**THE REPUBLIC OF TURKEY**
**BAHÇEŞEHİR UNIVERSITY**

# DATA MINING AND MACHINE LEARNING APPLICATIONS IN BANK AUDIT

**Master's Thesis**

**BARBAROS ŞENOCAK**

**İSTANBUL, 2021**

**T.C.**

**BAHÇEŞEHİR UNIVERSITY**

**GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES**

**ARTIFICIAL INTELLIGENCE MASTER'S PROGRAM**

# DATA MINING AND MACHINE LEARNING APPLICATIONS IN BANK AUDIT

**Master's Thesis**

**BARBAROS ŞENOCAK**

**Supervisor: ASSIST. PROF. DR. TAMER UÇAR**

**İSTANBUL, 2021**

T.C.
BAHCESEHIR UNIVERSITY
GRADUATE SCHOOL

02/08/2021

**MASTER THESIS APPROVAL FORM**

| | |
|---|---|
| **Program Name:** | ARTIFICIAL INTELLIGENCE MASTER'S PROGRAM |
| **Student's Name and Surname:** | BARBAROS SENOCAK |
| **Name Of The Thesis:** | DATA MINING AND MACHINE LEARNING APPLICATIONS IN BANK AUDIT |
| **Thesis Defense Date:** | 24.06.2021 |

This thesis has been approved by the Graduate School which has fulfilled the necessary conditions as Master thesis.

**Assoc. Prof. Dr. Burak KÜNTAY**
**Institute Director**

This thesis was read by us, quality and content as a Master's thesis has been seen and accepted as sufficient.

| | **Title/Name** | **Signature** |
|---|---|---|
| **Thesis Advisor's** | Assist. Prof. Dr. Tamer UÇAR | |
| **Member's** | Assist. Prof. Dr. Serkan AYVAZ | |
| **Member's** | Assist. Prof. Dr. Atınç YILMAZ | |

## ACKNOWLEDGMENT

First of all, I would like to thank my thesis supervisor Assist. Prof. Dr. Tamer UÇAR for his contributions during my thesis work.

And I would like to thank my managers Ali ÇALIŞ and Emrah ŞENTÜRK who supported me during my thesis work and also my colleagues Gülnihal ÇİÇEKDAĞI and Özge ŞENYÜZ.

Finally, I would like to thank my family for always being there for me.

İstanbul, 2021 Barbaros ŞENOCAK

## ABSTRACT

DATA MINING AND MACHINE LEARNING APPLICATIONS IN BANK AUDIT

Barbaros Şenocak

Artificial Intelligence Master's Program

Supervisor: ASSIST. PROF. DR. TAMER UÇAR

May 2021, 46 Pages

The digitalization process deeply affects economic systems, sectors, institutions and it becomes a necessity for institutions to keep up with technology at the point reached. Today, one of the concepts that has changed with digital transformation is the concept of "data". Every click made while surfing the internet, likes on social media, every transaction made on the phone have become usable data. It is critical for organizations to make computer-aided analysis in the enormously increasing data and to anticipate opportunities and risks. With this increase in data entries, the digitalization process accelerated the transaction volume and this process led to the change in the banking sector and the risk-oriented audit approach in banking. The increasing transaction volume and speed in the banking sector profoundly affect the risk perception. It has become important to use data analysis and data mining at a level to obtain optimum benefit by performing risk-based audits in internal audit methodologies applied in banks.

The aim of this thesis is to provide information about the machine learning algorithms that can be applied in internal audits with the technological changes in the banking field and to create models that will benefit the risk-oriented execution of audit activities practically through machine learning models. In the study, models that will contribute to audit techniques in the process of examining credits through different algorithms such as multiple regression, decision trees and random forest have been created and the success of the established models has been compared with different algorithms. As a result of the study, it has been concluded that integrating machine learning models into audit processes will contribute to risk-focused audit activities.



**Keywords**: Banking, Risk Based Auditing, Data Mining, Machine Learning

# ÖZET


BANKA DENETİMİNDE VERİ MADENCİLİĞİ VE MAKİNE ÖĞRENİMİ UYGULAMALARI

Barbaros Şenocak

Yapay Zeka Yüksek Lisans Programı

Tez Danışmanı: Dr. Öğretim Üyesi Tamer UÇAR

Mayıs 2021, 46 Sayfa

Dijitalleşme süreci ekonomik sistemleri, sektörleri, kurumları derinden etkilemekte, gelinen noktada teknolojiye ayak uydurmak kurumlar için bir zorunluluk haline gelmektedir. Günümüzde dijital dönüşüm ile birlikte değişen kavramlardan biri de 'veri' kavramıdır. İnternette gezinirken yapılan her tıklama, sosyal medyadaki beğeniler, telefondan yapılan her işlem kullanılabilir veri haline dönüşmüş ve kurumlar için muazzam ölçüde artan veri içerisinde bilgisayar destekli analizler yapmak, fırsat ve riskleri önceden görebilmek kritik düzeyde öneme sahiptir. Veri girişlerinde yaşanan bu artışla beraber dijitalleşme süreci işlem hacmini hızlandırmış, bu süreç bankacılık sektörünün ve bankacılıkta risk odaklı denetim anlayışının da değişimine yol açmıştır. Bankacılık sektöründe artan işlem hacmi ve hız risk algısını derinden etkilemektedir. Bankalarda uygulanan iç denetim metodolojilerinde risk odaklı denetim yaparak optimum faydayı elde edecek seviyede veri analizi ve veri madenciliği kullanmak önemli hale gelmiştir.

Bu tezde amaç bankacılık alanında yaşanan teknolojik değişimlerle birlikte banka içi denetimlerde uygulanabilecek makine öğrenim algoritmaları hakkında bilgi vermek, makine öğrenim modelleri üzerinden uygulamalı olarak denetim faaliyetlerinin risk odaklı yürütülmesine fayda sağlayacak modeller oluşturmaktır. Çalışmada çoklu regresyon, karar ağaçları, rasgele orman gibi farklı algoritmalar üzerinden kredilerin tetkiki sürecinde denetim tekniklerine katkı sağlayacak modeller oluşturulmuş ve kurulan modellerin başarısı farklı algoritmalar ile karşılaştırılmıştır. Çalışmanın sonucunda makine öğrenimi modellerinin denetim süreçlerine entegre edilmesinin risk odaklı denetim faaliyetlerine katkı sağlayacağı kanaatine varılmıştır.



**Anahtar kelimeler:** Bankacılık, Risk Odaklı Denetim, Veri Madenciliği, Makine Öğrenimi

**CONTENTS**

## TABLES

## FIGURES

## ABBREVIATIONS

KKB : Kredi Kayıt Bürosu

KNN : K-Nearest Neighboor

RF : Random Forest

SVM : Support Vector Machines

TİDE : Türkiye İç Denetim Enstitüsü

# 1. INTRODUCTION

Undoubtedly, the banking sector is one of the sectors where the effects of the digital transformation process in the world are best seen. Due to the development of computer-aided systems, the increase in banking activities through internet banking and mobile applications, the great change in financial technologies and the developments in artificial intelligence technologies have caused serious changes in banking activities in recent years.

Changes in banking activities have caused internal audit activities to change, replacing traditional audit methods with computer-oriented audit methods. With the rapid change in technological developments, the number of transactions and the speed of transactions increased and it became very important to direct the available resources to risk-oriented audit activities in the audits made for the bank activities. With regard to risk-focused audit activities, it has become very important to benefit from statistical models and to integrate artificial intelligence and in particular, machine learning models into audit procedures.
The aim of this thesis is to show practically the contribution of machine learning models to audit activities and show the models that give the best results comparatively through different machine learning models (Logistic Regression, Naive Bayes, Random Forest, etc.). The last part of the study also explains how the analysis results can be integrated into control systems.

In the continuation of the introduction of the thesis, brief information about the concept of data, data mining techniques, artificial intelligence, the effects of artificial intelligence technologies on the internal audit process and machine learning models will be given and then will proceed with the data and method parts. Finally, the study will be concluded by evaluating the findings obtained.

## 1.1 DATA CONCEPT

Data is the raw (unprocessed) real piece of information. (Bosij et al. 2003) Although data does not mean anything by itself, it turns into information as a whole and by being processed. An important issue regarding the data concept is the rapid change in technology, the decrease in storage costs and the use of it in bulk. This situation creates an environment for the data to be processed and transformed into meaningful information, institutions can integrate and use this information in all business processes.

The enormous increase in data volume has also increased the importance of the concept of "big data". It has become very important to organize and process data and to draw meaningful results. Big data must share at least one of the four characters listed below, called 4V. (Hurwitz and Kirsch 2018)

a) Extremely large *Volumes* of data
b) The ability to move that data at a high *Velocity* of speed
c) An ever-expanding *Variety* of data sources
d) *Veracity* so that data sources truly represent truth

The concept of big data refers to complex data as well as regular data. Nowadays, voice recordings, photos, online shopping, credit card expenditures, etc. all data are processed in batches, analyzed and used by decision makers. Developments in information technologies increase the speed of data processing. This enables large-scale data to be analyzed and transformed into useful information in a much shorter time.

## 1.2 DATA MINING

Data mining is the process of discovering meaningful patterns and rules from large amounts of data. (Linoff & Berry 2011) Data mining is useful to extract meaningful and valuable information from big data heaps, allows foreseeing the opportunities and risks that may occur and enabling decision-making mechanisms to work according to these opportunities and risks. Although statistics is the basis of data mining, nowadays,

advancement of data mining activities by using techniques such as machine learning, natural language processing and image recognition is becoming widespread in different sectors. Data mining techniques are used for many analyses such as measuring customer satisfaction, evaluating the profitability and performance of the company, making future planning of the company by creating predictive models for the future.

There are five basic stages in the functioning of the data mining process. These steps are given below (Restivo 1999).

i. Defining the Problem

ii.. Preparation of Data

iii. Establishment and Evaluation of the Model

iv. Using the Model

v. Model Monitoring

If these processes are briefly mentioned; the goal should be determined exactly at the beginning of the data mining process. The process should be started by determining the benefits and costs that this goal will provide.

One of the most difficult steps of the Data Mining process is the data preparation step. It is critically important that the analyzed data fits for purpose and it is inevitable that improperly generated datasets will lead to erroneous results. At this stage, it is necessary to control the quality of the data and whether it is suitable for analysis. The data preparation stage consists of the steps of collecting, combining and cleaning and transforming in itself (Domouchel 1999). While the data cleaning process ensures that the data that affects errors or distributions in the data set is cleaned and the data is made suitable for analysis, the data transformation process refers to the arrangement of the data according to the format suitable for analysis.

In the process of establishing and evaluating the model, the validity of the established model is tested and it is decided whether it is suitable for use or not. When the model passes the test phase successfully and starts to be used, the process of monitoring the

model begins. In this process, the results of the model are checked at regular intervals and the model continues to be used or necessary corrections are made.

## 1.3 ARTIFICIAL INTELLIGENCE TECHNOLOGIES

Artificial intelligence technologies have started to be used widely in many institutions in different sectors. The rapid development of information technologies, the decrease in data storage costs and the developments in the system infrastructures regarding the database have made it possible to reach large-scale data. In an increasingly competitive environment, it has become of vital importance for organizations to analyze these large-scale data correctly and to evaluate the opportunities and risks that may occur.

It is necessary to understand what artificial intelligence is before giving examples of machine learning models which are one of the artificial intelligence applications on internal audit processes. Artificial intelligence is hardware and software systems that learn like the human brain, establish cause-effect relationships, adapt to new situations, analyze, make decisions, and operate complex reasoning processes (TİDE 2017).

The benefits of artificial intelligence technologies to an organization and the risks that they may cause should be discussed beforehand and these benefits and risks should be evaluated. If the benefits and risks of artificial intelligence technologies are briefly mentioned;

a) The ability to reveal detailed analyzes that cannot be seen in large data sets due to the limitations of human analysis,

b) While the analyzes made by people take a long time while working with big data, the analysis made by the machines save a lot of time, (For example; after loading a machine learning algorithm created for the credits process into the system, newly created data sets can be automatically included in the algorithm and analyzed regularly without causing loss of time and effort.)

c) The predictions made using artificial intelligence algorithms can produce more detailed and successful results, in this case, the institutions contribute significantly to future planning by showing the risks and opportunities for decision makers,

d) Facilitating "personalization" processes with artificial intelligence applications, developing suggestion systems according to customers' behaviour and allowing marketing activities to be done on an individual basis,
e) Artificial intelligence technologies make a significant contribution to institutions in many processes such as increasing customer satisfaction and profitability with many technological processes like voice and image recognition, payment systems, mobile banking, etc.,

To mention the benefits of artificial intelligence as well as its risks;

a) The artificial intelligence models which are created by humans produce biased results and do not reflect the truth,
b) Installation of artificial intelligence technologies can be costly, as well as maintenance and repair costs,
c) When working with large data sets, there is a risk of human errors in the process,
d) The technologies created with artificial intelligence have the risk of causing damage when adequate tests and controls are not applied,
e) The potential of artificial intelligence technologies to create unemployment problems on the one hand, by changing the content of the business concept in general, while providing new job opportunities,
f) In cases where the investments made in this field do not provide benefits as expected, it is the risk that the cost exceeds the benefit and the institution suffers a loss.

## 1.4 THE IMPORTANCE OF ARTIFICIAL INTELLIGENCE TECHNOLOGIES FOR INTERNAL AUDIT

In terms of internal audit, it is very important that artificial intelligence technologies are widely used in organizations. In addition to the benefits arising from these systems, the risks that may arise are of great importance in terms of internal audit. Internal audit units must first of all have to master artificial intelligence technologies and applications and be careful about the risks that may arise in this regard. In addition, internal audit should

provide assurance in managing the risks related to the reliability of the underlying algorithms and the data on which these algorithms are based (TİDE 2017).

In addition, internal audit units should participate in all stages of the process by carrying out audit activities in areas where artificial intelligence technologies are applied, conduct studies for emerging and potential risks and establish control methods. It is very important to control the reliability of the data used in the application of artificial intelligence algorithms, data confidentiality, the errors that may arise from these data and the errors that may occur during the modelling and analysis stages and to evaluate the results.

At this stage, the creation of data used in artificial intelligence algorithms by humans creates problems such as data bias and carelessness. It is quite possible to have unintentional human errors as well as human mistakes that can be made intentionally. In this sense, internal audit units should evaluate whether there are manual and systemic controls that will minimize human errors and also whether these controls are effectively implemented.

Internal audit units should integrate these technologies into their internal audit practices while conducting risk assessments for the process of creating and implementing artificial intelligence algorithms throughout the organization.

When the change in internal audit processes in banks is examined, it is seen that the rapid changes in information technologies have created serious changes in the implementation of internal audit and this trend of change will continue in the upcoming period. With this process of change, the understanding of auditing with traditional methods is gradually decreasing and methods based on information technologies are increasing. In addition, it is seen that while manual controls and on-site inspection methods were used at the beginning in the implementation part, centralized control activities have gained great importance today and manual controls have been replaced by systemic controls. With the development of computer-aided systems, preventive controls are now applied quickly by computers and erroneous transactions can be prevented before the process begins.

Although the digital transformation process and technological changes have brought systemic controls to the forefront, the increasing volume and speed of transactions on the other side of the event have made it necessary to analyze correctly and interpret "big data" in the banking and audit sector. It is necessary to identify risky areas by mining data in large data sets, create sample sets that have the ability to represent the population and thus implement the necessary control mechanisms by identifying risky areas as soon as possible.

In order for artificial intelligence technologies to be used in internal audit of banks, first of all, there should be a systemic infrastructure where big data can be used and stored. Correct, consistent and healthy acquisition of data is only possible with a properly functioning system infrastructure.

Integration of artificial intelligence technologies into internal control systems is a very important issue for risk-oriented auditing. While the analysis made on big data provide significant time savings, it can also contribute to the emergence of detailed information that people cannot see in large data sets. The aim of this thesis is the integration of machine learning algorithms, which is a sub-branch of artificial intelligence technologies to control processes.

## 1.5 MACHINE LEARNING

Machine learning algorithms are a sub-form of artificial intelligence technology and are systems that perform learning over data instead of a programming path. Training data is placed in a model and a result is obtained by training the algorithm. In machine learning, it is aimed for the system to recognize the model by analyzing the training data with statistical and logical methods, to gain experience and make a decision. Machine learning models are frequently used in analysis such as predicting the future, clustering and classifying data accurately. In this section, brief information about commonly used machine learning techniques will be given and then the literature review section will be passed.

### 1.5.1 Machine Learning Techniques

#### 1.5.1.1 Supervised learning

In supervised learning method, examples are used by tagging to teach the algorithm (Silva 2017). The meaning of labelling the samples is that the results are included in the data set. The purpose of this method is to reveal the relationship network within the data by considering the results of the model. As an example of the supervised learning method, if the data set contains information about the characteristics that may cause the existing customers to leave the bank and whether they leave the bank for an analysis of the probability of leaving the bank of newly acquired customers of a bank branch, the existing customers are labelled. In supervised learning, the model takes into account whether the current customers leave the bank and produces results accordingly.

#### 1.5.1.2 Unsupervised learning

In unsupervised learning methods, unlike supervised learning, data are not labelled. Unsupervised learning models are very useful, especially when working with big data and if the results of the model to be applied are unpredictable. As an example, customers going to a bank branch can be divided into segments according to their transactions. Initially, there is no information about which segment the customers are in, but in unsupervised learning, customers belong to a cluster according to their behaviour. The basic point in unsupervised learning is to take into account the similarities of the data to each other.

#### 1.5.1.3 Reinforcement learning

Reinforcement learning is based on trial and error. The algorithm develops itself to get better results by re-evaluating each step.

### 1.5.2 Machine Learning Algorithms

In this section, brief information about commonly used machine learning algorithms will be given and detailed information about the algorithms to be used within the scope of the thesis study is given in the data and method section.

#### 1.5.2.1 Linear regression

Linear regression is a statistical method used to examine the relationship between two variables. It is used to estimate the value of another variable according to the value of one variable.

#### 1.5.2.2 Logistic regression

The logistic regression model is a classification algorithm and enables the data to be classified according to the interaction between the independent variables and the dependent variable. Logistic regression is one of the most popular machine learning models today. The most important difference of logistic regression from linear regression is that the dependent variable takes a continuous value in linear regression, while it takes categorical values in logistic regression. The aim here is to determine the class to which the variables belong.

#### 1.5.2.3 Naive Bayes classifier

This algorithm, named after the British mathematician Thomas Bayes, is based on probability and classifies according to this basis. Naive Bayes is an algorithm that is easy to use, can work with little data and give successful results. In addition, the Naive Bayes method, which is one of the methods with high data separation capacity, is widely used because it evaluates each independent variable within itself.

#### 1.5.2.4 Support vector machines

Support vector machines is an algorithm commonly used for classification purposes. Support vector machines are an effective method for parsing nonlinear data sets as well as linear data sets. Support Vector Machines draw a hyperplane from the points to which the data is closest to optimally separate the data from each other. It tries to maximize the distance between the data allocated to the classes. This algorithm is applied in many different fields such as text classification, image decomposition and sound analysis.

#### 1.5.2.5 Decision trees

The decision tree algorithm, one of the supervised learning models, is a very useful model in the analysis of complex data sets. When working with big data, it greatly facilitates the understanding of the data by dividing the data into small groups according to their specific characteristics.

#### 1.5.2.6 K-Nearest Neighbor algorithm

K-Nearest Neighbor algorithm, which is mostly used for clustering and classification processes, is used to place the data in a data set into a group according to their distance to the k value. This method, which classifies newly incoming data using distance measurement methods such as Euclidean, Manhattan, and Minkowski, is widely used due to its successful results in complex data sets.

#### 1.5.2.7 Random Forest algorithm

The random forest algorithm, which can also be used in classification and regression models, creates a random forest instead of a tree. The number of trees in the forest, i.e. the decomposed data, directly affects the result. The algorithm uses the votes (decisions) from these trees to achieve the best result.

#### 1.5.2.8 Gradient Boosting algorithm

This algorithm, which can be used in classification and regression methods, provides a strong prediction model by combining weak prediction models. In this model, which generally uses decision tree structures, it is intended to minimize the loss function and correct the errors in each step in the previous step.

#### 1.5.2.9 Artificial Neural Networks

Artificial neural networks (ANNs) are computer systems developed with the aim of automatically realizing the capabilities of the human brain, such as the ability to generate and discover new information through learning, without any assistance. (Öztemel, 2003) Artificial neural networks are very useful in complex problems, when the relationships between variables are not clear. It is one of the most widely used algorithms due to its ability to work with missing data and patterns, to use different algorithms and to give successful results in many fields.

# 2. LITERATURE REVIEW

Although the history of the studies on machine learning models goes back many years, it is seen that these technologies have found many application areas in different sectors, especially in recent years. Today, machine learning models significantly affect decision-making mechanisms in different areas and this increases the number of studies conducted in this field. Considering that machine learning models are applied in different fields, in this thesis, the application of machine learning models in banking and audit on previous studies and research on data mining have been focused.

In the article "The Use of Artificial Neural Networks as an Analytical Examination Technique in the Control Planning Stage: A Company Application" published by Serpil Senal and Mahmut Sami Öztürk in 2014, it was revealed that the use of artificial neural networks in the audit process is an effective method as an analytical examination technique.

In the study published by Hilmi Kırlıoğlu and İsmail Fatih Ceyhan in 2014 on "The Use of Data Mining as a Preliminary Analytical Review Technique in Financial Statement Auditing: Borsa Istanbul Application", the financial data of 40 companies were analyzed on Borsa Istanbul and the financial data of the companies were analyzed as a preliminary indicator in financial audits. It has been tried to prove that it is possible to use it.

In a study conducted by Kasım Baynal and Aslı Çalış in 2016, a cluster analysis was performed with data belonging to 200 customers of a bank branch, and accordingly, the effectiveness of cluster analysis was evaluated while determining sales strategies.

Regina Esi Turkson, Edward Yeallakuor Baagyere, Gideon Evans Wenya tried to find the best model for the credit worthiness of their customers by testing the effectiveness of different machine learning models through bank data in 2016, and it was concluded that there were no serious differences between algorithms.

Dr. Mohammad Miyan introduced the methods and application areas of data mining in the banking sector in the article "Applications of Data Mining in Banking Sector" published in 2017 and data mining applications in the future of banking were discussed. In the “Extreme Learning Machine for Credit Risk Analysis” article published by Mais Haj Qasem and Loai Nemer, model performance measurements were made by comparing Naive Bayes, Decision Tree and Extreme Learning Machine (ELM) models to determine the risk levels.

In the study titled "The Use and Evaluation of Big Data in Audits" published by İlhan Özdemir and Şeref Sağıroğlu in 2018, the use of big data in audits, how it can be integrated into audit plans and how to conduct economic and risk-oriented audits were discussed.

Nana Kwame Gyamfi, Dr Jamal-Deen Abdulai conducted a study on fraud detection with the Support Vector Machine (SVM-S) in 2018. In the study, the normal and abnormal behaviours of the customers were tried to be determined and it was revealed that the SVM model was an effective method in such studies.

Spătariu Elena Cerasela, Gheorghiu Gabriela evaluated the changes that may occur in the internal audit process and the importance of adaptation to this process in the “Internal Audit Role in Artificial Intelligence” article published in 2018.

Gabe Dickey, Sandra Blanke and Lloyd Seaton, in their article "Machine Learning in Auditing" published in 2019, discussed the benefits and problems of adding machine learning models to audit processes.

In the article titled “Trust in the European Central Bank: Using Data Science and Predictive Machine Learning Algorithms” published by Andrii Skirka, Bogdan Adamyk, Oksana Adamyk, Mariana Valytska in 2020, it was aimed to predict the trust in the European Central Bank by considering different factors.

In the article "Predictive Model for Cutting Customers Migration from banks: Based on machine learning classification algorithms" published by Kahlid Alkhatib, Sayel Abualigah, the use of machine learning models to prevent customers from leaving the bank is explained. In the study, different classification models were tested and their performance was measured comparatively.

Nisha P. Shetty, Jayashree Shetty, Rohil Narula, Kushagra Tandona made comparisons of different machine learning models on anomaly detection in their study titled "Comparison study of machine learning classifiers to detect anomalies" in 2020 and Principal Component Analysis (PCA) was used for feature selection in the study. It was stated that anomaly detection changes according to the type and characteristics of the data while comparing model and the use of hybrid models in real-time data can increase the accuracy rates.

In the article "Implications of Implementation of Artificial Intelligence in the Banking Business in Relation to the Human Factor" published by Krunoslav Ris, Željko Stanković, Zoran Ž. Avramović in 2020, it was examined that artificial intelligence applications reduce human dependency in banking and increase profitability and performance.

The purpose of this study is to show the benefits of machine learning algorithms in making the internal audit procedures of banks risk-oriented in addition to the existing studies and to measure the model success by comparing the models that can be created in the audit field through different algorithms.

The article and thesis information mentioned in the literature review section are shown in Table 2.1.

## Table 2.1: List of Publications Examined in the Literature Section

| AUTHOR | THESIS-ARTICLE | YEAR | METHOD |
|---|---|---|---|
| Serpil Senal, Mahmut Sami Öztürk | Using of Artificial Neural Networks as an Analytical Investigation Technique at Audit Planning Stage: A Company Application | 2014 | Artificial Neural Networks |
| Hilmi Kırlıoğlu, İsmail Fatih Ceyhan | The Use of Data Mining as a Pre-Analytical Review Technique in Financial Statement Auditing: Borsa İstanbul Application | 2014 | K-Nearest Neighbor and Naive Bayes Method |
| Kasım Baynal, Aslı Çalış | An Application in Banking Sector with Cluster Analysis in Data Mining | 2016 | K-Means and Clustering Analysis |
| Regina Esi Turkson, Edward Yeallakuor Baagyere, Gideon Evans Wenya | A Machine Learning Approach for Predicting Bank Credit Worthiness | 2016 | Comparison of different machine learning methods |
| Dr. Mohammad Miyan | Applications of Data Mining in Banking Sector | 2017 | A general study on data mining in the banking sector |
| Mais Haj Qasem and Loai Nemer | Extreme Learning Machine for Credit Risk Analysis | 2018 | Extreme Learning Machine, Naive Bayes, Decision Tree, Multi-Layer Perceptron |
| İlhan Özdemir, Şeref Sağıroğlu | An Assessment on the Use of Big Data in Audits | 2018 | General assessment on big data |
| Nana Kwame Gyamfi, Dr Jamal-Deen Abdulai | *Bank Fraud Detection Using Support Vector Machine* | 2018 | Support Vector Machine |
| Spătariu Elena Cerasela, Gheorghiu Gabriela | Internal Audit Role in Artificial Intelligence | 2018 | General evaluation of the role of artificial intelligence in internal auditing |
| Gabe Dickey, Sandra Blanke ve Lloyd Seaton | Machine Learning in Auditing | 2019 | Overview of machine learning's role in internal auditing |
| Andrii Skirka, Bogdan Adamyk, Oksana Adamyk, Mariana Valytska | Trust in the European Central Bank: Using Data Science and predictive Machine Learning Algorithms | 2020 | Logistic Regression, Decision Tree, Random Forest, Neural Network |
| Kahlid Alkhatib, Sayel Abualigah | Predictive Model for Cutting Customers Migration from banks: Based on machine learning classification algorithms | 2020 | Decision Tree, Random Forest, Naïve Bayes, K Nearest Neighbor and Logistic Regression. |
| Nisha P. Shetty, Jayashree Shetty, Rohil Narula, Kushagra Tandona | Comparison study of machine learning classifiers to detect anomalies | 2020 | Random Forest, Neural Network, SVM, Decision Tree |
| Krunoslav Ris, Željko Stanković, Zoran Ž. Avramović | Implications of Implementation of Artificial Intelligence in the Banking Business in Relation to the Human Factor | 2020 | General evaluation of the application of artificial intelligence in the banking sector |

# 3. MATERIAL AND METHODS

In this section, information on the selection of data used for analysis, data processing processes to organize data and machine learning algorithms that are planned to be used are included. In addition, information on the methods to be used in the evaluation of model results is also included in this section.

In this thesis, two different model studies have been conducted as personal loans and commercial loans. First of all, in order to examine the conversion of loans to non-performing loans in the personal loans process, the personal characteristics of the customers (age, gender, marital status, income, etc.) and the KKB scores created by the Credit Registry Bureau (KKB) by taking into account the customers' past payment habits were determined as independent variables. The effect of existing features on the conversion of loans extended to customers into non-performing loans was analyzed using logistic regression, naive bayes and k-nearest neighbor, support vector machines, decision tree and random forest algorithms and model comparison was made.

In the second part, the probability of firm risks turning into non-performing loans is analyzed over 43 financial ratios (current ratio, cash rate, etc.) for commercial loan processes.

The last part of the study includes information on how the results obtained from machine learning models can be used in internal audit processes.

In the continuation of the thesis, this study, which is carried out for two different processes, namely personal loans and commercial loans, is divided into two as modelling for the personal loans process and modelling for the commercial loans process.

## 3.1 MODELLING STUDY FOR THE PERSONAL LOANS PROCESS

In this section; data collection, data preparation and modelling studies related to personal loans process will be mentioned.

### 3.1.1 Data Collecting

The data used in the study on the personal loans process were created using the characteristics of a bank's individual customers and the data set included 1000 customers. While the risks of some of these customers have turned into non-performing loans, other customers do not have any risks that have turned into non-performing loans. Within the scope of the study, 11 different types of customer information were included in the analysis process. As can be seen in Table 3.1, the sector in which customers work (public-private), education level, marital status, gender, retirement or active employee, first-time loan availability, credit amount, KKB score, salary from the bank, the age of the customer and the total income of the customer are used in the process.

**Table 3.1: Sample Data of Individual Customers**

| Customer | Customer 1 | Customer 2 | Customer 3 | Customer 4 | Customer 5 |
|---|---|---|---|---|---|
| Credit status | 0 | 1 | 1 | 0 | 0 |
| Gender | 1 | 0 | 1 | 0 | 1 |
| Marital status | 4 | 2 | 1 | 2 | 2 |
| Level of education | 1 | 4 | 3 | 4 | 5 |
| Public-Private | 0 | 1 | 1 | 0 | 0 |
| Retired-Employee | 1 | 0 | 0 | 1 | 1 |
| Credit-Amount | 50000 | 246000 | 30000 | 400000 | 240000 |
| First loan -not | 0 | 0 | 1 | 0 | 0 |
| KKB_Score | 1656 | 1072 | 1039 | 1846 | 1799 |
| Salaried -not | 1 | 1 | 1 | 0 | 1 |
| Total-Income | 3798 | 4688 | 2952 | 3750 | 7911 |
| Customer-Age | 79 | 42 | 35 | 61 | 76 |

Some variables in the data set were categorized and their categorical values were taken into account in the analysis process. For example; In the gender line of the customer, the

value "0" is used instead of the word "male" and "1" is used instead of the word woman, while the categorical values in the education level line show the customer's education level such as secondary school, high school, university graduate.

### 3.1.2 Data Preparation Process

In the study on personal loans, it has been tried to classify whether the loans are transferred to non-performing loans or not. Within the scope of the study, logistic regression, naive bayes and k-nearest neighbor, decision tree, random forest and support vector machine algorithms were used and the model results of different algorithms were compared. Before using machine learning models, the data preparation process is crucial to the success and consistent results of the model results. Data sets can often contain many different problems such as null values, outliers, variables with high correlation. These should be checked and necessary arrangements should be made. In addition, checking the characteristics of the data set, the range of variables, whether the data distribution is normal or not are important points that should be checked in the preliminary process.

**Table 3.2: Features of Individual Customer Data**

| | mean | std | min | 25 percent | 50 percent | 75 percent | max |
|---|---|---|---|---|---|---|---|
| Credit status | 0.50 | 0.50 | 0.00 | 0.00 | 0.50 | 1.00 | 1.00 |
| Gender | 0.85 | 0.36 | 0.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Marital status | 2.016 | 0.56 | 1 | 2 | 2 | 2 | 4 |
| Level of education | 4.221 | 16.422 | 0 | 4 | 4 | 6 | 10 |
| Public-Private | 0.992 | 0.57 | 0 | 1 | 1 | 1 | 2 |
| Retired-Employee | 0.173 | 0.37 | 0 | 0 | 0 | 0 | 1 |
| Credit-Amount | 20.481 | 27.046 | 1000 | 38750 | 100000 | 250000 | 2040000 |
| First loan -not | 0.145 | 0.35 | 0 | 0 | 0 | 0 | 1 |
| KKB_Score | 1342.86 | 39.093 | 0 | 1106.5 | 1351 | 1651 | 1900 |
| Salaried -not | 0.759 | 0.42 | 0 | 1 | 1 | 1 | 1 |
| Total-Income | 5.426 | 67.73 | 630 | 3029 | 4779 | 66.145 | 200000 |
| Customer-Age | 48.517 | 12.351 | 21 | 39 | 49 | 58 | 86 |

Table 3.2 contains information about the customers with individual loans and the characteristics of the customers in the data set have been examined before proceeding to the analysis part. For example; before the analysis, information such as the average and standard deviation of the customers' total income gives an idea about which income group the customers included in the data set predominantly belong to. Similarly, the age range of customers etc. information is also available. While working with large data sets, data visualizations are performed and the visualizations of the relationship between the values in the data set and the dependent variable make a great contribution to the understanding of the data set.

During the preliminary preparation process of the analysis, it was first checked whether the data was suitable for the analysis and whether there were blank values or outliers in the data set. During the controls, it was observed that there were no null values in the data set and when the distribution of the data set was examined, it was understood that the data showed a distribution close to normal and the outlier data did not affect the analysis result.

The high correlation between the data, the distribution of the data can affect the result of the analysis and change the accuracy of the models. As can be seen in Table 3.3, in the data set, the correlation between data belonging to customers (age, KKB_Score, total income, etc.) was examined, and it was observed that there was a relatively high correlation between the data in the Public- Private column and the Retired-Employee columns in the data set.

**Table 3.3: Correlation Table**

| | Gender | Marital status | Level of education | Public-Private | Retired-Employee | Credit-Amount | Firstloan-not | KKB_Score | Salaried-not | Total-Income | Customer-Age |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Gender | 1.00 | -0.03 | -0.12 | 0.06 | -0.10 | 0.00 | -0.01 | -0.09 | 0.00 | -0.02 | -0.06 |
| Marital status | -0.03 | 1.00 | 0.00 | -0.06 | 0.13 | 0.05 | -0.12 | 0.10 | 0.05 | 0.06 | 0.22 |
| Level of education | -0.12 | 0.00 | 1.00 | -0.06 | -0.07 | 0.07 | -0.02 | 0.10 | 0.02 | 0.29 | 0.03 |
| Public-Private | 0.06 | -0.06 | -0.06 | 1.00 | -0.64 | -0.06 | 0.20 | -0.13 | -0.24 | -0.01 | -0.42 |
| Retired-Employee | -0.10 | 0.13 | -0.07 | -0.64 | 1.00 | -0.01 | -0.04 | 0.13 | 0.02 | -0.16 | 0.42 |
| Credit-Amount | 0.00 | 0.05 | 0.07 | -0.06 | -0.01 | 1.00 | -0.12 | -0.03 | 0.06 | 0.16 | 0.02 |
| First loan - not | -0.01 | -0.12 | -0.02 | 0.20 | -0.04 | -0.12 | 1.00 | -0.07 | -0.37 | -0.17 | -0.19 |
| KKB_Score | -0.09 | 0.10 | 0.10 | -0.13 | 0.13 | -0.03 | -0.07 | 1.00 | 0.05 | 0.17 | 0.30 |
| Salaried - not | 0.00 | 0.05 | 0.02 | -0.24 | 0.02 | 0.06 | -0.37 | 0.05 | 1.00 | 0.14 | 0.23 |
| Total-Income | -0.02 | 0.06 | 0.29 | -0.01 | -0.16 | 0.16 | -0.17 | 0.17 | 0.14 | 1.00 | 0.14 |
| Customer-Age | -0.06 | 0.22 | 0.03 | -0.42 | 0.42 | 0.02 | -0.19 | 0.30 | 0.23 | 0.14 | 1.00 |

### 3.1.3 Data Analysis and Modelling Process

After the data pre-examination and editing stage, it was seen that the data set was suitable for analysis and the modelling stage was initiated. In this section, different models have been studied and compared. In this section, more detailed information has been given about the applied models and information on the criteria used to measure model success is also included.

It is known that the Logistic Regression method gives successful results in studies on classification methods in the literature. This method was used first in the analysis part.

#### 3.1.3.1 Logistic regression

The purpose of the logistic regression model is to create a model that describes the relationship between the response variable and the explanatory variable or variables when the response variable is bivalent or classified (Önder and Cebeci 2002). The relationship between variables does not need to be linear in logistic regression. In cases where the dependent variable takes categorical values instead of continuous values, this model is frequently used in classification studies.

The main differences that distinguish logistic regression from linear regression are;
a) While the dependent variable to be predicted in linear regression can take a continuous value, it takes categorical (cut) values in the logistic regression model (Elhan 1997).
b) While it is important that independent variables show normal distribution in linear regression, there is no restriction on distribution in logistic regression.
c) The main difference between the two techniques is that while the value of the dependent variable in the multi-linear regression analysis is estimated, in the logistic regression analysis, one of the values that the dependent variable can take is estimating the probability of realization (Elhan 1997).
d) Logistic regression does not produce negative probabilities, all values vary between 0-1.

It is widely used for reasons such as the fact that the results of logistic regression are easy to apply and interpret and it has a package in many programs (SPSS, Python).

When working with the logistic regression model, it is very important that the number of dependent variables has the ability to explain the dependent variable and contains sufficient data. Another important point is that the high correlation between independent variables affects the model negatively. If there is a multicollinearity problem among the variables included in the analysis, it is recommended to remove one or more variables from the model in order to overcome this problem (Tabachnick and Fidell 1996).

### 3.1.3.2 Naive bayes classifier

Naive Bayes classification is a method based on probability theory. The Naive Bayesian method, which is often used among machine learning models, has been shown to yield successful results.

The probability theory of Bayes' theorem is expressed in the formula below.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \quad \textbf{(3,1)}$$

P (A | B) = Probability of A when B is known to be true,
P (B | A) = Probability of B when A is known to be true,
P (A) = probability of A,
P (B) = probability of B occurrence.

Here, a priori probability adds subjectivity to Bayes' theorem. In other words, for example, P (A) is the information about event A before data is collected. On the other hand, P (B | A) is the sequential probability because, after data collection, it provides information about the probability of event B occurring in cases where A has occurred (Theodoridis et al. 2009).

Considering the commonly used Naive Bayes Classifier models;
Gaussian Naive Bayes Classification: This method assumes that the sample is gaussian normally distributed if the data are of continuous value.
Multinominal Naive Bayes Classification: It is the model used when the number of categories is high when classifying categorical data.
Bernoulli Naive Bayes Classification: It is similar to Multinominal Naive Bayes and is a model used in binary classifications.

Main advantages of Naive Bayes algorithm;
a) Ability to give successful results in small data sets,

b) Gives better results since it evaluates each independent variable within itself (it can distinguish data),
c) Considering that it can give successful results with big data and it is easier to apply, it is convenient to use when working with real data or real-time analysis,

Disadvantages;
a) In some cases, it does not take into account the interaction of independent variables with each other.

#### 3.1.3.3 Decision tree

Decision tree algorithm is one of the tree-based algorithms and is widely used in both regression and classification methods.

The first distinction in decision trees is considered as root, and sub-nodes begin to form depending on the fulfilment of the condition here. Each node is reshaped according to the data whose previous condition has been realized. In the use of decision trees in classification, features such as the easy understanding of the created classification model thanks to the simple structure of the decision trees, the fact that the decision trees are not parametric, providing a suitable structure for information discovery and the creation of them in a faster way compared to other classification methods play a role. (Gehrke 2003)

Advantages of decision trees;
a) Easy to apply and the opportunity of obtaining detailed information about the data set when the tree structure is visualized,
b) The preparation time of the data set for analysis being relatively short,
c) Working with categorical and numerical data,
d) Possibility to determine the depth of the study in the analysis made,
e) Possibility to easily check the results on test data or real data which are not included in the analysis with statistical methods,

Disadvantages;

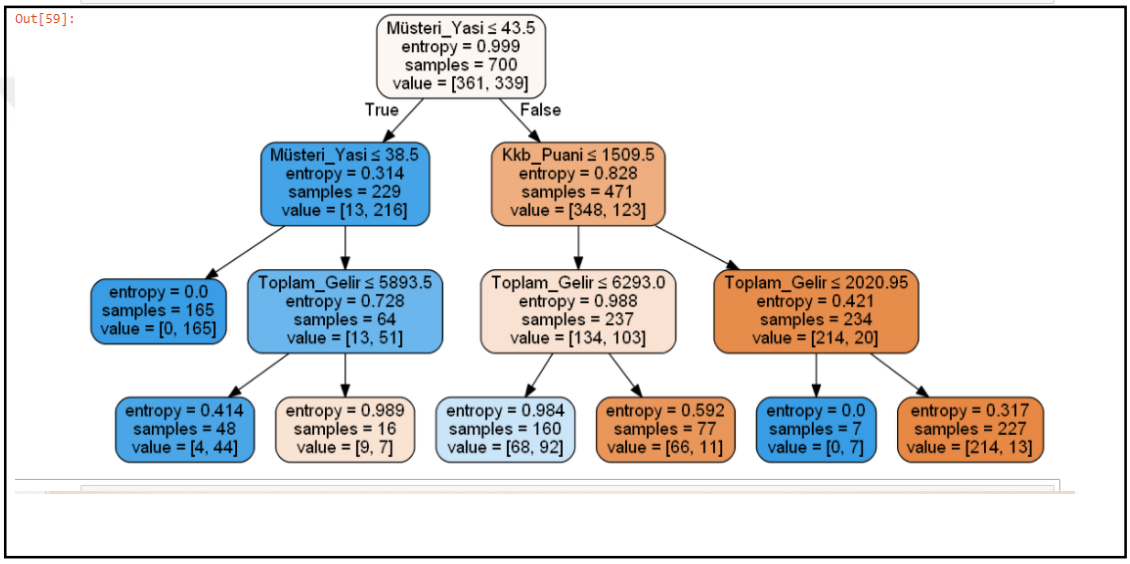
a) Possibility of the results being variable and the necessity of statistical methods to be verified,

b) Ability to produce complex results,

c) Having an overfitting problem.

In addition, the decision tree model can be very useful for the finding systems established within the scope of central controls during audits for the credit process.

**Table 3.4: Sample Decision Tree**




It is the visualization of the tree structure of the decision tree algorithm used in the current study in Table 3.4. The number of branches that the tree can move downward can be determined while setting up the model. While establishing a control mechanism for problem loans, these trees are very useful in finding out which criteria are problematic loans. It is useful to show the effect of different criteria on the dependent variable and the extent to which the loans that meet those conditions become non-performing loans. For example; In Table 3.4, it can be seen that the non-performing loan conversion rates of loans belonging to customers with a customer age above 43.5 and a total income of less than 2020.95 are high. This situation is also beneficial for early warning systems. Periodically running the model with updated data may be effective in determining which factors have how much effect.

### 3.1.3.4 K-Nearest neighbor algorithm

Although the KNN algorithm was first defined in the early 1950s, it did not become popular until the 1960s. Because the computing power at that time was not yet sufficient for a large-scale training dataset required by the k-nearest neighbor algorithm. However, with its increasing computational power, the algorithm is widely used, especially in pattern recognition (Han & Kamber 2001).

The KNN algorithm can be used in both regression and classification studies and it is seen that it is used mainly in classification in practice. The KNN algorithm ensures that the data in the data set is placed in the class it belongs to by matching it with its closest neighbors. In the Nearest Neighbor classification method, the entire training set is evaluated as a classifier, as no classifier model is created before and the raw training data is returned to each new data classification. The KNN algorithm, which is described as a lazy learner in terms of this feature, is an algorithm with a long classification process because it scans each sample one by one. (Khan et al. 2002).

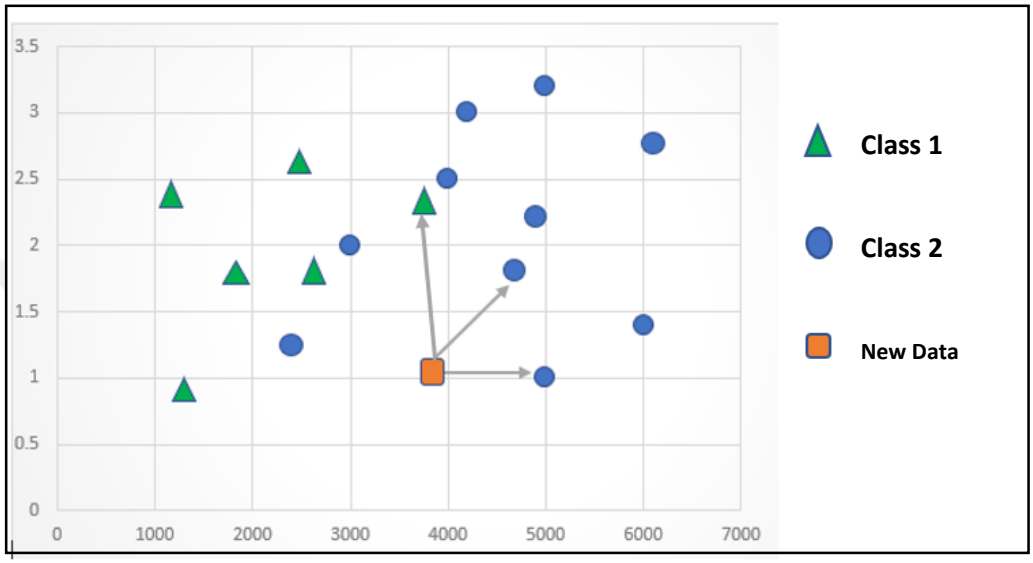
In KNN algorithm, k value represents k number of closest neighbors of a new value. New data are assigned to a class according to the prevalence of values assigned as their closest neighbors.

When using the KNN algorithm, different methods are used for distance measurements of data. Considering the commonly used methods, it is seen that the most widely used method in the literature is the Euclidean distance. Apart from this, methods such as Manhattan, Minkowski and Chebyshev that make distance measurements are also used. Ideally, distance measurement for KNN classification should be adapted to the problem being solved. There are studies showing that more accurate classifications can be made specific to the KNN algorithm by using different distance criteria. (Weinberger et al. 2006).

When classifying with the KNN algorithm, the k value is determined first. Then, the distance of the newly incoming data to the k value from the existing data in the data set

is measured by the above-mentioned Euclidian, Manhattan or Minkowski methods. After the distance measurement is made, the new incoming data is assigned to the class it belongs to and thus the data is labelled.

**Figure 3.1: K-Nearest Neighbor Classification**




As can be seen in Figure 3.1, labelling is made by calculating which class the newly assigned data belongs to according to the distance of the data with other data and classes.

Advantages of KNN algorithm;

a) It is an easy and fast model to apply,

b) Algorithm can produce good results despite noisy data,

Disadvantages;

a) Dependence of algorithm success on k value and attributes.

b) The need for diversity in the nature of classification within the data set.

### 3.1.3.5 Support vector machines

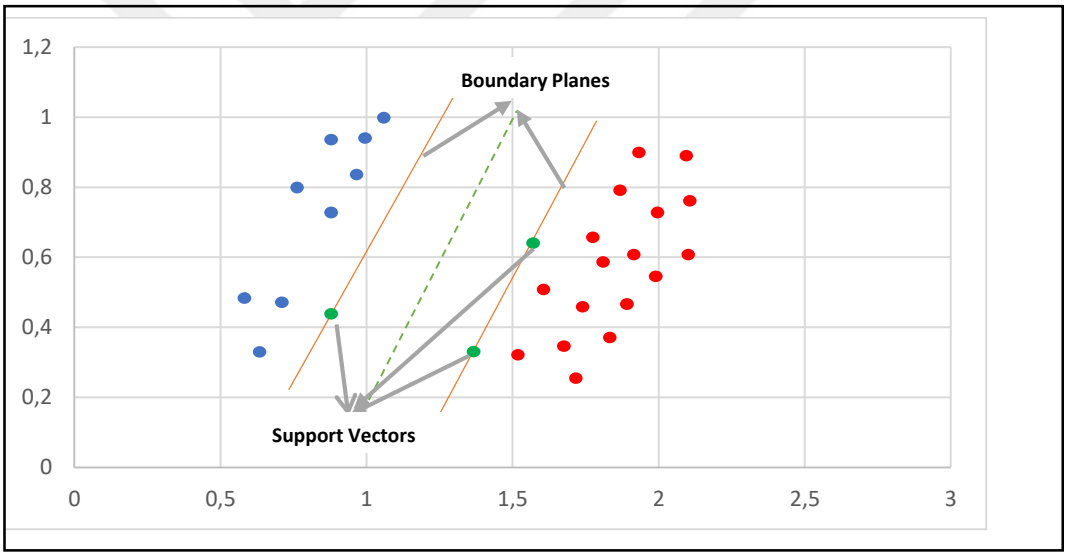
"Support Vector Machines (SVM)", founded in 1963 by Vladimir Vapnik and Alexey Chervonenkis, is a supervised learning algorithm based on the statistical learning theory.

Although its foundations date back to the 60's, it was developed in 1995 by Vladir Vapnik, Berhard Boser and Isabelle Guyon (Akpınar 2014).

Support vector machines are generally used for data classification and the basic point is to group the assigned data correctly. The support vector machine is a model that enables the separation of data from each other in the hyper plane, considering the maximum distance between classes for decomposition.

Support vector machines give successful results not only on linear datasets but also on nonlinear datasets with kernel identification method.

**Figure 3.2: Working System of Support Vector Machines**




As seen in Figure 3.2, support vectors represent the closest data to the hyperplane separating the classes. These support vectors define the boundary of the class to which they belong and are located on a plane parallel to the hyperplane of separation. (Burges 1998). Here it can be seen that the more open between the boundary planes, the better the separation.

It is not always possible for the boundary planes to be linear. In nonlinear data distributions, support vectors cannot draw a linear plane in the form of a straight line. In such cases, parsing can be done with the help of kernel tricks.

There are two commonly used kernel trick methods:
1-Polynomial Kernel: In this method, for the 2-dimensional data set, the 3rd dimension is processed and the data is separated by plane instead of a line.
2- Radial Basis Function (RBF) Kernel: It classifies the gamma value with normal distribution according to the similarities of the points in the data set.

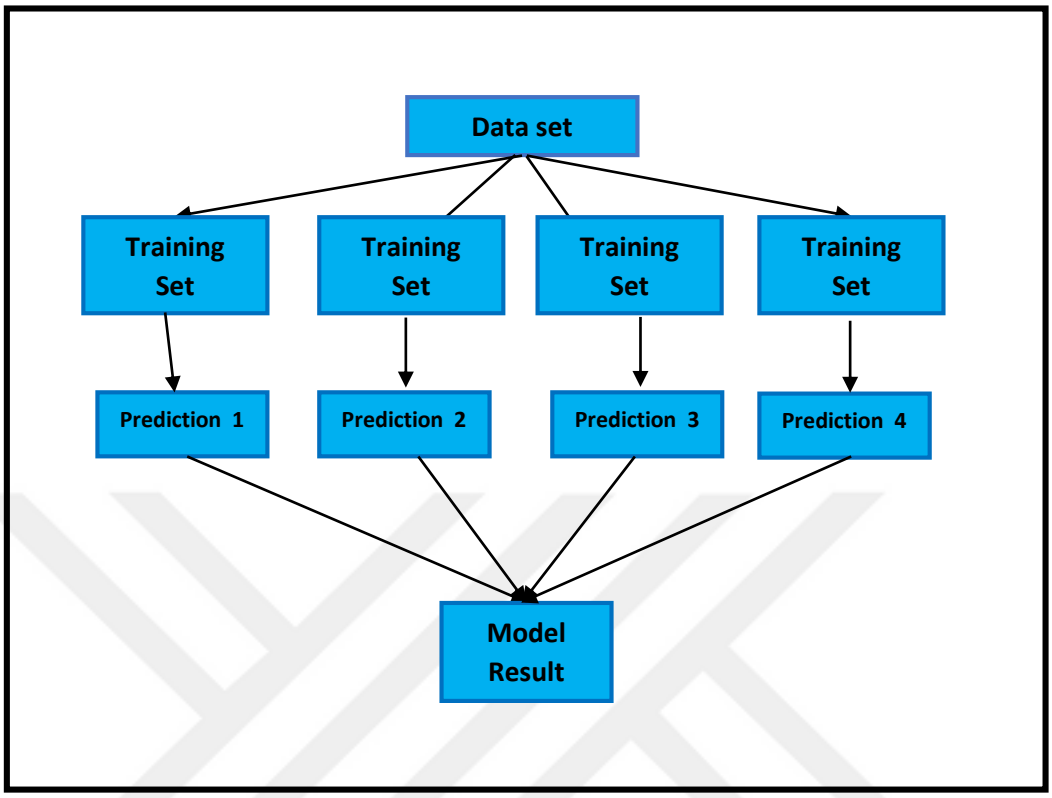
#### 3.1.3.6 Random forest algorithm

The Random Forest algorithm is a method that can generate more accurate estimates and enables more detailed analysis by randomly creating subsets-data sets within the data set. This method, which can be applied to regression and classification problems, turns the results obtained from randomly generated trees into a general result with the "voting" technique.

The biggest problem in models using decision trees is that there is an excessive amount of overfitting when data is scarce (Liao, Ju, & Zou 2016). The random forest method on the one hand increases the predictive power of the model due to the random distribution of the data set, on the other hand, it prevents the occurrence of overfitting problem.

Another advantage of the random forest method is that it can control erroneous predictions that may occur in single tree structures. In the random forest method, the wrong prediction that may occur in one of the training sets can be excluded from the evaluation according to the results of the other training sets. As the number of training sets increases, the possibility of making a biased estimate decreases.

**Figure 3.3: Random Forest Algorithm**




As can be seen in Figure 3.3., Problems such as biased estimation and high correlation that may occur in the data set, especially when analyzing large and complex data sets, can be overcome due to the separation of the data set into subsets in the RF algorithm. Estimates are made as a result of the model created with each training set. The estimations made are reached by the majority of votes method.

## 3.2 MODELLING STUDY FOR COMMERCIAL LOANS PROCESS

In the classification study conducted for the commercial loans process, the probability of companies to default over 43 different financial ratios (shown in Appendix-1) have been analyzed. Within the scope of the study, logistic regression, naive bayes, KNN, decision tree, RF and SVM algorithms were used as in personal loans and the model results of different algorithms were compared.

Since 43 financial ratios are interrelated and there are contradictory data in the data set, a detailed study was carried out in the data set during the data preparation process. First of

all, it was determined that there were empty values in the created data set and these empty values were changed by using statistical methods (assigning median values to empty values). Apart from this, it has been observed that there are outliers in some columns in the data set and these values distort the distribution of the data set and affect the success of the model.

The similarities of the financial ratios with each other were taken into consideration, some of the financial ratios included in the data set were eliminated by using the correlation values and the "Variance Inflation Factor" method in the analysis process. The balance sheets of 316 companies before the date of the last loan were taken as basis. Within the scope of the study, it was desired to work with companies whose risks were recently transferred to non-performing loans and this situation caused the data set to be relatively small.

**Table 3.5: Some Features of Commercial Loans Data Set**

| | mean | std | min | 25 percent | 50 percent | 75 percent | Max |
|---|---|---|---|---|---|---|---|
| Credit Statu | 0.51 | 0.5 | 0 | 0 | 1 | 1 | 1 |
| Current rate | 5.799 | 32.418 | 0.05 | 0.92 | 1293.5 | 2370 | 389640 |
| Acid Test Rate | 3.455 | 28.022 | -76964 | 0.27 | 0.69 | 1131 | 324555 |
| Cash Ratio | 1.689 | 26.730 | -92084 | 0.01 | 0.055 | 0.19 | 322709 |
| Stocks / Current Assets Ratio | 60.082 | 31.250 | 0 | 0.14 | 0.405 | 0.71 | 2274 |
| Stocks / Total Assets Ratio | 37.559 | 237.447 | 0 | 0.08 | 0.26 | 0.51 | 1975 |
| Stock Dependence Rate | 18.983 | 9.764 | -4943 | 0.57 | 1285 | 4194 | 1124408 |
| Short Term Receivables / Current Asset Ratio | 19.463 | 175.657 | 0 | 0.06 | 0.3 | 0.57 | 1930 |
| Short Term Receivables / Total Assets Ratio | 14.412 | 125.856 | 0 | 0.03 | 0.165 | 0.38 | 1222 |
| Foreign Registration Total / Assets (Active) | 124.237 | 418.100 | 0 | 0.53 | 0.79 | 0.9 | 2364 |

As can be seen in Table 3.5, it has been observed that the standard deviations of the outliers in some columns are very high and affect the distribution of the relevant columns. Arrangements were made for outliers in independent variables and the feature selection phase was initiated.

The correlation of 43 financial ratios in the data set was examined and it was observed that there were financial ratios with a high correlation between them and these ratios affect the success of the model. Variance Inflation Factor tests were carried out in order to expand the feature selection with correlation values and the model was established by reducing the data set to 22 independent variables. Although there are many methods for feature selection in the literature, there is no rule about which of these methods give better results and it is understood that it would be appropriate to try different methods and to compare the success rates of the algorithms over the data sets.

After the feature selections, models were created using different algorithms and a model comparison was made. Since the algorithms used in the personal loans process are informed, there is no further information in this section.

## 3.3 CRITERIA USED IN THE EVALUATION OF MODEL RESULTS

6 different algorithms have been studied for the individual loans and commercial loans process, the success and consistency of the models have been checked. In order to evaluate the success of the model, precision, recall, f-score values and cross-validation were used as well as model accuracy scores.

During the model building phase, the data set is divided into two as a train and a test set. The purpose here is to check the model result on the test data after the model has been trained with train data. With this method, which is applied before creating the model, it is useful to test the success of the model on the data not included in the model, to prevent possible erroneous or biased results. In the literature, it is seen that train and test set distinction is used as 70 percent train set-30 percent test set or 80 percent train set-20 percent test set.

After modelling, confusion matrix was used to control the decomposition power and accuracy rates of the established model.

**Table 3.6: Confusion Matrix**

| ACTUAL | PREDICTION | |
|---|---|---|
| | Positive | Negative |
| Positive | True Positive | False Positive |
| Negative | False Negative | True Negative |

Looking at the definitions of the values in the confusion matrix in Table 3.6;
True Positive; Correct estimation of the true value in reality,
True Negative; Wrong estimation of the actually wrong value,
False Positive; Incorrect estimation of the true value,
False Negative; True estimation of the actually wrong value.

The model accuracy rate is obtained by dividing the correctly predicted and actually correct values with the wrong predicted and actually wrong values by all values.

$$\text{Model Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad \textbf{(3.2)}$$

While checking the consistency of the model, besides the accuracy values, precision, recall and f1-score values were checked. The formulas for these values are shown below.

$$\text{Precision} = \frac{TP}{TP+FP} \quad \textbf{(3.3)}$$

Precision value shows the ratio of correctly predicted positive values within all positive values. It is an appropriate criterion to use in situations where incorrect estimation is critical.

$$\text{Recall} = \frac{TP}{TP+FN} \quad \textbf{(3.4)}$$

The recall value shows the ratio of correctly predicted positive values to the sum of correctly predicted positive and incorrectly predicted negative values. It is a criterion used in situations where false negative predictions should not be made.

$$\text{F1-Score} = 2* \frac{precision*recall}{precision+recall} \quad \textbf{(3.5)}$$

F1-Score value is the harmonic mean of precision and recall values. When the precision and recall values are both important, the result may not be accurate when using the simple average. F1-Score is an important indicator for the consistency of the model in cases where the margin of error is low.

Cross-validation is a statistical resampling method used to objectively evaluate the performance of machine learning models.

In the scope of the study, f1-score, recall and precision values were calculated separately for each class and the weighting method was used. The results of the analysis made on personal loans and commercial loans process using different machine learning algorithms are discussed in the next section.

# 4. FINDINGS

In this part, the results of the models created for personal loans and commercial loans processes are shown separately and then the algorithm successes are compared.

## 4.1 COMPARISON OF THE MODEL RESULTS CREATED FOR PERSONAL LOANS

The summary and result table of the logistic regression model created over 1000 customers within the scope of the study are shown below.

**Table 4.1: Personal Loans Logistic Regression Model Results**

| ACTUAL | PREDICTION | |
|---|---|---|
| | Positive | Negative |
| Positive | 92 | 3 |
| Negative | 17 | 88 |
| Accuracy Score : 92+88/92+3+88+17 = 0.9 | | |

As a result of the examination of the model results of the values in Table 4.1, it is understood that 92 out of 95 credits followed in non-performing in 200 data were estimated correctly, 3 of them were estimated incorrectly and 88 of the credits followed in 105 active accounts were classified correctly and 17 were classified incorrectly. Calculation of the accuracy rate of the model based on these results is calculated as follows.

The result in the table is the result by dividing true positive values and true negative values by the sum of all values. It is seen that the accuracy rate of the model created by logistic regression is 90 percent.

The results of the model created with the Gaussian Naive Bayes classification algorithm based on the available data are shown in Table 4.2.

**Table 4.2: Personal Loans Naive Bayes Model Results**

| ACTUAL | PREDICTION | |
|---|---|---|
| | Positive | Negative |
| Positive | 85 | 9 |
| Negative | 16 | 90 |
| Accuracy Score : 85+90/85+9+90+16 = 0.875 | | |

As a result of the analysis of the values model results in Table 4.2, it is understood that 85 credits followed in the non-performing accounts were estimated correctly, 9 of them were wrongly estimated among 200 data in the test set, 90 of the credits followed in the active accounts were classified correctly and 16 were incorrectly classified. In addition, it is seen that the accuracy rate of the model is 87.5 percent.

The results of the model created with the decision tree algorithm are included in the table.

**Table 4.3: Personal Loans Decision Tree Model Results**

| ACTUAL | PREDICTION | |
|---|---|---|
| | Positive | Negative |
| Positive | 74 | 20 |
| Negative | 14 | 92 |
| Accuracy Score : 74+92/74+20+92+14 = 0.83 | | |

As a result of the examination of the values model results in Table 4.3, it is understood that among 200 data in the test set, 74 credits followed in the non-performing accounts were estimated correctly, 20 of them were guessed incorrectly, 92 of the credits followed in the active accounts were classified correctly and 14 were incorrectly classified. In addition, it is seen that the accuracy rate of the model is 83 percent.

The results of the model created with the KNN algorithm are included in the table 4.4.

**Table 4.4: Personal Loans K-Nearest Neighbor Model Results**

| ACTUAL | PREDICTION | |
|---|---|---|
| | Positive | Negative |
| Positive | 97 | 10 |
| Negative | 29 | 64 |
| Accuracy Score : 97+64/97+10+64+29 = 0.805 | | |

As a result of the examination of the values model results in Table 4.4, it is understood that among the 200 data in the test set, 97 credits followed in the non-performing accounts were estimated correctly, 10 of them were guessed incorrectly, 64 of the credits followed in the active accounts were classified correctly and 29 were classified incorrectly. In addition, it is seen that the accuracy rate of the model is 80.5 percent.

The results of the model created with the SVM algorithm are included in the table 4.5.

**Table 4.5: Personal Loans Support Vector Machines Model Results**

| ACTUAL | PREDICTION | |
|---|---|---|
| | Positive | Negative |
| Positive | 78 | 16 |
| Negative | 21 | 85 |
| Accuracy Score : 78+85/78+16+85+21 = 0.815 | | |

As a result of the examination of the values model results in Table 4.5, it is understood that 78 credits followed in the non-performing accounts were estimated correctly, 16 of them were guessed incorrectly among 200 data in the test set, 85 of the credits followed in active accounts were classified correctly and 21 were classified incorrectly. In addition, it is seen that the accuracy rate of the model is 81.5 percent.

The results of the model created with the RF algorithm are included in the table 4.6.

**Table 4.6: Personal Loans Random Forest Model Results**

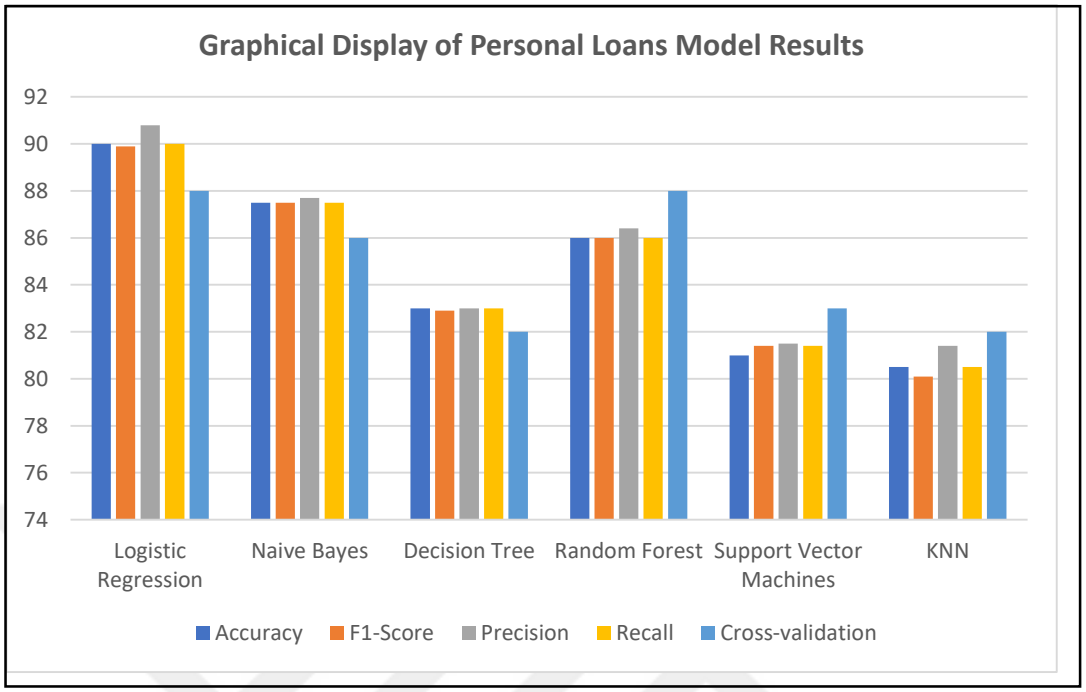
| ACTUAL | PREDICTION | |
|---|---|---|
| | Positive | Negative |
| Positive | 85 | 9 |
| Negative | 19 | 87 |
| Accuracy Score : 85+87/85+9+87+19 = 0.86 | | |

As a result of the examination of the values model results in Table 4.6, it is understood that 85 credits followed in the non-performing accounts were predicted correctly, 9 of them were wrongly estimated among 200 data in the test set, 87 of the credits followed in the active accounts were classified correctly and 19 were classified incorrectly. In addition, it is seen that the accuracy rate of the model is 86 percent.

**Table 4.7: Comparison of Model Results for Personal Loans**

| **Comparison of Model Results for Personal Loans** | | | | | |
|---|---|---|---|---|---|
| | **Accuracy** | **F1-Score** | **Precision** | **Recall** | **Cross-Validation** |
| **Logistic Regression** | 90 | 89.9 | 90.8 | 90 | 88 |
| **Naive Bayes** | 87.5 | 87.5 | 87.7 | 87.5 | 86 |
| **Decision Tree** | 83 | 82.9 | 83 | 83 | 82 |
| **Random Forest** | 86 | 86 | 86.4 | 86 | 88 |
| **Support Vector Machines** | 81 | 81.4 | 81.5 | 81.4 | 83 |
| **KNN** | 80.5 | 80.1 | 81.4 | 80.5 | 82 |

**Figure 4.1: Graphical Display of Personal Loans Model Results**




As can be seen in Table 4.7 and Figure 4.1, although it is seen that logistic regression, Naive Bayes classifier and RF algorithms produce more successful results when the accuracy rate is taken into consideration compared to other algorithms, it has been observed that all models produce close results in general. In addition, it is understood that the evaluations of the model results made with the f1-score, precision, recall and cross-validation methods are also consistent.

Regarding the work done on different algorithms for personal loans processes; it is concluded that the features included in the data set are useful in calculating the probability of default for customers' loans and when classifying them. Information on how to use the results of this study, which is made by comparing different algorithms, in internal control processes will be discussed in the last part of the study.

## 4.2 COMPARISON OF MODEL RESULTS CREATED FOR COMMERCIAL LOANS

Since information about the models is given in the study on personal loan processes, only the analysis results will be included in this section.

The results of the model created with the Logistic Regression algorithm are included in the table 4.8.

**Table 4.8: Commercial Loans Logistic Regression Model Results**

| ACTUAL | PREDICTION | |
|---|---|---|
| | Positive | Negative |
| Positive | 24 | 9 |
| Negative | 5 | 26 |
| Accuracy Score : 24+26/24+9+26+5 = 0.78 | | |

As a result of the examination of the values model results in Table 4.8, it is understood that among the 64 data in the test set, 24 credits followed in the non-performing accounts were predicted correctly, 9 of them were wrongly predicted, 26 of the credits followed in active accounts were classified correctly and 5 were incorrectly classified. In addition, it is seen that the accuracy rate of the model is 78 percent.

The results of the model created with the Naïve Bayes algorithm are included in the table 4.9.

**Table 4.9: Commercial Loans Naive Bayes Model Results**

| ACTUAL | PREDICTION | |
|---|---|---|
| | Positive | Negative |
| Positive | 24 | 9 |
| Negative | 8 | 23 |
| Accuracy Score : 24+23/24+9+23+8 = 0.73 | | |

As a result of the examination of the values model results in Table 4.9, it is understood that among the 64 data included in the test set, 24 credits followed in the non-performing accounts were predicted correctly, 9 of them were wrongly predicted, 23 of the credits

followed in active accounts were classified correctly, and 8 of them were classified incorrectly. In addition, it is seen that the accuracy rate of the model is 73 percent.
The results of the model created with the Decision Tree algorithm are included in the table 4.10.

**Table 4.10: Commercial Loans Decision Tree Model Results**

| ACTUAL | PREDICTION | |
|---|---|---|
| | Positive | Negative |
| Positive | 24 | 10 |
| Negative | 7 | 23 |
| Accuracy Score : 24+23/24+10+23+7 = 0.73 | | |

As a result of the examination of the values model results in Table 4.10, it is understood that among the 64 data in the test set, 24 credits followed in the non-performing accounts were predicted correctly, 10 of them were incorrectly, 23 of the credits followed in active accounts were classified correctly and 7 were incorrectly classified. In addition, it is seen that the accuracy rate of the model is 73 percent.

The results of the model created with the KNN algorithm are included in the table 4.11.

**Table 4.11: Commercial Loans KNN Model Results**

| ACTUAL | PREDICTION | |
|---|---|---|
| | Positive | Negative |
| Positive | 31 | 2 |
| Negative | 10 | 21 |
| Accuracy Score : 31+21/31+2+21+10 = 0.81 | | |

As a result of the examination of the values model results in Table 4.11, it is understood that 31 credits followed in the non-performing accounts were predicted correctly, 2 of them were guessed incorrectly among 64 data included in the test set, 21 of the credits

followed in active accounts were classified correctly and 10 of them were classified incorrectly. In addition, it is seen that the accuracy rate of the model is 81 percent.

The results of the model created with the SVM algorithm are included in the table 4.12.

**Table 4.12: Commercial Loans SVM Model Results**

| ACTUAL | PREDICTION | |
|---|---|---|
| | Positive | Negative |
| Positive | 28 | 6 |
| Negative | 0 | 30 |
| Accuracy Score : 28+30/28+6+30 = 0.90 | | |

As a result of the examination of the values model results in Table 4.12, it is understood that 28 credits followed in the non-performing accounts were estimated correctly, 6 of them were guessed incorrectly, and 30 of the credits followed in the active accounts were classified correctly among the 64 data included in the test set. In addition, it is seen that the accuracy rate of the model is 90 percent.

The results of the model created with the RF algorithm are included in the table 4.13.

**Table 4.13: Commercial Loans Random Forest Model Results**

| ACTUAL | PREDICTION | |
|---|---|---|
| | Positive | Negative |
| Positive | 31 | 3 |
| Negative | 5 | 25 |
| Accuracy Score : 31+25/31+3+25+5 = 0.875 | | |

As a result of the examination of the values model results in Table 4.13, it is understood that 31 credits followed in the non-performing accounts were estimated correctly, 3 of them were wrongly estimated among 64 data in the test set, 25 of the credits followed in

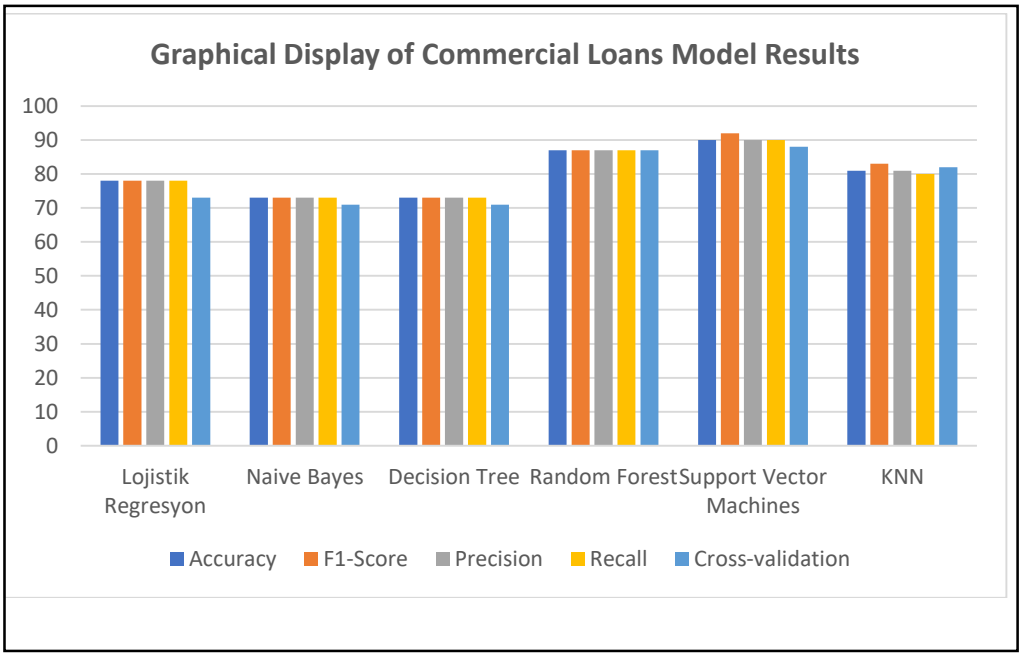
active accounts were classified correctly and 5 were classified incorrectly. In addition, it is seen that the accuracy rate of the model is 87 percent.

The result of the analysis made on the commercial loans process using different machine learning algorithms is shown in the table below.

**Table 4.14: Comparison of Model Results Created for Commercial Loans**

| Comparison of Model Results for Commercial Loans | | | | | |
|---|---|---|---|---|---|
| | **Accuracy** | **F1-Score** | **Precision** | **Recall** | **Cross-Validation** |
| **Logistic Regression** | 78 | 78 | 78 | 78 | 73 |
| **Naive Bayes** | 73 | 73 | 73 | 73 | 71 |
| **Decision Tree** | 73 | 73 | 73 | 73 | 71 |
| **Random Forest** | 87 | 87 | 87 | 87 | 87 |
| **Support Vector Machines** | 90 | 92 | 90 | 90 | 88 |
| **KNN** | 81 | 83 | 81 | 80 | 82 |

**Figure 4.2: Graphical Display of Commercial Loans Model Results**



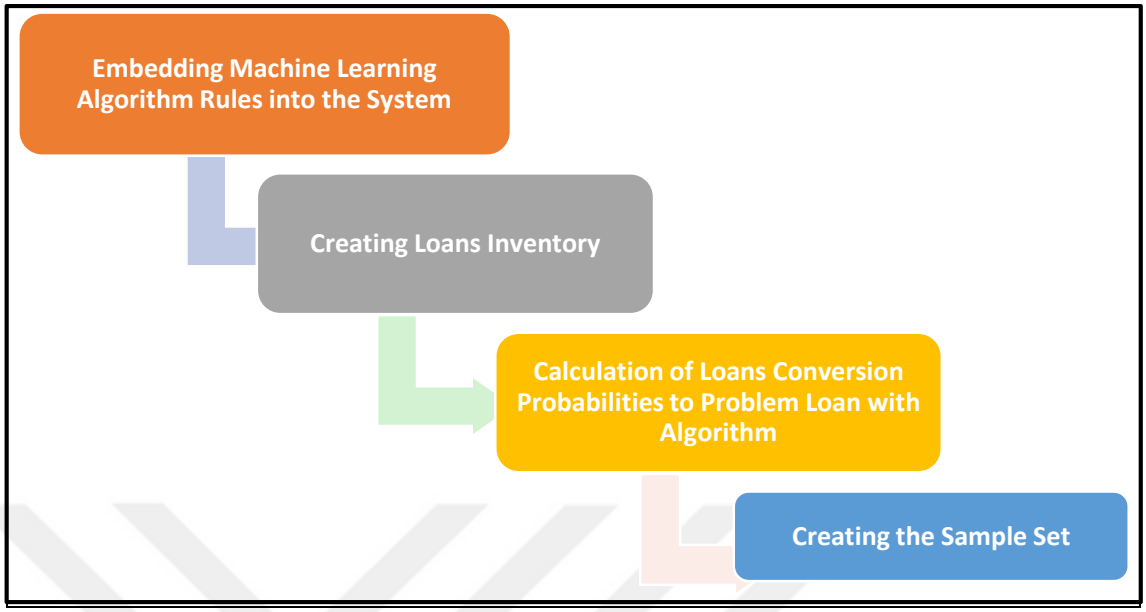
As can be seen in Table 4.14 and Figure 4.2, SVM algorithm gave more successful results than other algorithms with 90 percent accuracy. In addition, it has been observed that Random Forest and KNN algorithms have similar model results and perform better decomposition than other algorithms. In addition, it is understood that the evaluations of the model results made with the f1-score, precision, recall and cross-validation methods are also consistent.

## 4.3 USE OF MACHINE LEARNING ALGORITHMS IN INTERNAL AUDIT PROCESSES

Analysis made for commercial loans and personal loans processes using different algorithms in Chapter 1 and Chapter 2 reveal high accuracy models at the point of turning the loans of companies or customers into non-performing loans.

In the audits performed by the internal audit units, the methods of selecting samples from the large data sets are applied for the audited process (example: credits process, accounting process) and these methods are created by using statistical or non-statistical methods. With the increase in transaction volume due to technological developments, the ability of the inventories created to be controlled to represent the population and to be risk-oriented has become even more important. At this point, it is seen that taking advantage of artificial intelligence technologies and machine learning models will bring serious advantages. The above studies on the commercial loans and personal loans process can distinguish the probability of default of loans to a certain extent. Such models show which factors / characteristics increase the probability of default and it is seen that if such models are used while creating control inventories, they will significantly contribute to the detection of risky transactions.

**Figure 4.3: Creating a Sample Set with Machine Learning Algorithm**




The work flow diagram shown simply in Figure 4.3 proceeds similar to the process of evaluating and monitoring the model described in the data mining section. The model established with machine learning algorithms should be checked with real data and the analysis result should be checked to give consistent results when applied to real data. After the model is put into practice, the model should be retested periodically and the results should be compared.

# 5. DISCUSSION AND CONCLUSION

Within the scope of the study, logistic regression, naive bayes, KNN, decision tree, RF and SVM algorithms were used for personal loans and commercial loans processes. The model results reveal that the existing data sets and selected variables are successful in creating models with high discrimination power.

When the results of the study are evaluated together, it is seen that the models established with these algorithms can be useful in calculating the probability of default of individual customers and loans granted to companies. In addition, while the logistic regression model gave the best result in the study on the personal loans process, the SVM algorithm gave the best result in the modelling study for the commercial loans process. Although the results of the algorithms are generally close to each other in the studies conducted in the literature, it is thought that the results may vary according to the characteristics of the data set and it would be useful to work on different algorithms.

Although this study is explained over the models created for the loan process, it is thought that the integration of machine learning algorithms into different processes of internal audit can provide serious benefits. For example; when it is desired to inspect the branches of a bank periodically, it will be useful to use machine learning algorithms while classifying the risk of branches. In addition, machine learning algorithms are used in many areas such as anomaly detection in accounting audit processes, detection of risky transactions in deposit audit processes.

Artificial intelligence technologies are becoming widespread in different sectors and are now becoming a part of business processes. In this process of change, it is seen that banks benefit from artificial intelligence technologies in many processes such as evaluation and analysis. While it is very important for the internal audit units in banks to understand artificial intelligence technologies for the evaluation of these processes within the institution, it is thought that using these technologies will make a significant contribution while performing audits.

As a result; for the bank audit processes that are handled specifically in the thesis, the opportunities and risks created by artificial intelligence technologies for the institution should be taken into consideration by the audit units and actions should be taken against the risks that the institution may face. In addition, the issue of integrating artificial intelligence technologies in the audit processes should be considered and benefiting from these technologies should be considered.

# REFERENCES

***Books***


Akpınar, H. (2014), Data Veri Madenciliği Veri Analizi, Papatya Yayıncılık

Hurwitz, J. & Kirsch, D. (2018) Machine Learning for Dummies, IBM Limited Edition, s.6.

Linoff, G.S. & Berry, M.J.A., 2011. "Data Mining Techniques for Marketing, Sales and Customer Relationship Management", *Wiley*,3rd.ed., Canada.

Öztemel E. (2003), "Yapay Sinir Ağları", Papatya Yayıncılık.

***Periodicals***


Alina, C.M., Cerasela, S.E., Gabriela, G., et al. (2018) "Internal audit role in artificial intelligence", *Ovidius University Annals, Economic Sciences Series* **18**(1), ss.441–445.

Baynal, K. & Çalış, A. (2016), "Veri Madenciliğinde Kümeleme Analizi ile Bankacılık Sektöründe Bir Uygulama", *Beykent Üniversitesi Fen ve Mühendislik Bilimleri Dergisi*, **9**(1).

Bosij, P. & Chafey, D. & Greasley, A. ve Hickie. S. (2003). *Business Information Systems: Technology, Development and Management for the E-Business. 2. Basım.* S. 4-5. Financial Times-Prentice Hall.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition, data mining and knowledge discovery. Kluwer Academic Publishers, **2** (2), ss. 121-167.

Dumouchel, W.,1999. Bayesian Data Mining in Large Frequency Tables, With An Application To The FDA Spontaneous, American Statistician, **53**(3),190-196.

Khan, M., Ding, Q., Perrizo, W., (2002), "K-nearest Neighbor Classification on Spatial Data Streams Using P-trees", Advances in Knowledge Discovery and Data Mining, ss. 517-528.

Kırlıoğlu, H. & Ceyhan, İ.F. (2014), "Mali Tablo Denetiminde Ön Analitik İnceleme Tekniği Olarak Veri Madenciliğinin Kullanımı: Borsa İstanbul Uygulaması", *Akademik Yaklaşımlar Dergisi*, **5**(1).

Krunoslav, R. & Stankovic, Z. & Avramović, Z. (2020). Implications of Implementation of Artificial Intelligence in the Banking Business with Correlation to the Human Factor. *Journal of Computer and Communications*. **8**(11), ss. 130-144. 10.4236/jcc.2020.811010.

Miyan, M. (2017), "Applications of Data Mining in Banking Sector", *International Journal of Advanced Research in Computer Science*, **8**(1), p108-114. 7p.

Önder, H., Cebeci, Z., 2002. Lojistik Regresyonlarda Değişken Seçimi. Çukurova Üniv. Ziraat Fakültesi Dergisi, **17**(2),105–114.

Özdemir, İ , Sağıroğlu, Ş . (2018), "Denetimlerde Büyük Veri Kullanımı Ve Üzerine Bir Değerlendirme". Gazi University Journal of Science Part C: Design and Technology **6**(2), ss. 470-480.

Qasem, M. H. & Nemer, L. (2018), Extreme Learning Machine for Credit Risk Analysis", *Journal of Intelligent Systems,* **29**(1).

Restivo, K., 1999. The Drill on Data Mining, Computer Dealer News, **15** (14), 29-30.

Senal, S. & Öztürk, M.S. (2014), "Denetim Planlaması Aşamasında Analitik İnceleme Tekniği Olarak Yapay Sinir Ağları Kullanımı: Bir Firma Uygulaması", *Ege Akademik Bakış,* **14**(3).

Shetty, N. & Shetty, J. & Narula, R. & Tandona, K. (2020). "Comparison study of machine learning classifiers to detect anomalies" , *International Journal of Electrical and Computer Engineering (IJECE),* **10**(5).

Silva, A.P.D. (2017). "Optimization approaches to Supervised Classification*", European Journal of Operational Research*, **261** (2), 772-788.

Türkiye İç Denetim Enstitüsü, 2017, Tone at the Top, Aralık, İstanbul

Türkiye İç Denetim Enstitüsü, Özel Sayı Part 1: Küresel Bakış Açıları ve Anlayışlar: "Yapay Zeka- İç Denetim Mesleğine İlişkin Dikkate Alınması Gerekenler", 2017, https://www.tide.org.tr/file/documents/pdf/GPAI-Artificial-Intelligence-Part-I-Revised.pdf , [erişim tarihi 10 Mart 2021], s. 9.

***Other Publications***


Dickey, G. & Blanke, S. And Seaton, L. (2019), “Machine Learning in Auditing”, https://www.cpajournal.com/2019/06/19/machine-learning-in-auditing/, [accessed 02 December 2020].

Elhan, A.H. (1997), Lojistik Regresyon Analizinin İncelenmesi ve Tıpta Bir Uygulaması. (Biyoistatistik Yüksek Lisans Tezi) Ankara Üniversitesi, 4-29, Ankara.

Han, J. and Kamber, M., (2001), “Data Mining Concepts and Techniques”, Academic Press, New York.

J. Gehrke, (2003), “Decision Trees”, The Handbook of Data Mining, Editör: Nong Ye, Lawrence Erlbaum Associates Publishers, London, 149-175.

Liao, Z., Ju, Y., and Zou, Q. (2016). Prediction of G protein-coupled receptors with SVM-Prot features and random forest. Scientifica, 8309253. doi: 10.1155/2016/8309253.

N. K. Gyamfi and J. Abdulai, , (2018), "Bank Fraud Detection Using Support Vector Machine," *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pp. 37-41, doi: 10.1109/IEMCON.2018.8614994.

R. E. Turkson, E. Y. Baagyere and G. E. Wenya, (2016), "A machine learning approach for predicting bank credit worthiness," *2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR)*, 2016, pp. 1-7, doi: 10.1109/ICAIPR.2016.7585216.

S. Theodoridis and K. Koutroumbas, (2009), Pattern Recognition, 4. Edition. Burlington: Academic Press.

Skirka, A. & Adamyk, B. & Adamyk, O. and Valytska, M. (2020), "Trust in the European Central Bank: Using Data Science and predictive Machine Learning Algorithms", *2020 10th International Conference on Advanced Computer Information Technologies (ACIT)*, pp. 356-361, doi: 10.1109/ACIT49673.2020.9208857.

Tabachnick, B. G., Fidell, L. S. (1996). Using multivariate statistics (3rd ed.). New York, USA: HarperCollins College Publishers

Weinberger, K.Q., Blitzer, J., Saul, L.K., (2006), Distance Metric Learning for Large Margin Nearest Neighbor Classification, 20.11.2020, https://papers.nips.cc/paper/2005/file/a7f592cef8b130a6967a90617db5681b-Pape

# APPENDIX

.