam, S. R. Nikam, A. A. Aher, G. V. Shelke, and A. S. Chandgude, "Prediction for loan approval using machine learning algorithm," *Int. Res. J. Eng. Technol. (IRJET)*, vol. 8, no. 04, pp. 4089-4092, 2021.

[26] P. S. Saini, A. Bhatnagar, and L. Rani, "Loan approval prediction using machine learning: A comparative analysis of classification algorithms," in *2023 3rd Int. Conf. Adv. Comput. Innov. Technol. Eng. (ICACITE)*, May 2023, pp. 1821-1826.

[27] V. Singh, A. Yadav, R. Awasthi, and G. N. Partheeban, "Prediction of modernized loan approval system based on machine learning approach," in *2021 Int. Conf. Intell. Technol. (CONIT)*, Jun. 2021, pp. 1-4.

[28] Y. Diwate, P. Rana, and P. Chavan, "Loan Approval Prediction Using Machine Learning," *Int. Res. J. Eng. Technol. (IRJET)*, vol. 8, no. 05, 2021.

[29] M. Alaradi and S. Hilal, "Tree-based methods for loan approval," in *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*, Oct. 2020, pp. 1-6.

[30] V. S. Kumar, A. Rokade, and S. MS, "Bank loan approval prediction using data mining technique," *Int. Res. J. Modern. Eng. Technol. Sci.*, vol. 2, no. 05, pp. 965-970, 2020.

[31] N. Uddin, M. K. U. Ahamed, M. A. Uddin, M. M. Islam, M. A. Talukder, and S. Aryal, "An ensemble machine learning based bank loan approval predictions system with a smart application," *International Journal of Cognitive*

Computing in Engineering, vol. 4, pp. 327-339, 2023.

[32] J. Tejaswini, T. M. Kavya, R. D. N. Ramya, P. S. Triveni, and V. R. Maddumala, "Accurate loan approval prediction based on machine learning approach," *J. Eng. Sci.*, vol. 11, no. 4, pp. 523-532, 2020.

[33] H. V. Ramachandra, G. Balaraju, R. Divyashree, and H. Patil, "Design and simulation of loan approval prediction model using AWS platform," in *2021 Int. Conf. Emerg. Smart Comput. Informatics (ESCI)*, Mar. 2021, pp. 53-56.

[34] H. Meshref, "Predicting loan approval of bank direct marketing data using ensemble machine learning algorithms," *Int. J. Circuits. Syst. Signal Process.*, vol. 14, pp. 914-922, 2020.

[35] A. Gupta, V. Pant, S. Kumar, and P. K. Bansal, "Bank Loan Prediction System using Machine Learning," in *2020 9th International Conference System Modeling and Advancement in Research Trends (SMART)*, Dec. 2020, pp. 423-426.

[36] M. A. Sheikh, A. K. Goel, and T. Kumar, "An approach for prediction of loan approval using machine learning algorithm," in *2020 Int. Conf. Electron. Sustainable Commun. Syst. (ICESC)*, Jul. 2020, pp. 490-494.

[37] P. Tumuluru, L. R. Burra, M. Loukya, S. Bhavana, H. M. H. CSaiBaba, and N. Sunanda, "Comparative Analysis of Customer Loan Approval Prediction using Machine Learning Algorithms," in *2022 Second Int. Conf. Artif. Intell. Smart Energy (ICAIS)*, Feb. 2022, pp. 349-353.

[38] F. Y. Osisanwo, J. E. T. Akinsola, O. Awodele, J. O. Hinmikaiye, O. Olakanmi, and J. Akinjobi, "Supervised machine learning algorithms: classification and comparison," *Int. J. Comput. Trends Technol. (IJCTT)*, vol. 48, no. 3, pp. 128-138, 2017.

[39] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.

[40] M. N. Murty and R. Raghava, "Kernel-based SVM," in *Support vector machines and perceptrons: Learning, optimization, classification, and application to social networks*, 2016, pp. 57-67.

[41] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, pp. 20-28, 2021.

[42] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random forests and decision trees," *Int. J. Comput. Sci. Issues (IJCSI)*, vol. 9, no. 5, p. 272, 2012.

[43] Kaggle, "Loan Status Prediction," Available: https://www.kaggle.com/datasets/bhavikjikadara/loan-status-prediction/data.

[44] M. Cinelli et al., "Feature selection using a one dimensional naïve Bayes' classifier increases the accuracy of support vector machine classification of CDR3 repertoires," *Bioinformatics*, vol. 33, no. 7, pp. 951-955, 2017.

[45] A. S. Paramita and S. V. Winata, "A comparative study of feature selection techniques in machine learning for predicting stock market trends," *J. Appl. Data Sci.*, vol. 4, no. 3, pp. 147-162, 2023.

[46] D. M. Belete and M. D. Huchaiah, "Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results," *Int. J. Computers and Applications*, vol. 44, no. 9, pp. 875-886, 2022.

[47] A. T. Sarizeybek and O. Sevli, "A comparative analysis of bank customers' loan propensity using machine learning methods," *J. Intell. Syst. Theory Appl.*, vol. 5, no. 2, pp. 137-144, 2022. [Online]. Available: https://doi.org/10.38016/jista.1036047

[48] D. Dansana, S. G. K. Patro, B. K. Mishra, V. Prasad, A. Razak, and A. W. Wodajo, "Analyzing the impact of loan features on bank loan prediction using Random Forest algorithm," *Engineering Reports*, vol. 6, no. 2, p. e12707, 2024.

[49] J. Stavins, "Credit card borrowing, delinquency, and personal bankruptcy," *New Engl. Econ. Rev.*, pp. 15-30, 2000.

[50] C. L. Escalante, J. E. Epperson, and U. Raghunathan, "Gender bias claims in farm service agency's lending decisions," *J. Agric. Resour. Econ.*, pp. 332-349, 2009.

[51] S. Kuznets, "Economic growth and income inequality," in *The gap between rich and poor*, Routledge, 2019, pp. 25-37.

[52] D. Oh, E. A. Buck, and A. Todorov, "Revealing hidden gender biases in competence impressions of faces," *Psychol. Sci.*, vol. 30, no. 1, pp. 65-79, 2019.

[53] A. Bandyopadhyay, "Studying borrower level risk characteristics of education loan in India," *IIMB Management Review*, vol. 28, no. 3, pp. 126-135, 2016.

[54] C. Jamir and T. Z. Ezung, "Impact of education on employment, income, and poverty in Nagaland," *Int. J. Res. Econ. Soc. Sci. (IJRESS)*, vol. 7, no. 9, pp. 50-56, 2017.

[55] A. Lusardi, "Financial literacy and the need for financial education: Evidence and implications," *Swiss J. Econ. Stat.*, vol. 155, no. 1, pp. 1-8, 2019.

[56] E. Ravina, "Love & loans: The effect of beauty and personal characteristics in credit markets," SSRN Working Paper, 2019.
[57] O. Netzer, A. Lemaire, and M. Herzenstein, "When words sweat: Identifying signals for loan default in the text of loan applications," *J. Marketing Res.*, vol. 56, no. 6, pp. 960-980, 2019.
[58] H. K. Mutegi, P. W. Njeru, and N. T. Ongesa, "Financial literacy and its impact on loan repayment by small and medium entrepreneurs," *Int. J. Econ. Commerce Manag.*, vol. 3, no. 3, pp. 1-28, 2015.
[59] M. Li, A. Mickel, and S. Taylor, "'Should this loan be approved or denied?': A large dataset with class assignment guidelines," *J. Stat. Educ.*, vol. 26, no. 1, pp. 55-66, 2018.
[60] R. Van Ooijen, and M. C. Van Rooij, "Mortgage risks, debt literacy and financial advice," *Journal of Banking & Finance*, vol. 72, pp. 201-217, 2016.
[61] B. Venkatesh, and J. Anuradha, "A review of feature selection and its methods," *Cybernetics and Information Technologies*, vol. 19, no. 1, pp. 3-26, 2019.
[62] Y. Xiao, C. Xing, T. Zhang, and Z. Zhao, "An intrusion detection model based on feature reduction and convolutional neural networks," *IEEE Access*, vol. 7, pp. 42210-42219, 2019.
[63] P. Dhal, and C. Azad, "A comprehensive survey on feature selection in the various fields of machine learning," *Applied Intelligence*, vol. 52, no. 4, pp. 4543-4581, 2022.
[64] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A review of feature selection methods for machine learning-based disease risk prediction," *Frontiers in Bioinformatics*, vol. 2, pp. 927312, 2022.
[65] R. C. Chen, C. Dewi, S. W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," *Journal of Big Data*, vol. 7, no. 1, pp. 52, 2020.
[66] E. G. Adagbasa, S. A. Adelabu, and T. W. Okello, "Application of deep learning with stratified K-fold for vegetation species discrimation in a protected mountainous region using Sentinel-2 image," *Geocarto International*, vol. 37, no. 1, pp. 142-162, 2022.
[67] R. Valavi, G. Guillera-Arroita, J. J. Lahoz-Monfort, and J. Elith, "Predictive performance of presence-only species distribution models: a benchmark study with reproducible code," *Ecological Monographs*, vol. 92, no. 1, pp. e01486, 2022.