



**Sosyal Bilimler
Enstitüsü**

T.C.

MARMARA ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
EKONOMETRİ ANABİLİM DALI
İSTATİSTİK BİLİM DALI

**COMPARISON AND ANALYSIS OF LOGISTIC REGRESSION, NEURAL NETWORKS
AND NAİVE BAYES MACHINE LEARNING ALGORITHMS FOR PREDICTING
CHILD LABOR IN TURKEY**

M. A. Thesis

MEHMET ROHAT BİNGÖL

İSTANBUL, 2023

T.C.
MARMARA UNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
EKONOMETRİ ANABİLİM DALI
İSTATİSTİK BİLİM DALI

**COMPARISON AND ANALYSIS OF LOGISTIC REGRESSION, NEURAL NETWORKS
AND NAİVE BAYES MACHINE LEARNING ALGORITHMS FOR PREDICTING
CHILD LABOR IN TURKEY**

M. A. Thesis

MEHMET ROHAT BİNGÖL

Advisor: Prof. Ahmet Mete Çilingirtürk

İSTANBUL, 2023

ABSTRACT

COMPARISON AND ANALYSIS OF LOGISTIC REGRESSION, NEURAL NETWORKS AND NAÏVE BAYES MACHINE LEARNING ALGORITHMS FOR PREDICTING CHILD LABOR IN TURKEY

This thesis aims to compare and analyze the performance of three popular machine learning algorithms - logistic regression, Naïve Bayes, and random forest - in predicting child labor in Turkey. The statistics on working children status in Turkey in 2019 were derived from the micro database of the Turkish Statistical Institute (TURKSTAT). The study leverages big data analysis, artificial intelligence, and classification algorithms to gain insights into the complex dynamics of child labor and inform effective interventions. The research explores the intersection of these fields, emphasizing their significance in addressing societal challenges. Using a comprehensive dataset on child labor in Turkey, we apply supervised learning techniques to develop and evaluate predictive models. The findings reveal that while logistic regression and Naïve Bayes offer valuable insights, the random forest algorithm demonstrates superior performance in accurately predicting child labor instances. The model achieves an impressive accuracy rate of 97%, highlighting its ability to capture intricate feature interactions. The Study adds to the body of knowledge by demonstrating the effectiveness of machine learning algorithms in preventing child labor and by offering evidence-based advice to decision-makers. But it's important to identify some restrictions, such the generalizability and quality of the data. To improve prediction accuracy and broaden our understanding of the dynamics of child labor, future study should investigate more sophisticated algorithms and incorporate extra information.

Key Words: Machine Learning Classification Algorithms, Logistic Regression, Naïve Bayes, Random Forest, Child Labor

ÖZET

LOJİSTİK REGRESYON, SİNİR AĞLARI VE NAİVE BAYES MAKİNE ÖĞRENİMİ ALGORİTMALARININ KARŞILAŞTIRILMASI VE ANALİZİ: TÜRKİYE'DE ÇOCUK İŞÇİLİĞİNİN ÖNGÖRÜLMESİ ÖRNEĞİ

Bu tez çalışması, sosyal bilimlerde büyük bir etki yaratan makine öğrenme algoritmalarının önemine odaklanmaktadır. Makine öğrenme, büyük veri analizi, yapay zeka ve sınıflandırma gibi alanlarda önemli bir yer edinmiştir. Sosyal bilimler, karmaşık toplumsal sorunları anlamak ve çözümlmek için veri tabanlı yaklaşımlara dayalı güçlü bir araç sağlamaktadır. Makine öğrenme algoritmaları, büyük miktarda veriye dayalı örüntüleri tanımlama, tahmin yapma ve sınıflandırma gibi görevlerde büyük bir potansiyele sahiptir.

Özellikle sınıflandırma algoritmaları, sosyal bilimlerdeki birçok araştırma alanında önemli bir rol oynamaktadır. Lojistik regresyon, Naïve Bayes ve rastgele orman gibi makine öğrenme sınıflandırma algoritmaları, karmaşık veri setlerinde çeşitli kategorilere ayrılmış verileri etkili bir şekilde sınıflandırma yetenekleri ile bilinir. Bu algoritmalar, çocuk işçiliği gibi önemli sosyal sorunların analizinde ve tahmininde kullanıldığında, politika yapıcılarına ve araştırmacılara değerli bilgiler sunabilir.

Araştırmada kullanılan veri seti, Türkiye İstatistik Kurumu (TÜİK) tarafından 2019 gerçekleştirilen çocuk işgücü anketinin çıktısı olan çalışan çocuk istatistikleri mikro veri setinden alınmıştır. Türkiye genelinde çocuk işçiliği vakalarının kapsamlı bir görüntüsünü sunmaktadır. Veri seti çocuk işçiliği ile ilişkilendirilebilecek çeşitli özellikler içermektedir, örneğin yaş, cinsiyet, eğitim düzeyi ve hane halkı büyüklüğü gibi.

Denetimli öğrenme teknikleri kullanılarak, lojistik regresyon, Naïve Bayes ve rastgele orman algoritmaları üzerinde tahmin modelleri oluşturulmuştur. Bu modeller, çocuk işçiliği vakalarını tahmin etmek için kullanılan veri setini kullanarak eğitilmiş (%80) ve test (%20) edilmiştir.

Elde edilen sonuçlar, rastgele orman algoritmasının diğer iki algoritmadan daha yüksek bir tahmin doğruluğuna sahip olduğunu göstermektedir. Lojistik regresyon %92 doğruluk oranı sağlarken, Naïve Bayes %86 doğruluk oranı sunmuş ve rastgele orman %97 doğruluk oranına ulaşmıştır. Bu bulgular, rastgele orman algoritmasının çocuk işçiliği vakalarını

tahmin etmede daha etkili olduğunu ve daha güvenilir sonuçlar üretebildiğini göstermektedir.

Ayrıca, algoritmaların performansını değerlendirmek için farklı metrikler de kullanılmıştır. Hassasiyet, özgüllük, doğruluk ve F1 puanı gibi metrikler, her bir algoritmanın çocuk işçiliği vakalarını doğru bir şekilde sınıflandırma yeteneklerini değerlendirmek için kullanılmıştır. Bu analizler, rastgele orman algoritmasının en yüksek hassasiyet, özgüllük ve F1 puanı değerlerine sahip olduğunu göstermiştir.

Ancak, bu çalışmanın bazı sınırlamaları da göz önünde bulundurulmalıdır. Veri setinde eksik veya eksik olabilecek bazı önemli bilgiler olabilir, örneğin bölge bazlı aile geliri gibi. Bu eksik bilgilerin sonuçları etkileyebileceği ve daha tam bir resim sunabilmek için daha kapsamlı veri toplama çabalarının gerekliliği akılda tutulmalıdır.

Sonuç olarak, bu tez çalışması, Türkiye'deki çocuk işçiliğini tahmin etmek için lojistik regresyon, Naïve Bayes ve rastgele orman algoritmalarının performansını karşılaştırarak önemli bir katkı sağlamaktadır. Rastgele orman algoritmasının daha yüksek bir doğruluk oranı ve daha iyi sınıflandırma yetenekleri sergilediği bulunmuştur. Bu çalışma, çocuk işçiliğiyle mücadele etmek için veri odaklı ve etkili stratejilerin geliştirilmesinde bilimsel bir temel sunmaktadır. Gelecekteki çalışmalar, daha kapsamlı veri setleri kullanarak ve diğer gelişmiş makine öğrenme algoritmalarını dahil ederek bu alanda daha ileri araştırmalara yol açabilir.

Anahtar Kelimeler: Makine Öğrenmesi Sınıflandırma Algoritmaları, Lojistik Regresyon, Naïve Bayes, Rastgele Orman, Çocuk İşçiliği

FOREWORD

I am deeply grateful to my dear girlfriend, Begüm Tulgaroğlu, whose unwavering support and love have been my constant motivation. I also extend my appreciation to my friends and family for their encouragement and understanding throughout this endeavor.

Furthermore, I would like to thank my thesis supervisor, Prof. Dr. Ahmet Mete Çilingirtürk, for his invaluable guidance and support.

Mehmet Rohat BİNGÖL

İstanbul, 2023

TABLE OF CONTENTS

TEZ ONAY BELGESİ.....	
ABSTRACT.....	i
ÖZET	ii
FOREWORD	iv
LIST OF ABBREVIATIONS.....	vii
LIST OF TABLES	viii
LIST OF FIGURES	ix
CHAPTER 1 - INTRODUCTION.....	1
CHAPTER 2 - LITERATURE REVIEW	3
2.1. Big Data Analysis and Data Mining	3
2.2. Artificial Intelligence	5
2.3. Machine Learning.....	7
2.3.1. Supervised Learning	12
2.3.2. Unsupervised Learning	16
2.3.3. Semi-supervised Learning	18
2.4. Classification Algorithms.....	19
2.4.1 Logistic Regression.....	23
2.4.2 Naïve Bayes	29
2.4.3 Random Forest	33
2.5. Evaluation Performance Measures.....	37
2.5.1. Confusion Matrix	38
2.5.2. Accuracy	39
2.5.3. Precision.....	39
2.5.4. Recall	39

2.5.5. F1 Score	39
2.5.6. AUC-ROC Curve	40
2.5.7. A Comparative Analysis of Performance Measures: A Review of Literature	40
CHAPTER 3 – RESULTS	43
3.1. Purpose of Application.....	43
3.2. Child Labor in Turkey and Literature Review	43
3.3. Dataset Definition	45
3.4. Data Analysis	47
3.4.1 Descriptive Statistics.....	47
3.4.2 Examination of Child Labor in Turkey with Logistic Regression Model.....	51
3.4.3 Examination of Child Labor in Turkey with Naïve Bayes Model	55
3.4.4 Examination of Child Labor in Turkey with Random Forest Model	57
3.4.5 Performance Comparisons of Logistic Regression Model, Naïve Bayes Model and Random Forest Model.....	60
CHAPTER 4 - CONCLUSION	63
REFERENCES	66

LIST OF ABBREVIATIONS

CNN	Convolutional Neural Network
ILO	International Labor Organization
LRA	Logistic Regression Analysis
ML	Machine Learning
NB	Naïve Bayes
SVMs	Support Vector Machines
RF	Random Forest
TURKSTAT	Turkish Statistical Institute
THLFS	Turkish Household Labor Force Survey
US	United States
WCS	Working Children Statistics
WEF	World Economic Forum

LIST OF TABLES

Table 1. Breiman's Random Forest Predicted Value at X	35
Table 2. Confusion Matrix Template for a Two-Class Problem	38
Table 3. The Model with Measurement Items	45
Table 4. Mann-Whitney U test p-values for FACOM variables	48
Table 5. Frequencies, Percentages and Chi-Square p-values For Independent Variables.	49
Table 6. Logistic Regression Model Statistics.....	51
Table 7. Logistic Reg. Model (80% Trainig Data): Wald Chi-Square p-values and OR..	52
Table 8. Conflict Matrix for Logistic Regression Model	54
Table 9. Performance Evaluation for Logistic Regression Model.....	55
Table 10. Naïve Bayes Model Statistics	55
Table 11. Conflict Matrix for Naive Bayes model.....	56
Table 12. Performance Evaluation for the Naive Bayes model.....	56
Table 13. Random Forest Model Statistics.....	58
Table 14. Conflict Matrix for the Random Forest Model.....	59
Table 15. Performance Evaluation for the Random Forest Model	58
Table 16. Performance Comparison of Machine Learning Classification Algorithms.....	60

LIST OF FIGURES

Figure 1. Data Mining	5
Figure 2. Sub-Dimensions of Artificial Intelligence	7
Figure 3. Machine Learning Strategies	9
Figure 4. Machine Learning Algorithms	11
Figure 5. Supervised learning process	13
Figure 6. Supervised Learning Model.....	14
Figure 7. Illustrations of the semi-supervised learning assumptions	18
Figure 8. Classification Model for Loan Application Data	20
Figure 9. Classification Model for Loan Application Data	21
Figure 10. Naive Bayes Structure	31
Figure 11. Cramer's V correlation Heat Map.....	48
Figure 12. Random Forest Feature Importance.....	59
Figure 13. Accuracy Comparison of Machine Learning Classification Algorithms	61
Figure 14. ROC Curve of Machine Learning Classification Algorithms.....	61

CHAPTER 1 - INTRODUCTION

Machine learning (ML) classification algorithms play a crucial role in social science research by providing crucial insights into societal phenomena such as social patterns and societal phenomena due to their ability to analyze big and complicated datasets. As a result, today, a lot of real-world problems in different domains, stretching from retail and banking to medicine and healthcare, are tackled while using machine learning models (Linardatos et al., 2020). These algorithms enable researchers to categorize and classify diverse data points, such as survey responses, textual data, or social media posts, into meaningful groups. Social scientists may now uncover hidden patterns, spot trends, and make predictions that were previously difficult or impossible to do using only traditional statistical methods by using the power of machine learning. This makes it easier to understand how people interact, prioritize, and make decisions, which eventually advances knowledge in disciplines like sociology, psychology, political science, and economics. According to Ballestar et al. (2019), in order for big data to maximize its effectiveness in policy studies, it is crucial to employ multi-disciplinary approaches. These approaches should not only rely on the latest computational algorithms derived from machine learning literature but also incorporate the methodologies and practical knowledge gained from decades of multi-disciplinary research that utilizes empirical evidence to guide policy-making choices. Additionally, automated analysis is a potential benefit of machine learning classification algorithms. This cuts down on the time and labor needed for data processing and creates new opportunities for extensive research that may adequately represent the complexity and diversity of social phenomena. The advantage of ML methods over traditional statistical tools is that they provide new approaches to improve the estimation of causal effects, thus reducing the reliance of these estimates on modeling assumptions and enhancing the credibility of policy analysis (Mumtaz, 2022). Overall, the use of machine learning classification algorithms in social science research has enormous potential for improving our understanding of human behavior and guiding solutions that are based on solid research for societal improvement.

Child labor is a serious problem that affects everyone in the world and has negative social, economic, and ethical effects. It entails putting kids to work in jobs that are bad for their

social, mental, or physical growth. Like many other nations, Turkey has struggled with the issue of ending child labor and protecting the rights and welfare of its young people.

The purpose of this study is to develop a predictive model that can effectively identify instances of child labor in Turkey. We use the logistic regression, naive bayes, and random forest machine learning techniques to do this. These algorithms have been successfully used to classification problems in many different fields and have yielded encouraging results.

Chapter 2 presents a comprehensive literature review on big data analysis, data mining, artificial intelligence, and machine learning. It offers a summary of various machine learning approaches, such as supervised learning, classification, regression, unsupervised learning, and semi-supervised learning. Additionally, it delves into the Logistic Regression, Naive Bayes, and Random Forest classification techniques that were used in this study.

Chapter 3 focuses on the results obtained from applying these classification algorithms to predict child labor in Turkey. It begins by stating the application's goal and giving a brief summary of child labor in Turkey using published research. A detailed data analysis is then carried out after defining a dataset that is specific to child labor in Turkey. This investigation makes use of descriptive statistics as well as logistic regression, Naive Bayes, and random forest models to investigate child labor in Turkey. The effectiveness of various models is then compared and assessed.

Chapter 4 presents a discussion and draws conclusions based on the results obtained from the analysis. The results of this study help us understand child labor in Turkey better and provide light on the effectiveness of various machine learning algorithms in foretelling occurrences of child work. These findings have ramifications, which are explored along with suggestions for stakeholders and policymakers.

This study attempts to increase the precision and effectiveness of identifying child labor situations in Turkey by using machine learning techniques. The results of this study can ultimately guide targeted initiatives and policies meant to discourage child work, advance child welfare, and defend children's rights in Turkey.

CHAPTER 2 - LITERATURE REVIEW

2.1. Big Data Analysis and Data Mining

Data collection, commerce, science, technology, etc. are well recognized for having numerous advantages in specific fields. In 2012, the World Economic Forum (WEF) declared that data had a specific value similar to money. In this example, a significant amount of political and financial information is discussed. On the other hand, finding the data in question's financial equivalent is challenging. The financial compensation that people give to organizations in this situation is hard to quantify. For instance, "2.5 petabytes" of data were given by "Wal-Mart Stores" from its consumers during a 24-h cycle (Doğan and Arslantekin, 2016: 21).

Francis X. Diebold (2000) initially discussed the big data phenomenon in the title "Big Data Dynamic Factor Models for Macroeconomic Measurement and Forecasting," was released at a US financial conference. In a speech in the year 1998, a researcher by the name of Mashey referred to "Big Data and the Next Wave of InfraStress." In 2001, Meta Group Inc. went on to describe the relevant principles. In this sense, Meta Group has made significant contributions to the notion of big data (Gürsakar, 2013: 20).

The researchers claim that big data is not fully and precisely defined. Data that takes a great deal of room in a computer system is what the term "big data" originally referred to software developers require the latest technology to examine this data. Yet because of the software that browsers have created, massive data can now be controlled. The main cause of this is that technology has advanced through time in a way that makes data monitoring straightforward (Doğan and Arslantekin, 2016: 22).

Big data is often defined as "datasets that exceed the storage, administration, and processing capabilities of the utilized programs." Big data analysis could be categorized based on its size and substance. In order to address the complication, new technological instruments have been created. Big data is the term used to describe the technologies used to store, organize, and analyze data on an issue (Cackett, 2013: 14).

Big data offers the collection of various info, along with processing and evaluation. Also, Big data must be analyzed in accordance with the data's properties in order to expose its meaning. Simple data analysis reveals that this transaction's format is insufficient. Big data

analysis, in general, includes activities like combining several sets of data, handling massive amounts of data, and handling this amount of data quickly (Erl, Khattak, & Buhler, 2016: 19).

The map-reduce program has the following organizational structure: "a software that splits massive issues into comparatively tiny, controllable sub-problems, then disperses them to a variety of servers known as nodes, factions the data blocks mapped based on the keys, and after they are examined at each sub-node, they are rebuilt and significantly lowered to the intended outcome" is the structure. In the map phase, the master node in the distributed database concept gathers the data input, divides these inputs, separates them into sub-jobs, and then sends them on to the worker nodes. Worker nodes carry out the sub-jobs allocated to them based on the job follower's direction, classifying the outcomes as data sets (key, value), and storing the results in local storage systems that are available for the subsequent Reduction phase. The master node acquires the outcomes from the worker nodes and decreases the valued information to correspond with the key data during the reduction phase (Aktan, 2018: 13-14).

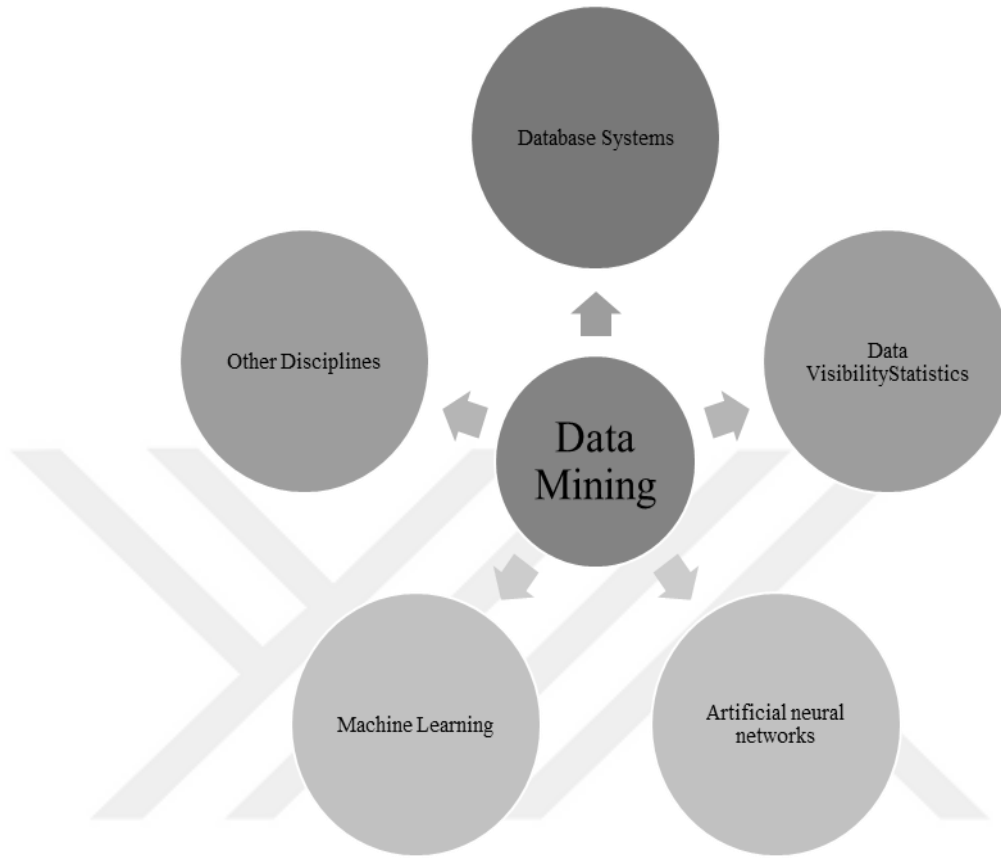
Doğan and Arslantekin (2016) used data mining to evaluate the connections, shapes, strategies, concepts, etc. that are concealed in digital word documents focused on the Semantic structure of Natural Languages. "A method of investigation designed to target and discover it."

Getting information from already-existing data is the definition of the term "data mining." The idea in issue makes it feasible to transmit information from large amounts of data automatically. The advantages of data mining may be shown in areas like analysis, modeling, and subject development across many scientific disciplines. In addition, through gathering, it offers important data to the surface (Oğuzlar, 2011, p.5-6).

According to Savaş, Topaloğlu, and Yılmaz's (2012) research, "Data mining is the discovery and evaluation of significant and practical relationships and rules using computer algorithms that will aid in future prediction from enormous volumes of data. Data mining is a data analysis method that also makes it possible to retrieve confidential details from database management systems and aids in connecting very vast volumes of data by looking at their connections. Figure 1. shows the infrastructures of data mining.

Figure 1. Data Mining

Source: Savaş, Topaloğlu, and Yılmaz's (2012)



Big Data is the presence of several distinct types of data. In order to create large data, new innovations are needed. The necessity for technology led to the creation of modern vehicles. Big data businesses have created these technologies (Salur, 2016: 20).

Data mining is a two-stage process. These stages are displayed as estimating and describing. Forecasting: It gives future data value predictions using parameters. Describe; it displays the values and presumptions that are dependent on estimations (Köseoğlu and Demirci, 2017: 2228).

2.2. Artificial Intelligence

The process of generating both human and animal behavior in a synthetic setting is referred to as artificial intelligence. Experiments using human intellect, as an illustration, have shown that technology includes properties like categorization and grouping in this

environment. Also referred to as artificial intelligence technologies are methods like "intelligent machines, genetic programming, fuzzy control, artificial neural nets, and computer vision." In addition to these tactics, living creatures are looked at in an effort to mimic nature, and comparable clever strategies are recommended. Artificial intelligence optimizing strategies employ procedures like ant colonies, particle swarms, and artificial bees. It is defined as "transferring the neurological and physiologic framework of the human intellect, such as the nervous system, genetic sequence, and natural occurrences to machinery (computers and software) via modeling." In a nutshell, artificial intelligence is a type of technology created using human intelligence. Today's commonly utilized "computer technologies" and "machine learning" are highlighted as key developments in artificial intelligence technology (Atalay and Çelik, 2017: 158-159).

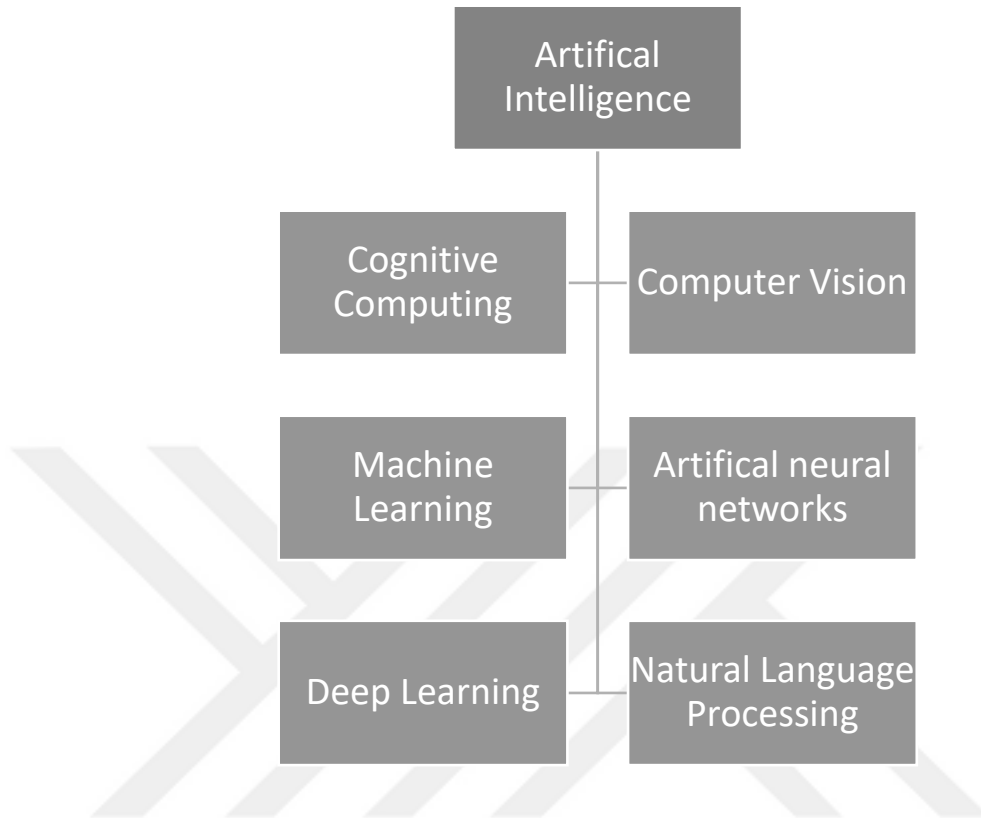
Artificial intelligence technology is referred to be technical items that do people's behaviors. Digital technologies enable individuals to think, feel, talk, and perform actions that they would not be able to do otherwise (Demirhan, Kılıç and Güler, 2010). Machines' reactions and activities in response to circumstances such as people can be attributed to artificial intelligence. The goal of artificial intelligence is to transmit humans' psychological and behavioral intelligence to machines (Atalay and Çelik, 2017).

“Artificial intelligence, according to McCarthy: "Intelligent robots, in particular, is the science and engineering of creating intelligent software programs." According to Silage's concept of artificial intelligence, it is viewed as a heuristic programming-based method. According to Popov, artificial intelligence is characterized as people's ability to make machines do the work that they have done. “Ax” described artificial intelligence as a component aimed toward clever programs in his study. Artificial intelligence research focuses on intelligent behavior. The primary purpose here is to develop a theory that seeks to artificially replicate the intelligent behavior of creatures in the environment”. (Aylak, Oral and Yazıcı, 2021, pp.74-93).

Artificial intelligence application is a complex and difficult technology. Artificial intelligence has its own sub-dimensions. These are shown in Figure 2.

Figure 2. Sub-Dimensions of Artificial Intelligence

Source: Aylak, Oral and Yazıcı, 2021, pp.74-93



2.3. Machine Learning

In 1959, Arthur Samuel defined computer vision and the core of machine learning as the capacity of computers and algorithms to operate on their own. The programming technique in the present system is not done by hand. The supplied data is automatically categorized when the model is determined during processing. For machine learning to be accomplished, the data must be vast enough. Methods based on machine learning emulate the motions and actions of human behavior in a sequence with artificial intelligence. Despite the disparities in learning challenges, it frequently attempts to emulate the environment according to the data and experiences collected by people since the origin. The way that humans use their brains, processes relating to human evolution, the theory of human intellectual growth and cognition, and the sociological theory underlying individual interactions all serve as motivations for machine learning approaches (Aylak, Oral and Yazıcı, 2021: 78).

The general name for software algorithms that define an issue depending on its data and related issues is machine learning. The representation created using the present collection of data and the method used is tuned for best performance. Many machine learning

techniques have been developed as a consequence, ‘‘including the algorithms; k-nearest neighbor, basic (naive) Bayes classification, decision trees, logistic regression analysis, k-means, support vector machines, and artificial neural networks.’’ Certain members of these structures have estimation skills, while others have clustering and classification skills (Atalay and Çelik, 2017: 161).

Research in machine learning focuses on two related questions: How are computer systems built that instantly get better with use? What are the core principles of statistics, computing, data, and theory that govern all learning systems including computers, people, and establishments? The research of machine learning is vital for tackling basic mathematical and technical concerns, such as for the immensely beneficial program, it has brought about and positioned in a wide range of industries. During the last twenty years, artificial learning has evolved from a scientific curiosity to a revolutionary technology in broad commercial usage. Machine learning has become the most common approach for producing realistic technology for machine learning, speech synthesis, processing of natural languages, motion control, and other usages within artificial intelligence. The majority of AI researchers now acknowledge that, to be used in many motives, training a program by offering it instances of wanted input output patterns is significantly simpler than programming it manually by estimating the feedback immediately for all likely inputs. The impact of machine learning is additionally experienced extensively throughout computer programming and a variety of facilities dealing with data intensive challenges, such as customer experience, deficiency indicatives in dynamic structures, and logistics network management. As machine-learning methods have been designed to assess high-throughput empirical values in unexpected ways, there has been a comparable wide range of consequences throughout experiential fields, from biology to cosmology to sociology. A learning challenge is the difficulty of increasing certain performance indicators when doing a function through a certain sort of training experience. For instance, while learning to identify credit card theft, the aim is to categorize each payment ecosystem as "fraud" or "not fraud." The reliability of these crime classifiers may be the output measure to be improved, and the training experience may consist of a database of past credit-card transactions, each identified as fraudulent or not in hindsight. Instead, a new performance metric could be defined that gives a larger consequence when "fraud" is wrongly labeled "not fraud" than when "not fraud" is erroneously classified as "fraud." An alternative form of training experience might also be

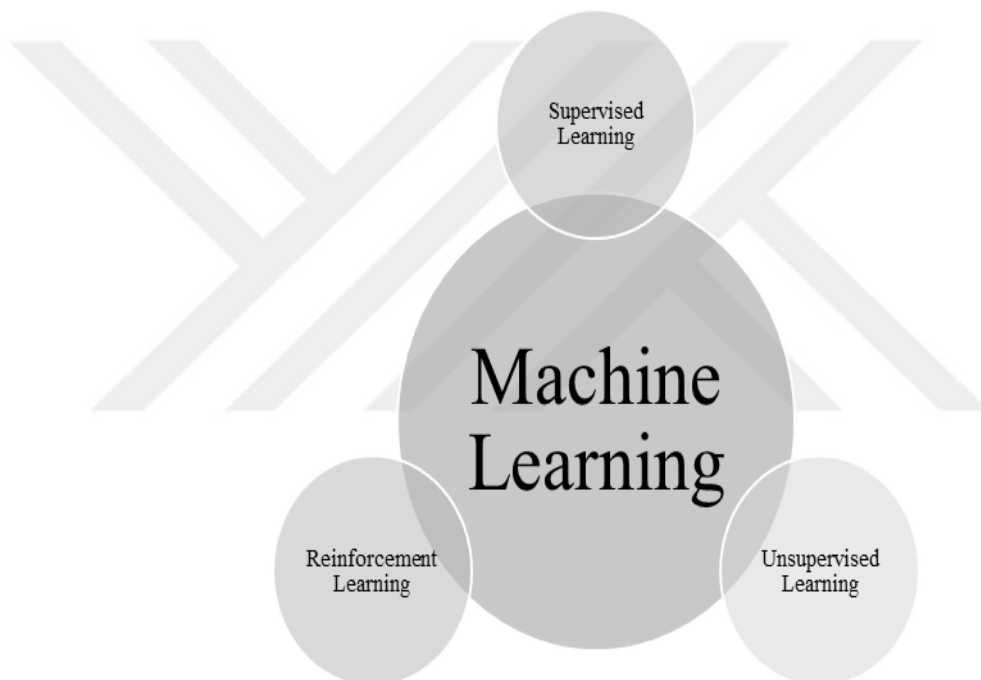
constructed, for example, by adding unmarked credit-card operations with annotated instances (Jordan and Mitchell, 2015: 255).

The machine learning mechanism is portrayed as the newest tech of artificial intelligence. Machine learning applications are referred to be a result of artificial intelligence.

Machine learning is divided into 3 in itself. These are shown in Figure 3.

Figure 3. Machine Learning Strategies

Source: Marsland, 2015



Supervised learning: Targets are generated between the available data in the framework and their counterparts, and the ones that are nearest to the values created between them are chosen. In a nutshell, it is the similarity between the present values and the desired values. (Hannah, Daniel, and Saskia, 2019).

The intelligence of an unsupervised learning system allows it to learn data. Expected values are not assigned to data in this application. Unsupervised learning performs exceptionally well at clustering algorithms. Resemblances may be recognized since the data-entering process is the same. (Marsland, 2015).

Reinforcement learning: Because the system does not know the ideal situation at the start of the process of learning, responses must be found repeatedly. Sensible techniques are rewarded in this process, while incorrect moves are penalized. The system may take into account several environmental impacts and respond properly. As a result, the system can design its alternatives with incentives and penalties" (Aylak, Oral and Yazıcı, 2021: 78).

Below, machine learning methods are shown in terms:

- Supervised learning
- Classification
- Regression
- Unsupervised learning
- Semi-supervised learning

There are 3 types of machine learning strategies in Figure 4. There are three types of supervision: supervised, unsupervised, and reinforced. In supervised learning, the data is recognized as well as the expected result from the data. Supervised learning provides the machine data and the outputs of that data. It enables the extraction of a component (a connection between the input data and the outcome data) from this information. As a result, the machine discovers the link between the data. Supervised learning tasks are classified as "regression" or "classification."

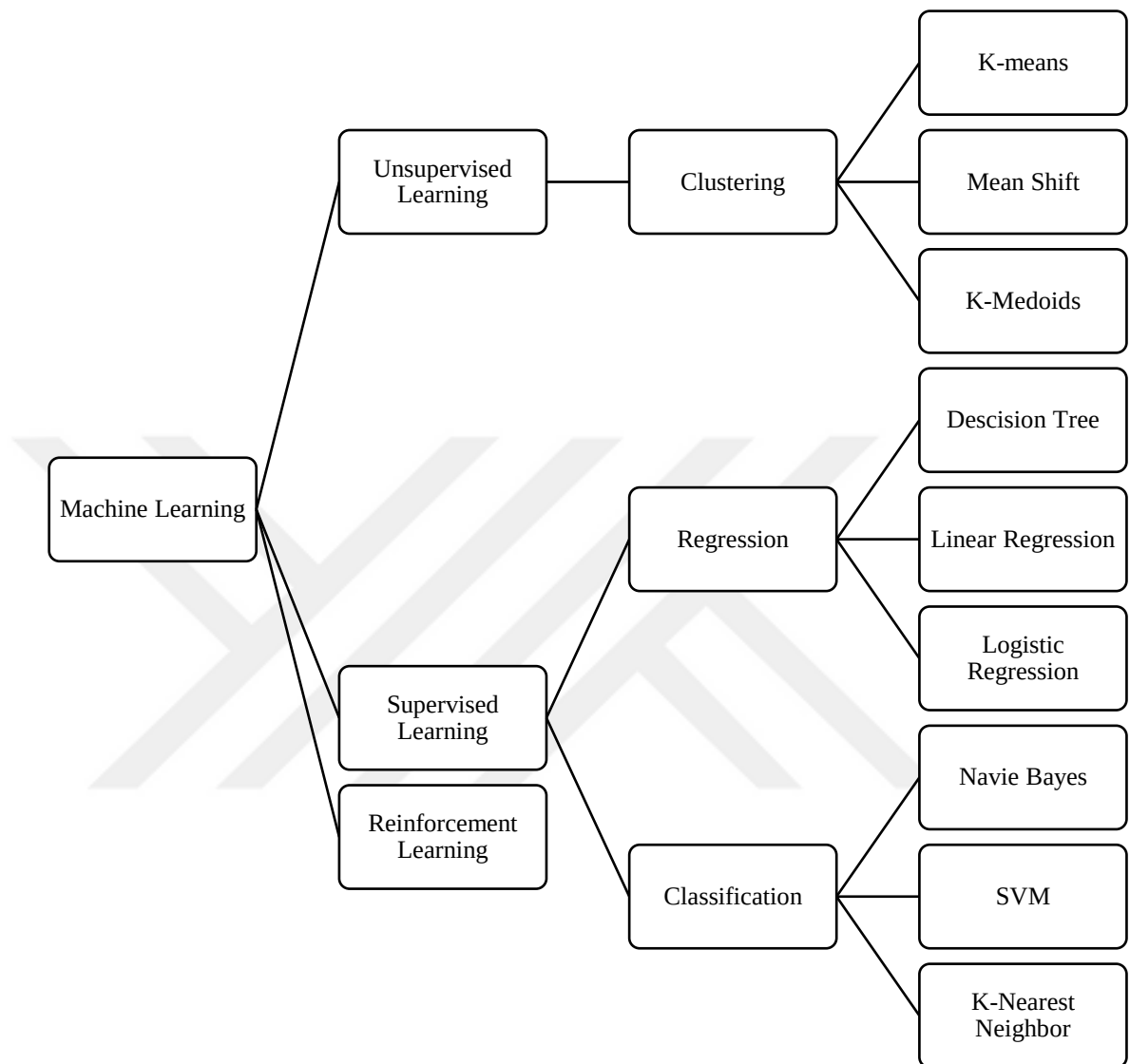


Figure 4. Machine Learning Algorithms

Source: Marsland, 2015

2.3.1. Supervised Learning

Predicting the conditional expectation of a vector result based on a set of variables or characteristics is a fundamental topic within the ML and econometric works of literature. Let Y_i represent the unit i result and represent X_i the K -component vector of variables or characteristics. The dependent assumption is:

$$g(x) = \mathbb{E}[Y_i | X_i = x].$$

There are some theoretical discrepancies between the ML literature and typical econometric textbooks. There are frequently numerous variables in the contexts studied in the ML literature, sometimes even more than there are divisions in the representative. In the ML literature, there is no assumption that the depending on distribution of the result of the present study the variables follow a certain parameterization. The differentiations of the conditional assumption for each control variable, which corresponds to the components in the linear regression framework, are not of inherent value. The emphasis is rather on out-of-sample forecasts and their validity. Additionally, contrasted to many economic applications, there is a smaller variance in the notion that the conditioned expectancy is monotonic in each of the variables. There is a frequent worry that the conditioned expectancy may be a very nonmonotone expression with significant higher-order correlations (Athey and Imbens, 2019).

The effectiveness disparities between certain older approaches, as an example of kernel regression, and contemporary techniques, such as random forests, are especially obvious in scant environments which have a high percentage of higher or lower extraneous variables. Random forests are good at detecting sparseness and disregarding unnecessary characteristics, even when there are lots of them, whereas standard kernel method applications effectively squander degrees of freedom on compensating for such variables. Although it is theoretically conceivable to adjust kernel techniques to the inclusion of unnecessary covariates by enabling covariate-specific bandwidths, minimal effort has been made in this regard in operation. The current approaches are also especially great at identifying extreme nonlinear effects and high order correlations. The inclusion of such high order correlations within a few of these methodologies' great achievements should not blind us to the reality that we strive for continuous improvement correlations to be of low consequence with many economic data. If we aim to estimate wages for people, we should anticipate the degenerate task to be monotonic in variety of the significant indicators, such

as schooling and past earnings, also for homogenous subsets. Regression function also indicates that projections that depend on linearizations may outperform alternative approaches in such circumstances, as opposed to contexts where monotonicity is essentially less realistic, such as in visual detection and recognition. This is also one of the reasons why locally linear random forests outperform ordinary random forests (Friedberg et al. 2018).

A straightforward machine-learning classification breaks the learning process into two phases: training and testing. In the training phase, portions of the training data are utilized as input. During this phase, the training algorithm or learner learns the components, which are then used to create the learning approach (Sandhya et al. 2016). Using the execution method, the learning model predicts the tested or produced data during the testing phase. The outcome of the labelled data is given by learning models that offer the final estimation or classified data.

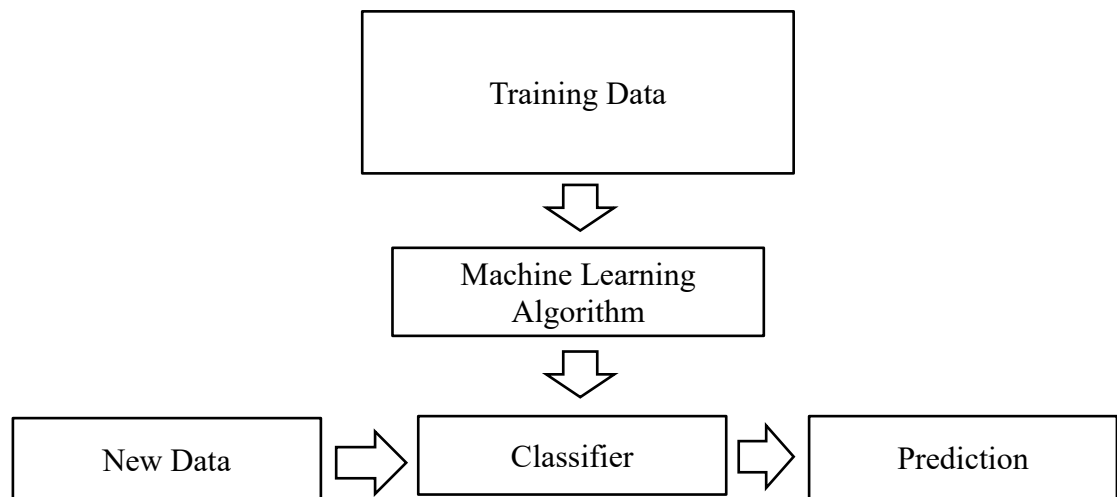


Figure 5. Supervised learning process

Source: Nasteski, 2018.

The essential aim is generally to stimulate the machine to absorb a scheme that we've assembled, supervised learning (Figure 5.) has become the most commonly used approach in classified challenges. Typically, supervised learning keeps the possibility for input unknown, like an input with a known predicted result. This procedure generates a dataset with features and labels. The primary aim is to build a predictor that can estimate the title of an item based on a collection of characteristics. The learning algorithm then gets a collection of characteristics as inputs, as well as the right outputs, and learns by contrasting

its real output to the rectified outputs to detect faults. It then adjusts the model as needed. The modelling is not required as long as the inputs are accessible, but it is impossible to deduce anything regarding the conclusions if part of the input data is absent. The most frequent approach for training neural networks and determining trees is supervised learning. Both rely on the data provided by the pre-determinate categorization. This method is also employed in scenarios where previous data is utilized to forecast likely featured instances. There are several practical applications of this learning, such as software that guesses the classification of an iris based on a set of flower dimensions. As previously noted, supervised learning projects are classified into two types: classification and regression. The label in classification is discontinuous, but the tag in regression is continual.

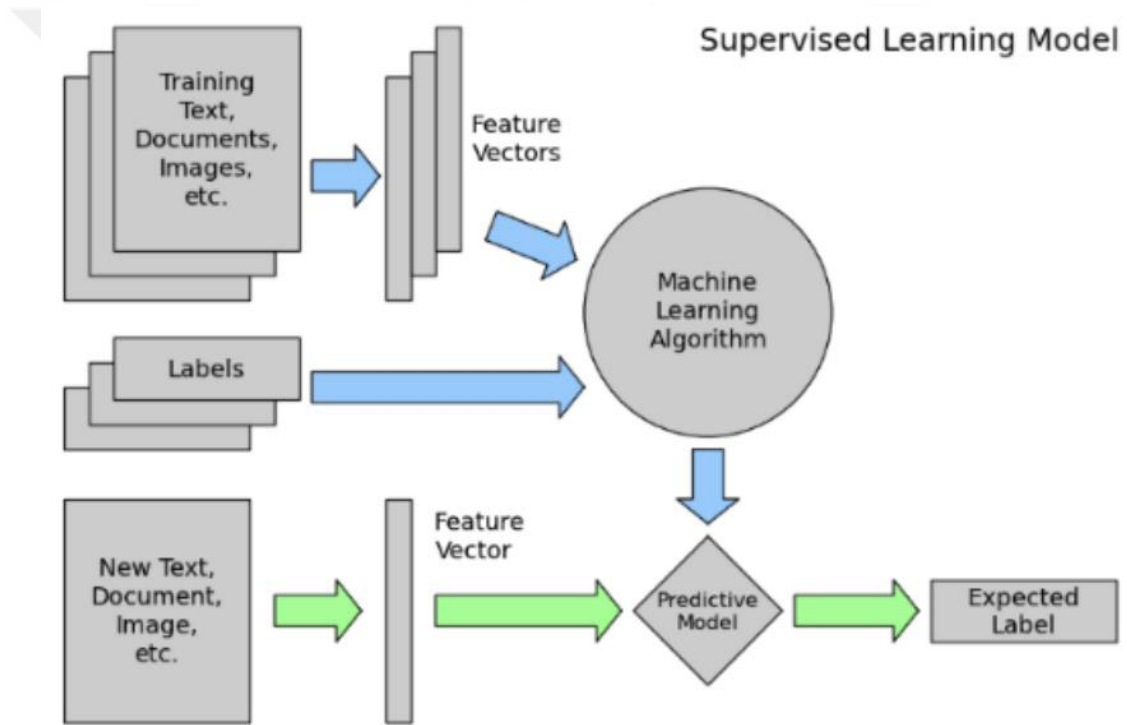


Figure 6. Supervised Learning Model

Source: (Sandhya et al. 2016).

The method makes a distinction between the observed values X and the testing dataset, which is frequently formatted data given to the algorithms during the learning phase, as shown in Figure 6. During this whole process, the learning algorithm for supervised instruction builds the forecasting Supervised Learning Model models. This implies that even after learning, the algorithm would seek to identify the most likely tags for a new batch of values X in the testing set. Depending on the target y , supervised learning may be classified as follows:

- The process of predicting y is referred to as classification if it has quantities in a predetermined set of classified outputs (integers).
- The process of predicting y is referred to as regression if y contains exponential notation values.

Guided learning forms the type of classification. The procedure of splitting an incoming data into different disciplines is known as classification. This classification procedure starts by predicting the category of produced data sets. Classification employs predictive analysis. As a result of the analysis, the task from the input parameter to the output parameter is generated. The primary purpose of classification is to determine which class/category the latest data belongs to. In other words, classification is utilized in this supervised learning approach, which is most typically used for predicting, to determine which class a representative is included to. Binomial and multi-class segmentation are the two forms of classification. Samples of categorization include whether an e-mail is superfluous, if a person has cancer, and if a sports team wins or loses (Lantz, 2015). This categorization style denotes binary categorization. Many class issues require determining which of the several groupings an input belongs to. Classification is a supervised learning strategy in which the outcome has predefined tags (discrete value). Decision trees, support machines, logistic regression, and Naive Bayes are the most common classification algorithms (Yavuz, 2021).

The other major area of the supervised learning literature focuses on classification difficulties. With series of findings on a vector of qualities X_i and a label Y_i (an unsorted separate result), the aim is to find a method that allocates fresh units to one of the tags based on their characteristics. This is quite similar to the discontinuous decision survey in econometrics, in which analysts construct statistical models that suggest a likelihood that the output would take on a specific amount based on the variables (attributes). With a similarity, it is obvious to forecast a particular tag, specifically the one with the best odds. There are, however, distinctions between the two methods. One key distinction is that in the studies on classifications, the emphasis is frequently exclusively on categorization or the selection of a target only one label. One can categorize provided a possibility for each labelling, but such a probability is not required for classification. Several of the categorization methods do not, in reality, predict a possibility for each tag and, as a result, are not immediately applicable in circumstances where such a possibility is necessary. A pragmatic distinction is that the categorizing research has frequently concentrated on

scenarios where the variables finally allow one to allocate the name with nearly full confidence, as contrasted to conditions where is the best attitude have significant error rates. Quantity recognition provides the most popular illustration. Depending on a visual encoded as a set of, say, 16 or 256 black and white pixels, the aim is to classify a visual as corresponding to one of the 10 integers from 0 to 9. In this situation, ML techniques have shown to be incredibly successful. Support vector machines (SVMs) considerably outperformed earlier techniques in the 1990s (Cortes and Vapnik 1995). Recent advances in multilayer convolutional neural network (CNN) modelling have increased error margins significantly (Athey and Imbens, 2019).

Regression is another supervised learning approach for forecasting. The regression dataset is made up of variables that serve as inputs and outputs. In other words, each time the algorithm is checked, it generates a result score in response to the inspection (Elitedatascience, 2019). If the intended output value is made up of continuous data, this learning issue is known as regression. The fundamental purpose of both continuous and discrete data regression algorithms is to estimate. Predictions that indicate a linear connection between the qualities can be utilized to explain some instances. For example, if sales rise, the expenditure should rise as well (Polamuri, 2014). The operational connection between input and output data is found by utilizing training data. When the result data is known, the function's variables are 12 learned. As of current, it may offer estimations without knowing the computer's input. Reconstruction is a continuous real outcome supervised learning method. Regression techniques used in supervised learning techniques include linear interpolation, decision tree correlation, and random forest regression (Yavuz, 2021).

2.3.2. Unsupervised Learning

In unsupervised learning, the set of data is not defined by a title. To alter, this system attempts to be trained. Segmentation, anomaly, uniqueness identification, and display are all part of the unsupervised learning technique. It may be used for a variety of reasons, including size reduction, and learning the ensemble method (Meltem, 2019). The purpose of supervised learning is to connect the inputs with the outputs. Unsupervised learning lacks a supervisor and is limited to input data. The goal is to identify variations in input. A framework is created for the supplied data, and this framework is attempted to be identified

(Abdallh et al., 2016). In a nutshell, the incoming data is unlabelled and has no anticipated outcomes. It is set up to identify forms from input data (Brownlee, 2016).

In this scenario, the aim is to divide the extracted features into subspaces given a collection of findings on elements X_i . Depending on subspace participation, these groups may be leveraged to develop new functionalities. For example, we might utilize splitting to estimate parsimonious frameworks inside each subdomain. We may also want to employ group affiliation to arrange the sample into distinct sorts of entities that may be exposed to different receptions. This is an uncommon challenge to the extent that there is no practical benchmark to determine if one solution is superior to another. The mixture model, which is more conventional in the econometrics and statistics pieces of literature, is a similarly comparable technique in which the allocation that produced the example is described as a combination of multiple distributions. The groups and the mixture elements are like one another in nature (Athey and Imbens, 2019).

The K-means technique is a crucial strategy (Hartigan & Wong, 1979). Consider a situation where we wish to segment the categorization framework into K clusters or subgroups. To assign cells to the scheme based on their proximity to the centroids, we should first want to select the clusters $b_1, \dots, b_K$. The basic algorithm is listed below. We start with such a set of K subspace parts called centroids, $b_1, \dots, b_K$, which are distributed evenly over this space. Allocate every unit to the groups that have the shortest distance between it and the cluster's centroid, given an assortment of centroids:

$$C_i = \arg \min_{c \in \{1, \dots, K\}} \| X_i - b_c \|^2$$

The cluster centres are then updated as the median of the X_i for every cluster:

$$b_c = \sum_{i: C_i=c} X_i / \sum_{i: C_i=c} 1$$

Recapitulate between the two stages several times. Selecting the cluster count K is problematic as it lacks a direct cross-validation technique to compare one quantity to another. This number is frequently determined on substantive concerns rather than on the basis of statistics.

2.3.3. Semi-supervised Learning

Certain data in semi-supervised learning contains a class label, but some have no data about which category it corresponds to. The majority of semi-controlled algorithms. It is made up of both supervised and unsupervised learning methods. Deep belief networks (DBN), for example, are referred to as restricted Boltzmann machines (RBM). It is built on combining unsupervised learning features. The entire system is then assessed using a supervised learning technique (Meltem, 2019).

For semi-supervised training to function, the underlying margin-based data $p(x)$ from the over-input vector should include data on the prior distribution $p(y|x)$. In such a case, one might be able to find additional information regarding $p(x)$ and, consequently, $p(y|x)$ by using unsupervised learning. Nevertheless, whether this criterion is not met, and $p(x)$ has no information about $p(y|x)$, it is almost difficult to increase the accuracy of forecasts made using further unsupervised approaches (Zhu, 2008).

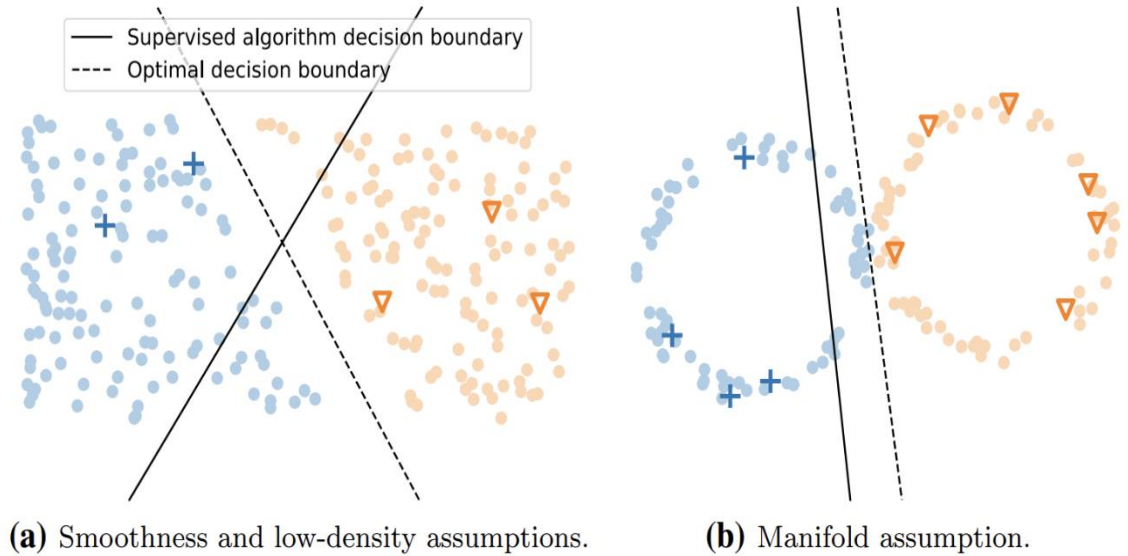


Figure 7. Illustrations of the semi-supervised learning assumptions

Source: Engelen and Hoos, 2020

Thankfully, most practical learning contexts appear to satisfy the condition, as evidenced by the successful use of semi-supervised learning techniques in actual practice. Nonetheless, the interaction between $p(x)$ and $p(y|x)$ need not be the same. Therefore, semi-supervised training concepts had also developed, defining the types of anticipated involvement (Chapelle et al. 2006b). The far more widely accepted assumptions in Figure

7 are the reduced structural rigidity assumption (the differentiator should not transmit through the use of greater intensive areas in the training dataset), the manifold presumption, and the cohesiveness assumption (if 2 offered sample data x and x are connected in the input data, their classifications y and y must correspond to one another) (data points on the exact low dimensional manifold must have the identical tag). The vast, at least the majority, of semi-supervised learning techniques are founded on the notion that such conditions have been satisfied, either explicitly or implicitly. We'll examine the essential premises that underpin every learning method throughout this test.

2.4. Classification Algorithms

Datasets have untapped knowledge that may be used to build smart systems for making decisions. (Witten et al. 2016) Most sorting and forecasting techniques are used in systems that support decisions. It is possible to determine major categories of info, derive modelling from data, or forecast foreseeable data patterns using the two types of data analysis known as classification and prediction. These analyses can aid us in comprehending the larger body of data. Categorization models foretell category labels. In contrast, continuous-valued variables are specified in estimating frameworks. As an illustration, the bank may develop a classification model to categorize applications for loans as safe or dangerous, and the model's output could be a categorized listing of new applications that are either safe or hazardous. The bank can develop a prediction algorithm to determine clients' daily or monthly costs by taking into account their income and line of work (Witten et al., 2016). Experts have created a variety of categorization and forecasting techniques that are utilized in machine learning, detection, diagnosis, and analytics. These algorithms are employed in data mining studies and the extraction of information, sequences, and meanings from large data sets.

In many significant studies nowadays, classification methods are applied. For instance, a company's marketing manager could want data analysis to assist in determining if a consumer with a specific profile is going to buy a new gadget based on their prior purchases. A doctor can forecast which of the three therapies a breast cancer patient ought to undergo by studying data about the disease. Classification methods are used in each of these cases to analyse the data. For marketing data, the model of classification is designed to foresee class labels like "yes" or "no," and for breast cancer, "treatment A," "treatment B," or "treatment C."

The numerical estimate is a crucial subject and is employed in many contexts in addition to the classification framework. For instance, the dosage of a medicine to be utilized during the course of treatment of breast cancer is a monetary amount that can aid in the recovery of the individual. A qualitative class label as well as a numerical value can be estimated by classifying techniques. The most popular statistical technique for making numerical forecasts is regression analysis. Consequently, the two names are frequently used interchangeably.

For the loan implementation, there are two steps in the data categorization phase. A classification model is built in the first phase, and it will be utilized for learning a set of training data with specified class labels. Using the class-labelled dataset as its training set, this categorization algorithm goes through a learning or training phase. Since each training sample is given a class name, this process is sometimes referred to as learning under supervision. Below in Figure 8 is a screenshot of the classification model developed using the loan application's data set for training.

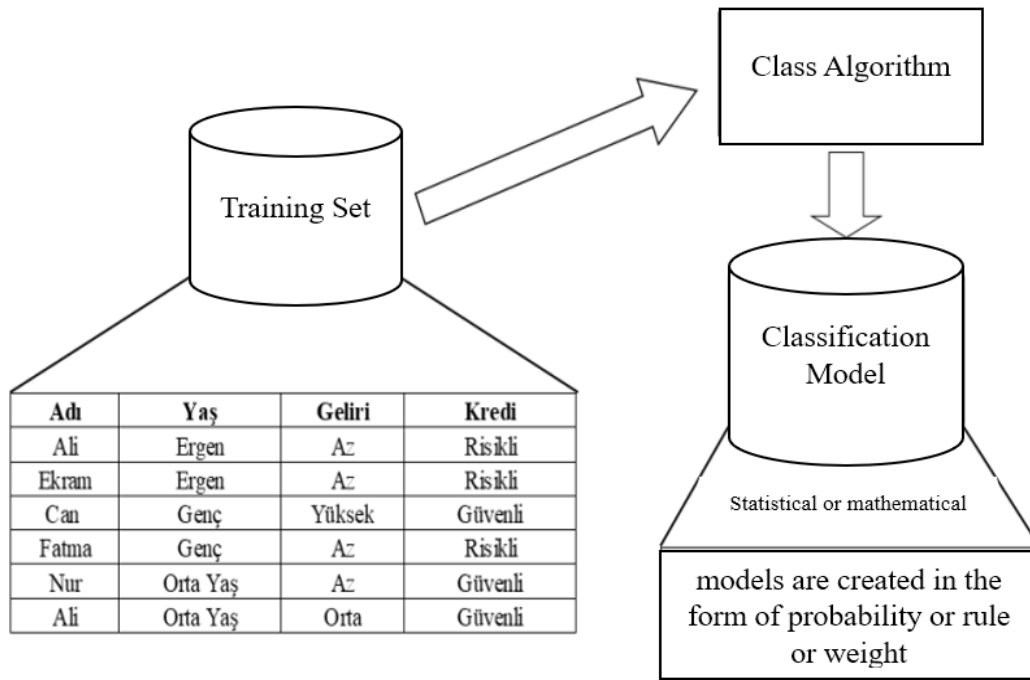


Figure 8. Classification Model for Loan Application Data

Source: İbrahim, 2019.

The dataset in the graphic is represented as $V = \{I_i\}_{i=1}^m$. Here $I_1, I_2, \dots$. The dataset's samples are represented by I_m , and every single one is denoted by $I_j = "A, C."$ Here, $A =$

a_j $j=1$ stands in for the feature set n and $C = c_k$ $k=1$ for the class set. Both discrete and continuous values are possible for attributes and class names. The examples that make up the training dataset are referred to as sample training sets, and the classification model is trained using these training samples. The following step involves testing the classification model developed with the training set using the test data set or a fresh sample. Below is an example of the classification system developed for loan applications.

The model built from the training data is utilized for forecasting and classification procedures, as shown in Figure 9. The very accurate categorization of samples that are not part of the set used for training and the category value unknown is what may be anticipated from the trained classification model. We would probably have excellent results if we tried to evaluate the classification model's performance using the training set. because samples were used in this training while the categorization model was being developed. Therefore, a test group of test samples and related class labels must be employed to appropriately assess the effectiveness of the classification model.

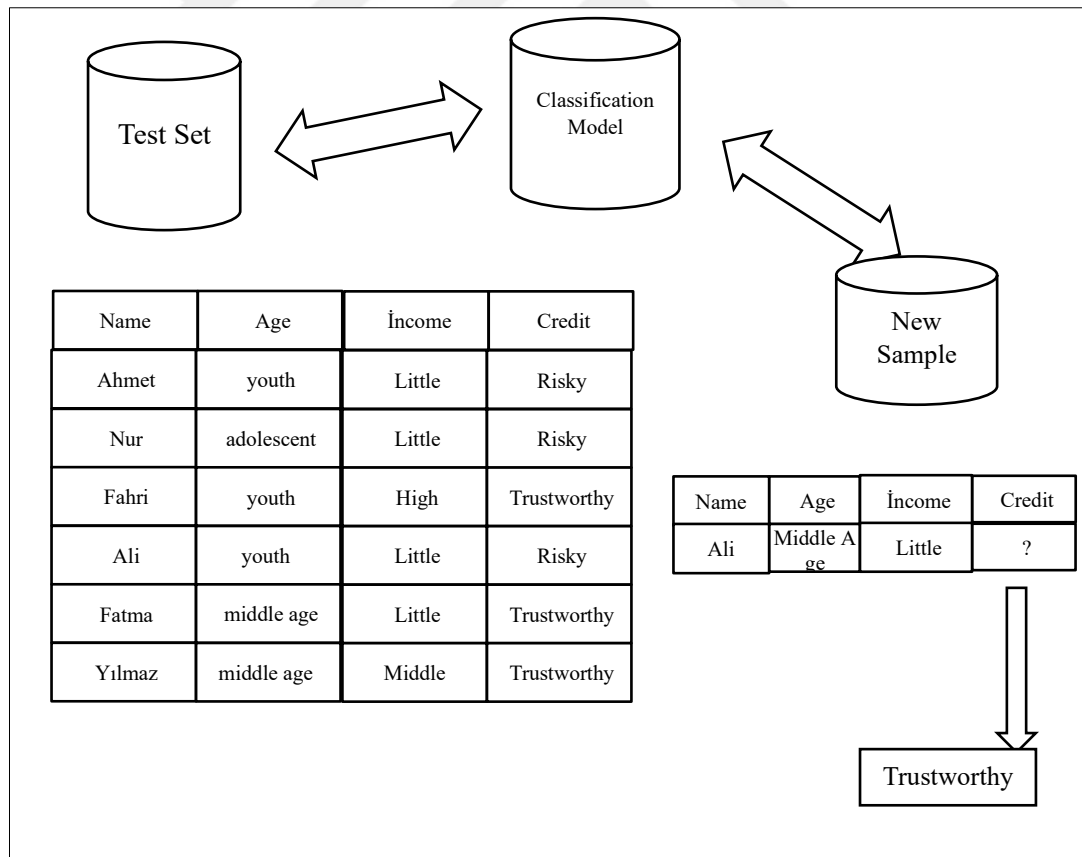


Figure 9. Classification Model for Loan Application Data

Source: İbrahim, 2019.

From the whole dataset, these test examples were chosen at random, and they should be separate from the samples used for training. They are thus not applied throughout the classification model's training process. The proportion of test set samples (for example, 90%) that the categorization algorithm properly classifies as belonging to a specific test set indicates the classification model's performance on that test set. Comparisons are made between the test samples' actual category labels and the classes that the classification model predicts for them. The Criteria for Evaluating the Performance of Classification Algorithms section explains how to determine the classification system's effectiveness.

Comparable to the classification model, the numerical forecasting approach has two steps. The class label, nevertheless, loses its meaning in the aforementioned model of prediction since the projected values are numerical forecasts rather than categorical ones. Let's say that a user of the categorization model determines that the example we just gave is safe. This estimation model produces a predicted numerical number for this customer's loan quantity. Their approaches for creating their categorization and models for prediction also vary. Similar to classification, the forecasting model's performance shouldn't be assessed using the training set that produced the prediction. Instead, a separate test set ought to be employed. By computing an error based on the discrepancy between the anticipated result from the model used for prediction and the known real value for each of the test samples, one may determine if the model used for prediction is successful. There are several errors in estimation metrics, including mean, mean percent, mean absolute, mean square, root mean square, and mean absolute percent error.

Data analysis gets harder when there are more instances, characteristics, and class labels in the classification datasets being utilized to solve classification issues. As additional steps in the classification process, speech recognition, recognition of handwriting, illness evaluation, document categorization, etc. are performed. Applications for it include (Han et al., 2011). The result was the development of classification algorithms such as rule-based (ID3, C4.5, CART, REPPER, etc.), probability-based (NB, etc.), and weight-based (ANN, Logistic, etc.) algorithms (Berry and Linoff, 1997; Wu et al., 2008). The use of machine learning rules-driven and probability-based methods for classification may provide a classification model with outstanding accuracy on continuous and categorical info sets. Nevertheless, since the data from practical uses is continuous, it could be difficult to develop classification models that function well in rule- and probability-based classification techniques. Deterministic and probabilistic approaches to classification

should thus be applied after the continuous information generated from applications has been discretized using discretization techniques. The setup of the data set and the parameters for weighting used in the weight-based classification procedure are both important factors in the learning of weight-based algorithms for classification that determine the effectiveness of the classification model that is produced. As a result, it's important to prepare the data set carefully and pick a strong training (weight update) technique to get a decent performance in the weight-based classification algorithm.

2.4.1 Logistic Regression

In academic disciplines, factors are examined at several scale levels. student achievement: Equal interval can be stated as a continuous variable at the scale level with scaled studies, as successful-unsuccessful at the categorization level, or as a discontinuities constant. The link between these factors, which may be categorized as either ongoing or inconsistent, has been modelled using a variety of statistical analyses. Which approach will be utilized depends on whether the factors being studied are either continuous or discontinuous to characterize the connections between them in a way that has the most appropriate fit. Regression modelling is one of the types of statistical methods widely used to investigate the connections between parameters. Regression analyses' primary goal is to develop a model that accurately captures the link between the forecasted and predictor variables while using the fewest possible variables (Atasoy, 2001).

As with many methods of statistics, the development of the linear regression approach was based on the suppositions regarding the typical distribution, linearity, and correlation of the variable to be predicted. However, we may create a regression model using logistic regression analysis (LRA) without needing to make assumptions about multivariate normality, continuity, or normality in general (Tabachnick & Fidell, 2001, p. 521). In an LRA, the variable that is predicted is inconsistent, the predictor variables can coexist and be continuous or discontinuous, and they can both exist simultaneously. A continuum result variable, or the expected variable, can if required be converted into a discontinuous variable to create the LRA model (Tabachnick & Fidell, 1996).

The LRA and some statistical analyses are comparable. Discriminant analysis is carried out with a discontinuous dependent variable and is a type of grouping analysis, whereas cluster analysis is an analysis that splits data into groups. Numerous linear regression analysis is an assessment similar to LRA in that its goal is to create a regression model. In general,

both multiple regression and discriminant modelling vary from logistic regression analysis in that they make assumptions, whereas cluster analysis has the disadvantage that it is unknown how many groups there are and who belongs to them (Çokluk, Şekerciolu, & Büyüköztürk, 2012).

LRA, which offers a more flexible structure compared to similar analysis, is mathematically based on the logarithm of probability, odds, and odds. Probability is simply the ratio of the number of outcomes of a particular type to the total possible outcomes. For example, the probability of rolling a 3 when a dice is rolled is 1/6. Because there is only one '3' on the dice and there are six possible outcomes. This ratio can be expressed as .167 or 16.70%. In the LRA, odds are defined as the probability of an event occurring divided by the probability of that event not occurring. Odds has no upper limit, but its lower limit is 0. Odds remove the problem of probability estimation being within the range of 0-1, but there is still another problem to be solved, which is how to ensure that the odds ratio does not take a value below zero. This problem is solved by calculating the logit value. Logit is the natural logarithm of the odds ratio. By calculating the logit value, a metric variable that can always be reverted to a probability between 0-1 and can take positive and negative values is obtained (Çokluk, Şekercioğlu, & Büyüköztürk, 2012; Mertler & Vannatta, 2005). The issue of possibility estimations falling inside the range of 0–1 is handled by odds, yet there is still the issue of how to prevent the odds ratio from falling below 0. The logit value is used to fix this issue. The odds ratio's natural logarithm is called logit. A metric variable that may take both positive and negative values and can always be converted to a probability between 0 and 1 is generated by computing the logit value (Çokluk, Şekerciolu, & Büyüköztürk, 2012; Mertler & Vannatta, 2005).

By utilizing the ideas of the odds ratio, the logistic function, and maximum likelihood estimation, logistic regression makes it easier to forecast binary events. This method has applications in many different fields since it provides insights into the connections between input data and binary outputs (Haghighat & Belyadi, 2021).

Odds Ratio:

The odds ratio is a fundamental idea in probability theory. It represents the proportion of an event's likelihood of happening (p) to its probability of not happening ($1-p$).

$$Odds(p) = \frac{p}{1-p}$$

Logistic Regression and Prediction Equation:

This concept is encapsulated in logistic regression by modeling the log-odds of event occurrence linearly in terms of input variables, which are then converted into probabilities by the logistic function. The following is a matrix representation of the prediction equation:

$$\ln\left(\frac{p}{1-p}\right) = X \cdot W + b$$

The alternative representation is as follows:

$$\ln\left(\frac{p}{1-p}\right) = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} + \begin{bmatrix} b \\ \vdots \\ b \end{bmatrix}$$

Here:

- X is the feature matrix representing the input data
- W is the weight vector associated with each feature
- b is the bias term that accounts for any offset in the prediction

Maximum Likelihood Estimation (MLE):

Maximum likelihood estimation (MLE) is a key component of the logistic regression process. This stage comprises determining the ideal weights (W) and bias (b) that, given projected probabilities, maximize the likelihood of witnessing actual outcomes. A definition of the likelihood function L is:

$$L(\mathbf{W}, b) = \prod_{i=1}^m p_i^{y_i} \cdot (1 - p_i)^{1-y_i}$$

By transitioning to the logarithmic space, the log-likelihood (LL) is derived:

$$LL(\mathbf{W}, b) = \sum_{i=1}^m (y_i \cdot \ln(p_i) + (1 - y_i) \cdot \ln(1 - p_i))$$

The model's ability to match predictions with observed results is then improved by employing optimization techniques like gradient descent to iteratively change W and b .

Logistic Regression Analysis Classification in the categorization of the study, the framework of the factor that is a dependent (predictive) parameter in which the logit transformation is used in logistic regression analysis plays a significant role. According to the dependent variable's structure, LRA is split into three groups. The 'Binary Logistic Regression Model' is employed when the categorical dependent variable is binary (for example, gender; female or male), among these models. While the 'Multinomial Logistic Regression Model' is used when the categorical dependent variable has multiple categories (for example, marital status; married-single-divorced), the 'Ordinal (Ordinal) Logistic Regression Model' is used when there are multiple categories and an ordinal structure (for example, Likert-type scales, less-middle-many). (Barak, Karahan, & Saraçbaşı, 2005).

There are instances where the LRA is categorized based on how many independent (predicted) variables there are. This categorization refers to LRA as either "Univariate Logistic Regression" or "Multiple Logistic Regression Analysis" depending on the number of independent variables it contains (Cook, Dixon, Duckworth, Kaiser, Koehler, Meeker, & Stephenson, 2001). When using Linear Regression Analysis (LRA), there are no specifications regarding the distribution of the independent variables that must be met. This indicates that the LRA can be used without assuming anything about the distribution of the independent variables. However, according to Çokluk, Şekerciolu, and Büyüköztürk (2012), Field (2009), Hosmer & Lemeshow (2000), and Tabachnick & Fidell (1996), there are a few presumptions and prerequisites when using LRA:

- The data should be reviewed for omissions and excessive values, and any required changes should be performed. Although this criterion is not regarded as an assumption of the pertinent analysis, it is said that it ought to be taken into consideration before the analysis since excessive values and losses are believed to have the potential to skew the outcomes of statistical tests. The predicted frequency should be larger than 1 for all categorical variable pairs, and 20% of pores should have an expected frequency of less than 5.
- There shouldn't be a problem with multicollinearity between the variables because the LRA is particularly sensitive to strong correlations between independent variables. According to Tabachnick and Fidell (1996), multicollinearity develops when the correlations between variables ($r > .90$) are large. It is sometimes possible to overcome a

multicollinearity issue between the variables used in the study by adding more data, although this is not always feasible. Another approach is to use component analysis to combine these variables into a single variable that is then included in the model (Kalaycı, 2010). It is one of the suggested approaches to eliminate all or some of the elements that create the issue from the framework if the issue of multicollinearity could not be solved with other options. (Tabachnick & Fidell, 1996; Kalaycı, 2010).

- The logit value for the result variable and predictors are assumed to have a linear relationship by LRA. By examining the significance of the link between the predictor and the log transformation, which is one may verify or refute this supposition.
- One of the LRA's underlying presumptions is the independence of mistakes. Overdispersion in LRA is described as the variance that is observed to be higher than the predicted variance (Cox, 1983; Dean, 1992). Since conventional deviations tend to decrease and the confidence ranges for the variables in the regression model are shrunk, excessive dispersion is an issue. By measuring the ratio of the chi-square value to its unique levels of freedom, it is possible to identify overdispersion for the underlying assumption of error independence; Pearson, or Deviation in the goodness of fit values (the one that will result in a lower distribution parameter is desired). The distribution parameter (ϕ) is the name given to this ratio. Overdispersion is indicated when the scatter parameter is larger than 1. Overdispersion can be viewed as a concern, though, (ϕ) if is more than or near to 2 (Field, 2009, p. 273, 276).

Choosing which elements or variables to include in the model is the major goal of the model construction procedure. Instead of simply one greatest model, there will probably be several decent ones in use. The study's objectives will determine how to adapt the model selection technique. There are several ways to choose or identify the factors that significantly improve the model's ability to explain the variation of the dependent variable. When there are three or more independent variables, variable selection becomes crucial (Kalaycı, 2009). The choice of which variable to include in the model is crucial for its formation. There are more potential models available the more independent variables there are.

The approach to be used in the LRA application should be chosen first. LRA can be carried out using either the usual (entry) or stepwise approaches. The two forward and backward stepwise approaches are likewise separated (Field, 2009).

Based on the findings of earlier studies, variables should be included in the model in a perfect investigation. It may be said that the stepwise technique is the most suitable approach for exploratory investigations. According to the conventional approach, parameter estimates are computed for each block of variables that are incorporated into the regression model as a covariate. The usual approach, according to some academics, is only appropriate for testing theories. This is because stepwise approaches are susceptible to arbitrary data changes, and even when the model is run again using the same samples, it is uncommon to have the same outcomes (Field, 2009). The conventional technique can be applied if none of the independent variables are thought to be more significant than the others.

Reporting and analysing LRA results One of the most crucial phases of the research is reporting and interpreting the findings. To meet the fundamental requirements of scientificity, such as transmittance, repeatability, and testability, it is deemed necessary to inform the reader about the findings and comments, to provide evidence that will give the reader confidence in the research, to demonstrate the correct analysis and use it appropriately, and to address the key points regarding the research findings.

The categories to be taken into account while presenting the findings of LRA can be mentioned as follows after reviewing the literature (Field, 2009; Hosmer & Lemeshow, 2000; Tabachnick & Fidell, 1996): The statistical model predictors (Wald statistics), critiques on goodness-of-fit statistics (Hosmer-Lemeshow, Pearson, or Deviation fit statistics), remarks on likelihood or odds ratio, data on model capabilities (from classification tables), and competence interpreting by examination of residuals or with the so-called R^2 statistic are all aspects of general model evaluation.

General Evaluation of the Model

The probability value is the fundamental metric used to assess how effectively the LRA's maximum likelihood estimation suggests a satisfactory match. A model fit index is equal to $-2 \log \text{likelihood} (-2LL)$. By using the likelihood value's -2 logarithm, logistic regression calculates the expected model fit.

Zero, which is the smallest value that $-2LL$ can accept, is a perfect match. In other words, the $-2LL$ value drops as compatibility increases. It is compared to determine how the likelihood value differs from the fit variations across various models. Three phases make up this fundamental methodology (Hair, Black, Babin, Anderson, & Tatham, 2006).

Estimation of the Initial Model

The fundamental idea behind the initial model is to generate an initial value using only the constant (dependent variable) so that you can compare it to the model once the predictor variables have been included in the subsequent phases.

Estimation of Intended Model

This model was developed through study and incorporates predictive factors. The value of -2LL lowers as they fit in the proposed model gets better.

Evaluation of the -2LL Difference

The -2LL discrepancy between the planned model and the baseline model is assessed for statistical significance. It is understood that "the predictor variable or variables in the intended model significantly contributed to the improvement of the fit of the predicted model" if statistical tests indicate that the difference is significant. By applying similar reasoning, it is possible to examine the differences between the two proposed models. The -2LL difference in this instance denotes the variance in fit between the two models. However, because statistical significance is more challenging to achieve in small samples than it is in big ones, these tests are very sensitive to sample size.

2.4.2 Naïve Bayes

The sum rule, which is the two basic rules of probability theory, is stated in Equation 1 and the product rule is specified in Equation 2 (Deng, et al. 2015).

$$P(X) = \sum Y P(X, Y) \quad (1)$$

$$P(X, Y) = P(Y \vee X) P(X) \quad (2)$$

$P(X, Y)$, $P(Y|X)$, and $P(X)$ are the condition, marginal, and common probabilities, respectively, in formulas. The Bayes rule given in Equation 3 may more easily be obtained based on the product rule when the symmetry property $P(X, Y) = P(Y, X)$ is applied (Deng, et al. 2015).

$$P(Y \vee X) = \frac{P(X \vee Y) P(Y)}{P(X)} \quad (3)$$

By using the rule of sum, the equality that appears as a numerator in Bayes' theorem can also be expressed in the denominator (Equation 4) (Deng, et al. 2015).

$$P(X) = \sum Y P(X \vee Y) P(Y) \quad (4)$$

The variable Y serves as the class label for the given data samples, and $P(Y=y_j)$ is used to describe the previous likelihood for the class y_j when determining the chances of the class in the samples of data that are provided. As a characteristic of the data samples, the variable X is represented. The previous probability values of the likelihood of the X data seen for various classes, $P(X=x_i | Y=y_j)$ and $P(Y=y_j)$, are required to estimate the following probability value $P(Y|X)$, which computes the odds of being present in the Y class despite the X characteristic in the data collection. (Deng et al. 2015)

Using Bayes' theorem, the Naive Bayes (NB) method is a probabilistic model that only needs a little amount of training data to estimate the statistical parameters (like mean and variance) required for classification. The degree of independence between data points is taken into account by NB. Numerous classification issues (such as text analysis, document classification, signal segmentation, natural disasters, and medical diagnostics) are solved using the NB classifiers. This classifier is a popular ML approach because it is straightforward to build, which is a significant benefit of NB over the other ML methods for classification (Demir and Şahin, 2022).

Bayes' theorem serves as the foundation for naive Bayes categorization. J classes, $j \in 1, \dots, J$, are each represented by a discrete variable called y_j . The expression for feature X is the m -feature feature vector $X=(x_1, x_2, \dots, x_m)$. Equation 5 (Deng et al. 2015) provides the posterior probability $p(y_j | X)$ for the hypothetical y_j value.

$$P(y_j \vee X) = \frac{P(X \vee y_j) P(y_j)}{P(X)} \quad (5)$$

The fundamental equation in Equation 6 for the Naive Bayes classifier is created when the product of the conditional probabilities for all characteristics is determined in the estimate of the target class Y' (Deng, et al. 2015).

$$Y' \leftarrow \underset{j}{\operatorname{argmax}} y_j \frac{P(y_j) \prod_{i=1}^m P(X=x_i \vee y_j)}{\sum_{j=1}^J P(y_j) \prod_{i=1}^m P(X=x_i \vee y_j m)} \quad (6)$$

Since the Naive Bayes classification technique used for estimating the value of the target class Y' does not depend on a single class, the formula in Equation 6 can be eliminated to become Equation 7, and the value of the denominator is used as a common denominator throughout the process (Deng et al. 2015).

$$Y' \leftarrow \underset{j}{\operatorname{argmax}} y_j P(Y = y_j) \prod_{i=1}^m P(X = x_i \vee Y = y_j) \quad (7)$$

$$P(y_j) = \frac{N_j}{N} \quad (8)$$

Equation 8 states that $|N|$ is the total number of educational credentials in the class label, and $|N_j|$ is the total number of educational credentials in class y_j (Bai et al., 2023).

The target class is chosen after applying Equation 7 to each class and determining the probabilities for each, according to the result values.

The Naive Bayes (NB) Classifier is one of the Bayesian theory-based probability-based classification algorithms. Using conditional probabilities, the Bayes Network is a type of visual model that depicts potential links between occurrences in Figure 10. Variables are depicted as nodes in the graphical model, while potential links between nodes are indicated as lines. If a node is unconnected, it is conditionally independent of other variables (Bai et al., 2023) and is hence conditionally independent.

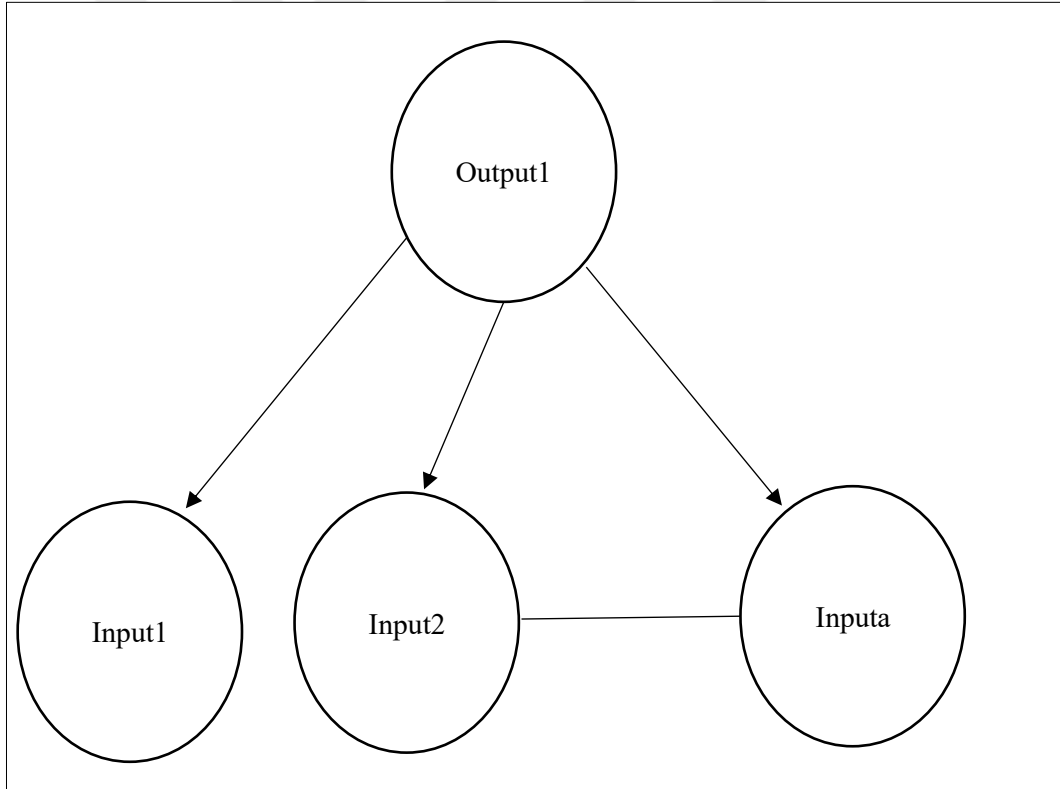


Figure 10. Naive Bayes Structure

Kaynak: Haltaş ve Alkan, 2013.

The probability of the second event happening in the case of the first event, given two independent random occurrences, is denoted by the equation $P(AB)$. The following two equations can alternatively be used to represent this statement. (Equation 1)

$$P(A \cap B) = P(A|B) P(B) = P(B|A) P(A) \quad (1)$$

The conditional probability of Bayes' theorem for two variables A and B is defined by Equation 2.

$$P(A \vee B) = \frac{P(B \vee A)P(A)}{P(B)} \quad (2)$$

In Formula 2, P(A) is the problem's intake chance, P(B) is the possibility of an event occurring as a result, P(A|B) is the chance that B will happen when A happens, while P(B|A) is the likelihood that A is going to appear when B happens.

Here, X input values are expressed as consisting of n features $X = \{X_1, X_2, \dots, X_n\}$ and the common probability of the variables in the node is calculated as in equation 3 with the chain rule.

$$P(X \vee YJ) = \prod_{i=1}^n P(X_i \vee YJ) \quad (3)$$

The Yj function in Equation 3 represents a node's immediate surroundings, while n is the Naive Bayes structure's total number of nodes. Due to the equality of the denominators in Equation 3, categorization is done by comparing the numerator values. The biggest result is selected throughout the categorization procedure. Equation 4 below shows how the Naive Bayes formula is computed in this situation.

$$Y' = \text{argmax} \quad (4)$$

The scikit-learn package for Python contains a classification function called Gaussian Naive Bayes. It makes predictions by mixing probabilities and presupposes that features within classes are distributed normally.

GaussianNB implements the Gaussian Naive Bayes algorithm for classification. The likelihood of the features is assumed to be Gaussian:

$$P(x_i | y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp \left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2} \right)$$

The parameters σ_y and μ_y are estimated using maximum likelihood.

2.4.3 Random Forest

In contrast to conventional trees used for classification and regression, the Random Forest (RF) technique allows for the creation of the necessary number of trees throughout the model's distribution creation process, as well as the expansion of these trees' value for prediction at every node in the tree. There is a lengthy process involved in carrying out such tasks on the RStudio application. Since many distinct tasks must be completed to complete the operations, from data preparation and transmission to the program to the phase where the map is created, to the stage where complicated instructions have to be mixed and written (Özdemir, 2018).

Let's start by saying a warning. The phrase "random forests" is a little vague. Nevertheless, no matter how the trees are produced, it is only a general term for some writers to aggregate random decision trees. Others refer to the original method developed by Breiman (2001). In this poll, we primarily adhere to the 2nd point of view.

As was already established, the forest method is adaptable enough to handle problems requiring both supervised regression and classification. making matters simple, we simply briefly discuss the classification case in this beginning and instead concentrate on regression analysis. Our goal in this part is to provide the technique for creating a random forest in a clear, theoretically accurate manner. In the basic framework of nonlinear regression calculation, an arbitrary intake vector $X \in \mathbb{R}^p$ is seen. The goal is to anticipate the square-compatible randomized response $Y \in \mathbb{R}$ by estimating the function of the regression $m(x) = E[Y | x = x]$. To do this, we suppose that we've got the use of a training set $D_n = ((X_1, Y_1), \dots, (X_n, Y_n))$ of different random variables that are distributed as the solitary prototypical pairing (X, Y) . An estimate of the variable m called $m_n: \mathbb{R}^p \rightarrow \mathbb{R}$ is to be built using the data set D_n . Thus, if $E[m_n(X) - m(X)]^2 \rightarrow 0$ as $n \rightarrow \infty$ (the expectancy is evaluated over X and the sample size D_n) then the regression equation's approximation m_n is (mean squared error) consistent.

M -generated trees of regression make compose an indicator known as a random forest. The notation $m_n(x; j, D_n)$ represents the expected value at the requested location x for the j th tree in the family, where $1, \dots, M$ is independent of the variables that are scattered similar to a general random number that is not influenced by D_n . Later on, more detailed definitions will be offered. In reality, the parameter is utilized to decide the direction of splitting as distinct trees. In mathematical terms, the j -th tree estimate takes the form.

$$m_n(x; \Theta_j, D_n) = \frac{1}{N_n(x; \Theta_j, D_n)} \sum_{i \in D^*_n(\Theta_j)} 1_{X_i \in A_n(x; \Theta_j, D_n)} Y_i$$

where $D^*_n(\Theta_j)$ is the set of data points selected before the tree construction, $A_n(x; \Theta_j, D_n)$ is the cell containing x , and $N_n(x; \Theta_j, D_n)$ is the number of (preselected) points that fall into $A_n(x; \Theta_j, D_n)$.

At this stage, we note that the trees are combined to form the (finite) forest estimate

$$m_{M,n}(x; \Theta_1, \dots, \Theta_M, D_n) = \frac{1}{M} \sum_{j=1}^M m_n(x; \Theta_j, D_n). \quad (1)$$

The R package randomForest's $ntree = 500$ default value for M (the forest's tree count) is the standard value. The (infinite) forest estimate and letting M reach infinite making sense from a modelling standpoint because M may be selected to be any size (restricted only by existing computer resources)

$$m_{\infty,n}(x; D_n) = E_{\Theta} [m_n(x; \Theta, D_n)].$$

According to this formulation, E stands for what is expected about the randomized variable, subject to D_n . The rule of big numbers, which states that virtually certainly, conditional on D_n ,

$$\lim_{M \rightarrow \infty} m_{M,n}(x; \Theta_1, \dots, \Theta_M, D_n) = m_{\infty,n}(x; D_n),$$

justifies the operation " $M \rightarrow \infty$." For further details on this limit calculation, see, for example, Breiman, 2001, and Scornet, 2015. The next section will use $m_{\infty,n}(x)$ rather than $m_{\infty,n}(x; D_n)$ to simplify the notation.

We now discuss the construction of the individual trees and the introduction of randomization. Every node of one tree is connected to a hyperrectangular cell in Breiman's (2001) initial forests. The tree's root is X itself, and at each stage of its development, a node (or its matching cell) is divided in half. A partition of X is made out of the terminal nodes (or leaves), when combined.

The procedure creates a set of M distinct (random) trees. Before each tree is constructed, a randomly selected observation from the initial data set is used with (or without) replacement. These observations—and just these observations—are taken into account during the tree building procedure (perhaps with repetitions). Then, each tree's cells are divided by maximizing the CART criterion across m_{try} routes that are uniformly distributed at random from the p original ones. The resulting subgroup of selected locations is known as M_{try} .

Each of the trees also stop growing when the total amount of points within each cell is less than the size of a node. For each query point $x \in X$ where the linked X_i falls inside the cell

of x , each regression tree predicts the average of the Y_i (that were amongst the points). We want to draw attention to the fact that the tree may form and the final assessment can be performed with just a small number of well picked points of data. Algorithm 1 gives detailed instructions on how to compute a forest's prediction.

Although Algorithm 1 may appear a little challenging at first, its fundamental concepts are straightforward. We begin by observing that this method has three crucial inputs:

1. $a_n \in \{1, \dots, n\}$: the number of sampled data points in each tree;
2. $m_{try} \in \{1, \dots, p\}$: the number of possible directions for splitting at each node of each tree;
3. node size $\in \{1, \dots, a_n\}$: the number of examples in each cell below which the cell is not split.

The parameters m_{try} , a_n , and node size are set to $dp/3e$, n , and 5, respectively, by default in the method of regression of the randomForest R package. de is the ceiling constant. In the next part, we will go into great detail on how these three characteristics affect the method's accuracy.

There is much more to be said about the CART-split criteria in Table 1. In order to render matters easier to comprehend, we are going to assume for the duration being that the tree is built using the entire and distinctive collection of data D_n without any subsampling. A additionally serves as a placeholder cell. $N_n(A)$ displays the overall number of points of data that fit into the category A . A cut in A is represented by the pair (j, z) , where j is one of the values (dimensions) from $\{1, \dots, p\}$, and z stands for the cut's beginning position along the j -th coordinate that exists, inside of A . Make CA a database of all technically possible cuts of this kind in A .

Table 1. Breiman's Random Forest Predicted Value at X

Algorithm 1: Breiman's random forest predicted value at x .	
Input	Training set D_n , number of trees $M > 0$, $a_n \in \{1, \dots, n\}$, $m_{try} \in \{1, \dots, p\}$, $nodesize \in \{1, \dots, a_n\}$, and $x \in X$.
Output	Prediction of the random forest at x .
1 for	$j = 1, \dots, M$ do
2	Select a_n points, with (or without) replacement, uniformly in D_n . In the following steps, only these a_n observations are used.

```

3      Set P = (X ) the list containing the cell associated with
4      the root of the tree.
5      Set Pfinal = ∅ an empty list.
6      While P ≠ ∅ do
7          Let A be the first element of P.
8          if A contains less than nodesize points or if all Xi ∈
9          A are equal then
10             Then
11                 Remove the cell A from the list P.
12                 Pfinal ← Concatenate(Pfinal, A).
13             else
14                 Select uniformly, without replacement, a subset
15                 Mtry ⊂ {1, . . . , p} of cardinality mtry
16                 Select the best split in A by optimizing the CART-split
17                 criterion along the coordinates in Mtry (see text for
18                 details).
19                 Cut the cell A according to the best split. Call AL and
20                 AR the two resulting cells.
21                 Remove the cell A from the list P.
22                 P ← Concatenate(P, AL, AR).
23             end
24         end
25     end
26
27     Compute the predicted value mn(x; Θj , Dn) at x equal
28     to the average of the Yi falling in the cell of x in partition
29     Pfinal.
30
31 end
32
33 Compute the random forest estimate mM,n(x; Θ1, . . . , ΘM, Dn) at the query point
34 x

```

Kaynak: Biau and Scornet, 2015.

$X_i = (X_{(1) i}, \dots, X_{(p) i})$, for any $(j, z) \in \mathcal{C}_A$, the CART-split criterion takes the form

$$L_{reg, n}(j, z) = \frac{1}{Nn(A)} \sum_{i=1}^n (Y_i - \bar{Y}^A)^2 1_{X_i \in A}$$

$$\frac{-1}{Nn(A)} \sum_{i=1}^n Y_i - Y_{AL} 1_{X_j^z} - Y_{AR} 1_{X_i^{(j)} > z}^2 1_{X \in A} \quad (2)$$

where $AL = \{x \in A : x_{(j)} < z\}$, $AR = \{x \in A : x_{(j)} \geq z\}$, and $\bar{Y}^A$ (resp., $\bar{Y}^{AL}$, $\bar{Y}^{AR}$) is the average of the Y_i belonging to A (resp., AL , AR), with the convention that the average

is equal to 0 when no point X_i belongs to A (resp., AL , AR). For each cell A , the best cut (j^*n, z^*n) is selected by maximizing $L_{reg,n}(j, z)$ over M_{try} and CA ; that is

$$(j^*n, z^*n) \in \begin{cases} \text{argmax} \\ j \in M_{try} \\ (j, z) \in CA \end{cases} L_{reg,n}(j, z)$$

(To break up certain ties in the argmax, the optimal cut is always made during two data points.) Let's conclude by noting that the aforementioned optimization technique applies to the reprocessing scenario with ease by optimizing over a selected group of observations rather than the initial data set D_n .

The method evenly chooses random m_{try} places in every single cell of every tree and presents the best one in order to evaluate criteria (2) overall viable cuts along the directions in M_{try} . The standard measurement (2) is used as a criterion in Breiman et al.'s (1984) influential CART approach. Using this criterion, it is possible to compute the (renormalized) variance compared to the real fluctuation in the the node's value prior to and following a cut. There are three key differences between CART and a tree in Breiman's forest. Breiman (2002)

Instead of using the whole range $1, \dots, p$, the condition (2) in Breiman's forests is first evaluated over a subset M_{try} of picked at random locations. The final cells only include the number of node size assessments, and none of the trees are pruned (unless each information point in the particular cell has identical X_i). At the very least, instead of utilizing the complete sample D_n , each tree is constructed using a selected group of cases picked from the initial samples; the estimated value is only based on these findings. The procedure runs in bootstrapping mode when $a = n$ (and replace sampling is used); when an n , sub-samples (with or without replacement) is utilized.

2.5. Evaluation Performance Measures

This section contains data on the Confusion Matrix, Accuracy, Precision, Recall, F1 Score, and AUC-ROC Curve. It has also involved research on machine learning methods and their efficiency evaluation.

2.5.1. Confusion Matrix

The confusion matrix describes the outcome of categorization issues. It indicates the accuracy of the model predictions. The amount of right and wrong predictions can be categorized by calculating each outcome and categorizing them. To put it another way, the confusion matrix reveals how confused the model is while making predictions. The matrix provides information about the model's flaws along with their categories based on the values that occur. Table 2. shows the general format of a confusion matrix for a two-class problem.

Table 2. Confusion Matrix Template for a Two-Class Problem

ACTUAL CLASS	PREDICTED CLASS	
	True	False
True	TP	FN
False	FP	TN

The confusion matrix is a 2x2 matrix, as shown in Table 3.1, with rows for the true class and bars for the anticipated category. True Positive, False Negative, False Positive, and True Negative are acronyms for True Positive, False Positive, and True Negative, correspondingly. The following elements in the confusion matrix are clarified as follows:

- True Negative (TN): The number of parameters estimated by the equation to be negative that are really negative.
- True Positive (TP): The number of variables that were estimated by the model to be Positive and found to be Positive.
- False Negative (FN): The number of factors estimated by the model appears to be negative but is positive.
- False Positive (FP): The number of parameters estimated by the model appears to be positive but turns out to be negative.

2.5.2. Accuracy

Precision is an innate performance metric. It indicates how near the outcomes are to the genuine value. As shown in Equation 1, accuracy is indicated as the number of right estimates divided by the entire amount of guesses.

$$Accuracy = \frac{TN+TP}{TN+TP+FN+FP} \quad (1)$$

The correctness issue is that it fails to offer all of the information about the data. To illustrate, consider COVID-19 cases: the accuracy of persons in the community who test negative for the virus can be 99%, but this does not ensure security, and the one percent that remains does not need to be handled. The high-accuracy findings do not provide all of the information required to analyze data. As a result, more assessments and evaluations using indicators are required.

2.5.3. Precision

Precision is the level of repeatability with which a procedure can be repeated. The model's accuracy is the number of times it successfully detects a fractured picture and a non-cracked picture. According to Equation 2, it is the proportion of genuine positivity compared to the total expected positivity from the confusion matrix.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

2.5.4. Recall

The level of responsiveness is referred to as recall. The recall is the number of positive examples that a model catches by identifying them as such. As seen in Equation 3, this is represented as the percentage of really recognized positive cases per real outcome.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

2.5.5. F1 Score

F1-result measures the level of reliability using Recall and Precision consideration, providing greater importance to false positives and false negatives while preventing a

significant number of genuine negatives from influencing the result. The F1-Score computation is detailed in Equation 4.

$$F1 = 2x \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

2.5.6. AUC-ROC Curve

In the research we conducted, an essential sort of metric was the area under the ROC curve (AUC), which we employed in both the undersampled data and the whole dataset. AUC outperforms correctness in assessing algorithms for learning [103], with AUC assessed on positive category (fraudulent) FP and TP. The 'Area Under the ROC Curve' (AUC) is always used to contrast. The region underneath the curve (AUC) is determined by the formulae below:

$$AUC = \frac{u_1}{n_1 n_2} \dots \quad (5)$$

Where (U1) signifies the (U) value calculated with sample 1, (n1) indicates the quantity of example 1, and (n2) indicates the quantity of sample 2.

2.5.7. A Comparative Analysis of Performance Measures: A Review of Literature

Hussain et al. (2019) give an evaluation of a range of machine learning categorization approaches for recognizing activities in their research. To evaluate the efficacy of five machine learning algorithms for classification, a previous hierarchical activity identification system was modified. Numerous studies conducted and discovered that classification techniques surpass classic activity recognition algorithms substantially. The work's objective is to promote independence amongst the older society specifically individuals with Alzheimer's disease.

In the study we conducted, Khanam and Foo (2021) combine data mining methods, machine learning (ML) techniques, and Neural Network (NN) methodologies are applied to anticipate diabetes. For this study, we utilized the Pima Indian Diabetes (PID) data from the UCI Machine Learning Repository. The collection includes information on 768 patients and their nine distinct characteristics. On the data set, we utilized seven ML algorithms for

estimating diabetes. We discovered that the model based on Logistic Regression (LR) and Support Vector Machine (SVM) predicts diabetes effectively. We created the NN model with each hidden layer having a distinct epoch and found that the NN with two hidden layers delivered 88.6% accuracy.

Each year, crashes involving vehicles cause countless casualties, injuries, and deaths as well as financial losses, making it one of the world's most pressing challenges. Effective algorithms to anticipate the severity of traffic accidents, according to Al et al. (2019), is a key job for transportation networks. This inquiry attempt provides methods for selecting a collection of significant criteria and developing a model for categorizing the severity of accidents. Several machine learning is used to create these models. On traffic accident info, machine learning techniques with supervision such as AdaBoost, Logistic Regression (LR), Naive Bayes (NB), and Random Forests (RF) are used. The SMOTE method is used to deal with data imbalances. The results of the research suggest that the RF model might be a useful tool for forecasting the extent of traffic injuries. The RF method outperformed the LR algorithm (74.5% accuracy), the NB algorithm (73.1% accuracy), and the AdaBoost algorithm (74.5% accuracy).

Dropout is a severe problem in countries with limited resources like India, especially at the secondary level, and it is caused by a variety of socioeconomic and educational variables. Algorithms based on machine learning have been applied to learning for predicting student dropout so that probable dropouts may be identified, and corrective steps can be implemented to avoid the dropout. Mirza and Hassan (2020) assess the primary variables that cause school dropout and give a comparison of five supervised classification methods used for predicting school dropout pupils. Such algorithms can be utilized to rapidly identify and predict dropouts or students in danger of leaving schools, allowing corrective strategies to be deployed to decrease the possibility of dropout or preventing students from actually dropping out.

Stress has an impact on mental as well as physical wellness, and as a result, stress is one of the leading causes of pregnancy issues such as hypertension. As a result, it is vital to determine the impact of stress on the health of both the mother and the baby to identify potential difficulties while the course of pregnancy and delivery. It may also be effective in predicting and avoiding birth malformations, as there have been several cases of stress-related difficulties resulting in cognitive impairments in children. Madhusri et al. (2023)

used social, physical, environmental, and biological components to construct a model for the prediction of stress-related problems during pregnancy. The data set was created by administering a customized interview-based assessment to women who had previously undergone pregnancies and deliveries. The inquiries depended on the previously indicated variables. The produced data is utilized to validate the hypothesis and assess the performance of the suggested stress prediction framework utilizing machine learning methods such as Support Vector Machine (SVM), Naive Bayes (NB), KNearest Neighbor (KNN), and Decision Tree (DT). When the Naive Bayes approach was utilized, the experimental findings showed that the suggested model attained an accuracy of 90%. The other methods yielded less accurate but still near findings.

Gebreyohans and Gandhi (2018) divide children into four categories: orphans, solitary abandoned children, vulnerable kids, and protected children. The objective of this categorization is to assist the Love for Children Organization's outside supporters and to obtain complete information about each kid for internal purposes. To do this, three types of categorization approaches were applied within the context of the KDD (Knowledge Discovery in Databases) data mining model: decision tree, Bayesian learning, and neural network. For testing with the categorization method, the children's datasets were gathered, cleaned, processed, and merged. The final dataset contains 17044 recordings that have been examined and compared to their effectiveness. The dataset gathering was tested using a ten-fold cross-validation procedure and divided into 70/30%, 66/44%, and 50/50% for both testing and training, correspondingly. Furthermore, the total accuracy of classification and advantages of decision tree (98.83%), bayesian learning (98.32%), and neural network (98.86%) models were compared. The study finds that a decision tree (98.83%) ought to be used as a framework since it outperforms Bayesian learning (98.32%) and outperforms Neural Networks (98.86%) in the categorization of organizational children.

S and Dornadula (2019) The primary goal of this article is to build and create a unique fraudulent activity detection algorithm for Streaming Transaction Data, to analyse customers' prior purchase information and extract patterns of behaviour. Cardholders are divided into categories according to the volume of their transactions. Then, utilizing a sliding window method, combine the transactions done by cards from various categories to extract the behavioural patterns of the groupings. They used European credit card theft datasets in their article.

CHAPTER 3 – RESULTS

3.1. Purpose of Application

In practice, the Child Labor Force Survey data, which is not a wide application area in data mining with using "classification" methods of Turkey Statistics Institute (TURKSTAT) in 2019 were obtained from this database. The purpose of this study is to demonstrate how logistic regression, Naive Bayes, Random Forest, and machine learning classification models may all be used to better understand, recognize, and forecast patterns and characteristics associated with child labor. The development of precise, equitable, and efficient methods for diagnosis and prevention is made easier by comparing machine learning classification models on child labor. In order to effectively eliminate child labor, it helps to understand the underlying causes, minimize prejudices, guide policy choices, and encourage cooperation. In the continuation of this study, logistic regression, which is one of the classification tree methods, was used first, while the naïve bayes and random forest algorithm were used in the other stages.

3.2. Child Labor in Turkey and Literature Review

Child labor definitions vary across countries due to cultural factors. The International Labor Organization (ILO, 2021) defines child labor as work that poses risks and harm to the physical, mental, or moral well-being of children, and that interferes with their education. As per the labor legislation in Turkey, children who have finished their primary education and are no longer enrolled in school are permitted to work for up to 7 hours per day or 35 hours per week, with the possibility of extending to 40 hours per week. However, children who are currently attending school are only allowed to work for a maximum of 2 hours per day or 10 hours per week. During school closures, their work hours may be the same as those of children who are not attending school (Koçali, 2022).

In Turkey, despite efforts to combat it, child labor persists as a complex issue with socio-economic and cultural roots. According to the International Labor Organization (ILO, 2022), “A total of 160 million children – 63 million girls and 97 million boys – are in child labor globally, accounting for almost one in ten of all children worldwide. Nearly half of

all those in child labor – 79 million children in absolute terms – are in hazardous work that directly endangers their health, safety, and moral development.” Based on the results of the Turkish Household Labor Force Survey (THLFS) 2020 survey (TURKSTAT, 2021), it was found that the labor force participation rate for children aged 15-17 in Turkey was 16.2%, which corresponds to approximately 600,000 individuals. Among these children, boys had a labor force participation rate of 23.4%, while girls had a rate of 8.6%. The total number of children engaged in economic activity in Turkey was 720,000, with boys comprising 70% and girls 30% of this population. Among the working children status, the majority (80%) were aged 15-17, with 16% aged 12-14 and 4% between 5-11 years old. The primary sectors of employment for these children were agriculture (31%), manufacturing (24%), and services (45%). However, the sector distribution varied based on age, with 61% of children aged 5-14 working in agriculture and 51% of those aged 15-17 working in services. Despite their economic activity, a significant proportion of working children (66%) also attended school, with a higher attendance rate among females (66%) compared to males (57%). The main reasons for children engaging in economic activity included assisting with household economic activity (36%), seeking vocational training and occupational skills (35%), contributing to household income (23%), and meeting personal needs (6%). The International Labour Organization (ILO) reported a decline in the number of children engaged in economic activity in Turkey over the years. In 1994, approximately 2.3 million children were involved in economic activities, accounting for around 15% of the total child population. This number decreased to nearly 1 million (6%) in 2006 and further dropped to 720,000 children (4.4%) in 2019. The period between 1994 and 2006 witnessed a significant decrease in the number of children engaged in economic activity.

Child labor in Turkey has been extensively studied in the literature, covering various aspects. Gerşil (2017) has examined the occurrence of child labor in Turkey in different sectors, including agriculture, manufacturing, and informal urban settings, and have identified different forms of child labor, such as hazardous work, domestic labor, and street work. The agricultural sector is one of the worst in the country for seasonal migrant and agricultural labor in the workplace and in the environment, where occupational diseases and accidents are most prevalent (Koçali, 2022). The literature highlights several contributing factors, such as poverty, limited access to education, weak enforcement of labor laws, migration, and the presence of an informal economy. Numerous studies conducted globally and within Turkey emphasize that poverty is widely recognized as the

primary underlying factor leading to child labor when examining its causes (Gürses, 2009; Kahraman & Sallan Gül, 2015; Öngel, 2017).

3.3. Dataset Definition

The data used in practice, Turkey Statistics Institute (TURKSTAT) in 2019 "Working Child Statistics" were obtained from the micro database. The joint implementation of the Working Children Survey (WCS) and the Turkish Household Labor Force Survey (THLFS) took place in the year 2019. The HLFS involved interviewing 14,277 households, and data from 25,190 children was obtained as a result of this study. The definition of child labor for the purposes of this study according to TURKSTAT is children between the ages of 5 and 17. In this thesis, the application of a machine learning classification algorithm was implemented using the dataset of 7,934 children aged between 12 and 17 years. In this binary dependent variable (Working Children Status: WCS) given in the classifier scale, W: if the child works (1); NW: Coded as if child does not work (0). While the number of not working children was 6,860 (NW = 86%), the number of those who work was found to be 1,072 (W = 14%).

Table 3. The Model with Measurement Items

Measurement items	
Variable	Description
Dependent Variables	
Working Children Status (WCS)	1 = If the child works 0 = If the child does not work
Continuous Independent Variables	
Household Size (HHS)	Household Size (Min:2, Max:17)
Family composition (FACOM)	The proportion between the count of adults aged 18 or above and the count of children below 18 years old within the family
Categorical Independent Variables	
Maleness (MALE)	1 = If the child is male 0 = If the child is female
Age Group (AG)	1 = 15-17 age group 0 = 12-14 age group

Mother Age Group (MAG)	1 = 25-34 age group 2 = 35-54 age group 0 = 55 + age group
Father Age Group (FAG)	1 = 25-34 age group 2 = 35-54 age group 0 = 55 + age group
The Status of Attending School (SAS)	1 = Attending School 0 = Not Attending School
Educational Status of the Mother (ESM)	1 = Illiterate 2 = Less than high school 3 = High school 4 = Vocational or technical high school 5 = Higher education graduate
Educational Status of the Father (ESF)	1 = Illiterate 2 = Less than high school 3 = High school 4 = Vocational or technical high school 5 = Higher education graduate
Sector of the household Leader (SHL)	1 = Agriculture 0 = Not-Agriculture
Homework 1 (HW1) (Shopping for households, laundry-dishwashing, ironing, cooking, cleaning of household items etc.)	1 = Yes 0 = No
Homework 2 (HW2) (Caring for younger children in the household)	1 = Yes 0 = No
Working Status of the Mother (WSM)	1 = Employed 0 = Unemployed
Working Status of the Father (WSF)	1 = Employed 0 = Unemployed

Working Children Status is a binary variable and is considered as the dependent variable given in Table 3. As seen in Table 3., the model includes 14 independent variables, 2 of which are continuous variables and 12 are categorical variables. These continuous variables

are household size and family composition. According to Libaque-Saenz et al., (2018), the neural network model shows that geographical indicators, income levels, gender, family composition and educational levels significantly predict child labor. Family composition is computed and included in the model since it is an important variable in predicting child labor. The categorical variables are; gender, age group, mother age group, father age group, the status of attending school, educational status of mother, educational status of father, sector of household leader, homework 1(Shopping for households, laundry-dishwashing, ironing, cooking, cleaning of household items etc.), homework 2 (Caring for younger children in the household), working status of the mother and working status of the father.

3.4. Data Analysis

The present section of the analysis focuses on investigating child labor in Turkey using descriptive statistics, logistic regression, naive Bayes, and random forest models. These algorithms' abilities to comprehend and forecast patterns of child labor are also evaluated through a performance comparison.

3.4.1 Descriptive Statistics

The initial step in data analysis involves organizing and defining the data. Considering that certain variables associated with Working Children Status (WCS) exhibit similarities and correlations, the number of variables was reduced by examining the correlation matrix. Based on this matrix, it was determined to include categorical variables such as “gender”, “age group”, “mother age group”, “father age group”, “the status of attending school”, “educational status of mother”, “educational status of father”, “sector of household leader”, “homework 1”, “homework 2”, “working status of the mother” and “working status of the father “ as well as dependent variable for "Working Children Status" classifier. The significance coefficients were set at alpha levels of 0.10 and 0.05. Python-based software Jupyter Notebook was used to analyze the data.

The results in Table 4. indicates a statistically significant difference in the distributions of the continuous variable across the categories ($p < 0.05$). Therefore, we reject the null hypothesis and conclude that there is evidence of a significant difference in the distributions of the continuous variable among the categories. These findings contribute to our understanding of factors influencing child labor and have implications for formulating policies to combat child labor.

Table 4. Mann-Whitney U test p-values for FACOM variables

Variables	U-statistics	P-value
WCS*FACOM	62881036.5	0.000
Degrees of Freedom: 15866, Independent Sample T-test: 240.08, Z-test Statistics:36.56		

Given that the variables in our study do not exhibit a normal distribution except for Family Composition (FACOM), and are categorical in nature, we utilized the Cramer's Vu correlation coefficient in Figure 11 as a more suitable measure. Recognizing the significance of the statistical properties of the dataset in both data analysis and pre-processing stages, we present the chi-square p-values in Table 5. This table displays the frequencies of the selected categorical variables and their associations with the Working Children Status (WCS) variable.

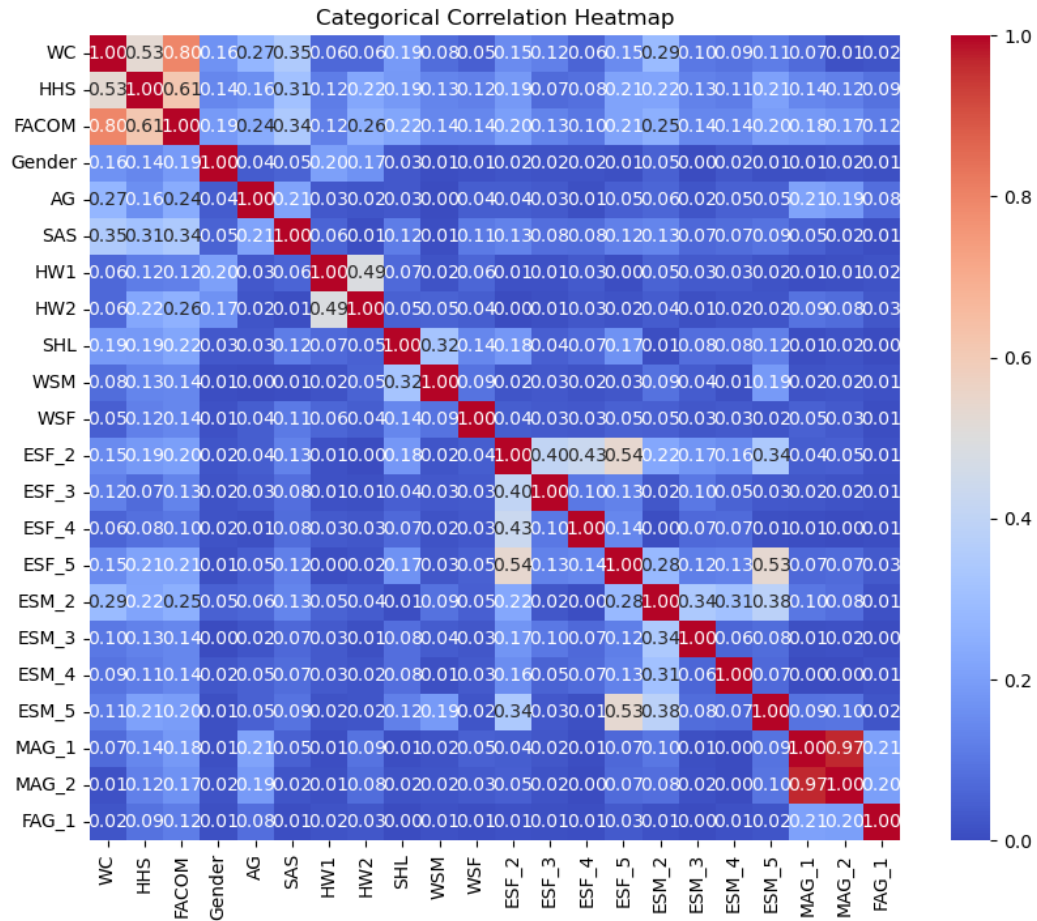


Figure 11. Cramer's V correlation Heat Map

Table 5. presents a comprehensive analysis of categorical independent variables, showcasing frequencies, percentages, and chi-square p-values. Moreover, it assesses the connection between these independent variables and Working Children Status (WCS)

through chi-square testing. The table provides a clear representation of the statistical information, enabling a thorough understanding of the relationship between the variables.

Table 5. Frequencies, Percentages and Chi-Square p-values For Independent Variables

	Working (W) n (%)	Not Working (NW) n (%)	Total n (%)	P-values
Gender				
1: Male	779	3332	4111(51.8)	0.000
0: Woman	294	3529	3823(48.2)	
Age Group (AG)				
1: 15-17	902	3025	3927(49.5)	0.000
0: 12-14	171	3836	4007(50.5)	
HHS				
5=<: HHS	666	4785	5451(68.8)	0.000
6=>: HHS	407	2076	2483(31.2)	
Mother Age Group (MAG)				
1: 25-34	71	929	1000(12.6)	0.013*
2: 35-54	943	5932	6875(86.6)	0.000
0: 55 +	59	0	59(0.8)	
Father Age Group (FAG)				
1: 25-34	9	120	129(1.62)	0.000
2: 35-54	950	6741	7691(96.9)	
0: 55 +	114	0	114(1.48)	
The Status of Attending School (SAS)				
1: Attending School	669	6432	7101(89.5)	0.000
0: Not Attending School	404	429	833(10.5)	
Educational Status of the Mother (ESM)				
1: Illiterate	769	585	1354(17.0)	0.000
2: Less than high school	304	4727	5031(63.4)	
3: High school	0	507	507(6.3)	
4: Vocational or technical high school	0	419	419(5.6)	
5: Higher education graduate	0	623	623(7.7)	
Educational Status of the Father (ESF)				
1: Illiterate	129	169	298(4.0)	0.000
2: Less than high school	873	4124	4997(62.9)	
3: High school	0	692	692(8.7)	
4: Vocational or technical high school	59	731	790(9.9)	
5: Higher education graduate	12	1145	1157(14.5)	0.007*
Sector of the household Leader (SHL)				

1: Agriculture	356	896	1252(15.8)	0.000
0: Not-Agriculture	717	5965	6682(84.2)	
Homework 1 (HW1)				
1: Yes	523	3975	4498(56.6)	0.000
0: No	550	2886	3436(43.4)	
Homework 2 (HW2)				
1: Yes	236	2100	2336(29.5)	0.000
0: No	837	4761	5598(70.5)	
Working status of the Mother (WSM)				
1: Employed	531	2641	3172(40.0)	0.000
0: Unemployed	542	4220	4762(60.0)	
Working status of the Father (WSF)				
1: Employed	932	6232	7164(90.2)	0.000
0: Unemployed	141	629	770(9.8)	

The statistical relationships between the variables were examined using the chi-square test. The chi-square test hypotheses for each relationship are provided below, indicating the proposed statements that were tested for statistical significance.

The chi-square test was conducted to examine the statistical independence between variables. The hypotheses for each relationship are as follows:

H0: Status of Attending School and Working Children Status are statistically independent.

H1: Status of Attending School and Working Children Status are not statistically independent.

The chi-square test revealed a significant statistical relationship between Status of Attending School and Working Children Status at a 95% confidence level ($p\text{-value} = 0.000$, which is less than the significance level $\alpha = 0.01$).

Similarly, for the other variables, the hypotheses and results are as follows:

H0: Sector of Household Leader and Working Children Status are statistically independent.

H1: Sector of Household Leader and Working Children Status are not statistically independent.

The chi-square test indicated a significant statistical relationship between Sector of Household Leader and Working Children Status at a 95% confidence level (p-value = 0.000).

In summary, according to the chi-square test, all variables exhibited a statistically significant relationship with Working Children Status at a 95% confidence level. This result confirms our main thesis about the variables that influence the likelihood of child labor. The chi-square analysis further highlights these determinants' statistical independence, underscoring their crucial role in determining the dynamics of child work.

3.4.2 Examination of Child Labor in Turkey with Logistic Regression Model

In the regression phase of supervised machine learning applied to working child statistics data, significant variables were determined using the backward logistic regression model. The results of this analysis, including the significant variables, are presented in Table 6 and in Table 7.

Table 6. Logistic Regression Model Statistics

Model Measure	Statistics
AIC	2500.9
Pseuda R2	0.61
Chi-Square Statistics	2514.8
Log-Likelihood (LL)	-1228.4
P value	0.0

The evaluation of child labor in Turkey is made possible by the statistical analysis of the logistic regression model, which is shown in Table 6. An AIC value of 2500.9 suggests a tolerable model fit, and the model shows a fair fit. The model accounts for around 61% of the variation in outcomes related to child labor, according to the pseudo R2 of 0.61. Further highlighting the statistical importance of the model's predictors in comprehending child labor is the Chi-Square score of 2514.8 and the low p-value (0.0). Despite a negative log-likelihood (-1228.4), which suggests some mismatch between predicted and observed values, these statistics demonstrate the model's applicability in examining the determinants causing child labor in the Turkish context.

It is necessary to convert categorical data with numerous categories into binary variables, sometimes referred to as dummy variables, in logistic regression. Each dummy variable corresponds to a particular category of the primary variable. For the "Mother Age Group (MAG)" variable in our example, we will make the following two dummy variables:

- MAG_25_34: This dummy variable takes the value 1 if the individual is in the "25-34" age group, and 0 otherwise.
- MAG_35_54: This dummy variable takes the value 1 if the individual is in the "35-54" age group, and 0 otherwise.

Now, based on an individual's age group, a logistic regression model may be developed to forecast their likelihood of falling under the category of "Working." The following is one way to represent the model:

$$\text{logit}(p) = \beta_0 + \beta_1 * \text{MAG_25_34} + \beta_2 * \text{AgeGroup_35_54}$$

Where:

- $\text{logit}(p)$ represents the natural logarithm of the odds of an individual being in the "Working Children Status" category.
- β_0 is the intercept term.
- β_1 is the coefficient associated with the "MAG_25_34" dummy variable.
- β_2 is the coefficient associated with the "MAG_35_54" dummy variable.

Interpretation of Coefficients:

- If β_1 is positive, it means that being in the "25-34" age group increases the log-odds of being in the "Working" category.
- If β_2 is positive, it means that being in the "35-54" age group increases the log-odds of being in the "Working" category.

Table 7. Logistic Reg. Model (80% Trainig Data): Wald Chi-Square p-values and OR

Variables	Coefficients	P-value	Odd Ratio (OR)
Gender	1.452594	0.0000	4.2742
Age Group (AG)	1.976648	0.0000	7.2185
Mother Age Group (MAG)			
1: 25-34	-0.122176	0.7969	0.8850
2: 35-54	0.096548	0.8221	1.1014
Father Age Group (FAG)			

1: 25-34	-2.766952	0.0000	0.0629
2: 35-54	-2.555942	0.0000	0.0776
The Status of Attending School (SAS)	-0.926122	0.0000	0.3961
Educational Status of the Mother (ESM)			
2: Less than high school	-2.825883	0.0000	0.0593
3: High school	-23.974611	0.9981	0.0000
4: Vocational or technical high school	-18.536824	0.9815	0.0000
5: Higher education graduate	-19.554207	0.9854	0.0000
Educational Status of the Father (ESF)			
2: Less than high school	-0.297466	0.1291	0.7427
3: High school	-21.0142	0.9961	0.0000
4: Vocational or technical high school	-0.346125	0.2069	0.7074
5: Higher education graduate	-0.981961	0.0150	0.3746
Sector of the Household Leader (SHL)	0.306906	0.0229	1.3592
Homework 1 (HW1)	0.224551	0.0487	1.2518
Homework 2 (HW2)	-0.18332	0.2062	0.8325
Working status of the Mother (WSM)	0.695768	0.0000	2.0052
Working status of the Father (WSF)	0.633951	0.0004	1.8850
Constant	0.11347	0.0003	1.1202

Reference: Male, 15-17, +55, +55, Attending School, Illiterate, Illiterate, Agriculture, Yes, Yes, Employed and Employed in the order of table above

In the logistic regression model (Table 7.), which was established in the form that 80% of the data would be training data and 20% test data, The chances of a child who attends school becoming a child laborer was 3.9% less than a child who does not attend school (OR = 0.39). In the situation of working status of mother and father, children whose mothers are working have a 20% and 18% less chance of becoming child laborers than those who are not working, respectively. In education status of mother and father, it was found that all categories starting from Less than high school (2) to Higher education graduate (5) has a lower incidence of child labor compared to those who did not attend school. It has been determined that fathers who have completed a higher education are approximately 4% less prone to having child labor in their households compared to those who do not receive any formal education. On the other hand, the chance of children who assist with household chores getting involved in child labor is approximately 13% higher than children who do

not help with household chores. In the factor of gender, the chances of being child worker are approximately 4.27 times higher for man than for woman. In the age group, it has been found that the chance of working child age between “15-17” compared to “12-14” was found to be approximately 72% higher.

The analysis of the logistic regression model sheds important sociological light on the complicated problem of child work. The odds ratios and coefficients show that having a male gender increases the likelihood of child labor, showing gender-based inequities. Additionally, when kids get older, their likelihood of working grows considerably, highlighting the effects of financial constraints or lack of educational opportunities. Higher parental education is associated with less child labor, highlighting the importance of education in breaking this cycle. The educational standing of parents also emerges as a critical factor. It is noteworthy that children of working parents, particularly working moms, are more prone to child labor as a result of economic demands. Additionally, the sector of the household head affects the prevalence of child labor, showing that some economic sectors are more vulnerable than others. The analysis emphasizes the crucial function of education in preventing child labor because it makes it less likely to occur when children are in school. Overall, these findings highlight the complex interplay between social injustices, gender roles, and economic factors that sustain child labor, highlighting the urgent need for all-encompassing interventions that place a priority on social welfare, gender equality, and education to address this issue.

The logistic regression model was evaluated using performance metrics such as Conflict matrix, accuracy, sensitivity, specificity, precision, F score, and AUC. The evaluation results, including the mentioned metrics, are presented in Table 8. and in Table 9. The probability cut-off point for the logistic regression model was set at 0.5.

Table 8. Conflict Matrix for Logistic Regression Model

		PREDICTED CLASS	
ACTUAL CLASS	W	W	NW
	NW	1335	37
		104	111

Table 9. Performance Evaluation for Logistic Regression Model

Algorithm	Accuracy	Sensitivity	Precision	F-score	AUC
Log Reg	0.91	0.57	0.71	0.63	0.74

Based on the provided Table 9., the accuracy value was reported as 91%, sensitivity as 57%, precision as 71%, F-Score as 63%, and AUC value as 74%.

3.4.3 Examination of Child Labor in Turkey with Naïve Bayes Model

The dataset is split into two parts: a training set and a testing set, with a ratio of 80% for training and 20% for testing. The Naive Bayes model is utilized to train on the training set, leveraging conditional probability principles. By analyzing the training data, the model establishes connections between predictor variables and the target variable, which in this case is child labor, by estimating probabilities. During the training process, the Naive Bayes assumption of feature independence is considered.

The dataset used contains both categorical and continuous variables. To convert categorical variables into numerical discrete variables, a technique called dummy variable encoding is typically employed. However, in this case, dummy variable encoding was not used in order to work with binary variables, as the Naive Bayes (NB) algorithm does not perform efficiently with this approach. Instead, models were created using only the target mean technique to handle continuous variables. As a result, GaussianNB was chosen as the algorithm for this purpose.

Table 10. presents the Naive Bayes model Statistics:

Table 10. Naïve Bayes Model Statistics

Model Measure	Statistics
AIC	54
Pseuda R2	0.98
Chi-Square Statistics	265.3
Log-Likelihood (LL)	-377.4
P value	0.0

Table 10. provides significant statistics demonstrating the Naive Bayes model's superior performance in the analysis of child work in Turkey. A balanced complexity-fit trade-off is

indicated by the low AIC value (54) of the model. The significant variance explanation is suggested by the high Pseudo R² (0.98), and the significant variable correlations are highlighted by the high Chi-Square value (265.3). Effective data fit is indicated by the negative Log-Likelihood (-377.4) value. The extremely low P value (0.0) denotes strong model stability and important results. Collectively, these figures show how well the Naive Bayes model captures and forecasts Turkey's dynamics of child labor.

To assess the model's performance, multiple evaluation metrics are computed using the testing set. These metrics encompass accuracy, sensitivity, precision, and F1 score. Additionally, a confusion matrix is constructed to provide a detailed breakdown of true positive, true negative, false positive, and false negative predictions.

Table 11. and Table 12. presents the performance evaluation metrics of the Naive Bayes model:

Table 11. Conflict Matrix for Naive Bayes model

ACTUAL CLASS	PREDICTED CLASS		
		W	NW
	W NW	1259 94	113 121

Table 12. Performance Evaluation for the Naive Bayes model

Metric	Value
Accuracy	0.86
Sensitivity	0.55
F1 Score	0.52
Precision	0.50
AUC	0.67

The evaluation of the Naive Bayes model reveals its impressive performance in accurately predicting instances of child labor. With an overall accuracy of 86%, the model demonstrates a strong ability to make correct predictions across the entire dataset. This indicates that approximately 86% of the instances are classified accurately as either child labor or non-child labor. The sensitivity of the model is determined to be 55%. This indicates that the model correctly identifies 55% of the positive cases or instances related to child labor in Turkey. Sensitivity is a measure of the model's ability to detect true

positives and is particularly important when the focus is on identifying the presence of child labor accurately. The model may not be able to capture every occurrence of child labor, as indicated by the sensitivity value of 55%, and there may be potential for improvement in terms of precisely identifying positive cases. In-depth analysis of the evaluation reveals that the model's accuracy is just 50%, meaning that only 50% of the occurrences it predicted as positive actually are positive in terms of child labor in Turkey. Precision serves as an essential measure in minimizing false positives, highlighting the accuracy of positive predictions. Simultaneously, the calculated F1 score of 52% represents a balanced evaluation of the model's performance, considering both precision and recall (sensitivity). This score, which represents a suitable trade-off between the two criteria, is the harmonic mean of precision and recall. The acquired F1 score indicates a moderate level of ability, but there is still opportunity for growth in both recall and precision. Increasing recall would assure the identification of a higher percentage of true positives while improving precision would reduce the number of false positives. It is conceivable to improve the model's overall performance and attain higher precision and recall rates by adjusting the model's parameters or investigating alternate modeling strategies. These results highlight the need for further development and improvement in the model's capability to precisely detect instances of child labor in the setting of Turkey. Additionally, the analysis reveals that the model achieves an Area Under Curve (AUC) value of 0.67. The AUC is a measure of the model's ability to discriminate between positive and negative instances, with a higher value indicating better discrimination. The AUC value of 0.67 suggests that the model demonstrates reasonable discriminative power in distinguishing between cases of child labor and non-child labor in Turkey. These performance metrics provide valuable insights into the model's performance in examining child labor instances. Despite the model's good accuracy, sensitivity and precision could be increased. A bigger percentage of genuine positives might be detected by the model with increased sensitivity, while fewer false positives could be detected with increased precision. It might be possible to improve the model's overall performance and get higher precision, recall, and AUC values by tweaking it or considering other strategies.

3.4.4 Examination of Child Labor in Turkey with Random Forest Model

A training set and a testing set are created from the dataset in an 80:20 split. Using an ensemble of 100 decision trees, the Random Forest model is trained on the training set. To

maintain diversity and reduce overfitting, each tree is built using a random selection of features and data samples.

Table 13. presents some Random Forest Model Statistics:

Table 13. Random Forest Model Statistics

Model Measure	Statistics
MSE	0.12
Pseuda R2	0.81

The results from Table 13. provide insight into the model's performance when used to analyze child labor in Turkey using a Random Forest model. The MSE score of 0.12 indicates that the model is capable of estimating child labor indicators with only modest prediction errors. Notably, the high Pseudo R-squared score of 0.81 demonstrates the model's remarkable capacity to reveal patterns within the data and demonstrates its aptitude for understanding the complex elements influencing trends in child labor in Turkey.

Several metrics are computed using the testing set to assess the performance of the model. To evaluate the model's predictive skills, the accuracy, sensitivity, precision, and F1 score are computed. The area under the ROC curve (AUC), which is another statistic, is used to determine how well a binary classification model can distinguish between positive and negative data. The most important elements influencing child labor in Turkey are also determined using a feature importance analysis.

Table 14. and Table 15. presents the performance evaluation metrics of the Random Forest Model:

Table 5. Conflict Matrix for the Random Forest Model

PREDICTED CLASS			
ACTUAL CLASS			
	WC	WC	NWC
	NWC	1355	17
		27	188

Table 15. Performance Evaluation for the Random Forest Model

Metric	Value
Accuracy	0.97
Sensitivity	0.87

F1 Score	0.89
Precision	0.90
AUC	0.92

These findings show that the Random Forest model performs effectively in forecasting child labor. The accuracy of 97% demonstrates the model's ability to make correct predictions overall. The sensitivity of 87% indicates its capability to correctly identify positive instances of child labor. The precision of 90% highlights the model's capability to provide reliable positive predictions, and the F1 score of 89% indicates a balanced trade-off between precision and recall. An AUC value of 0.92 indicates that the binary classification model has good performance in distinguishing between positive and negative samples. The model's capacity to distinguish between the two groups is improved by the AUC value's proximity to 1. A result of 0.92 thus indicates that the model is fairly successful in its categorization duty.

According to the feature importance analysis, household size and family composition are the most important factors influencing child labor in Turkey, as shown in Figure 12:

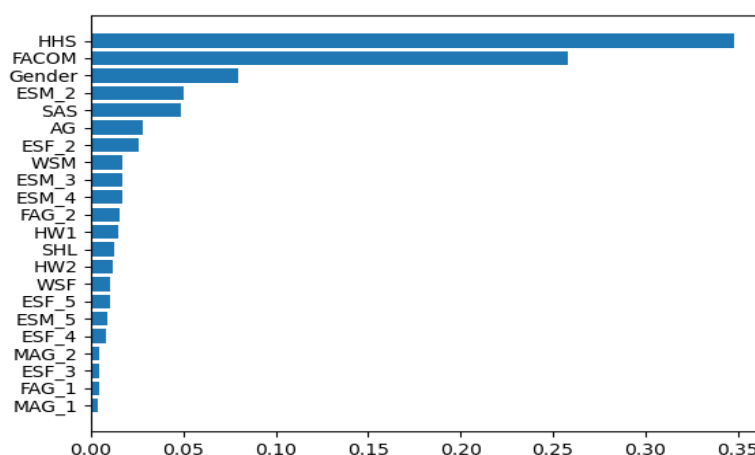


Figure 12. Random Forest Feature Importance

These findings underscore the need for focused interventions and policies aimed at raising parental education levels and strengthening household structures to reduce child work in Turkey and emphasize the significance of addressing educational and family structure elements in the fight against child labor.

3.4.5 Performance Comparisons of Logistic Regression Model, Naïve Bayes Model and Random Forest Model

This section aims to compare the performance of three classification algorithms, namely Logistic Regression, Random Forest, and Naive Bayes, in predicting child labor. The dataset used in the analysis includes features described in table 2.3. Prior to model training, the dataset undergoes thorough preprocessing steps, including handling missing values, outliers, and categorical variable encoding. The pre-processed data is then split into training and test sets, with 80% of the data used for training and the remaining 20% for testing.

The training set is used to train the algorithms. A linear model called logistic regression presupposes a linear relationship between the features and the desired outcome. An ensemble method called Random Forest combines the forecasts of several decision trees to increase accuracy. A probabilistic classifier called Naive Bayes bases its predictions on the conditional probabilities of the target variable given the characteristics, assuming independence between the features.

To assess algorithm performance, commonly used metrics include accuracy, precision, recall, F1 score, and AUC-ROC. The ability to properly identify all positives and prevent false negatives is measured by recall, while the ability to identify positive instances while avoiding false positives is measured by precision, which strikes a balance with recall, and accuracy, which evaluates total correctness. At different classification thresholds, AUC-ROC measures how well the algorithm distinguishes between positive and negative cases.

Table 16. provides a summary of the evaluation metrics for each algorithm:

Table 16. Performance Comparison of Machine Learning Classification Algorithms

Algorithm	Accuracy	Precision	Recall	F1 Score	AUC-ROC
Logistic Regression (LR)	0.91	0.71	0.57	0.63	0.76
Naive Bayes (NB)	0.86	0.5	0.55	0.52	0.67
Random Forest (RF)	0.97	0.9	0.87	0.89	0.92

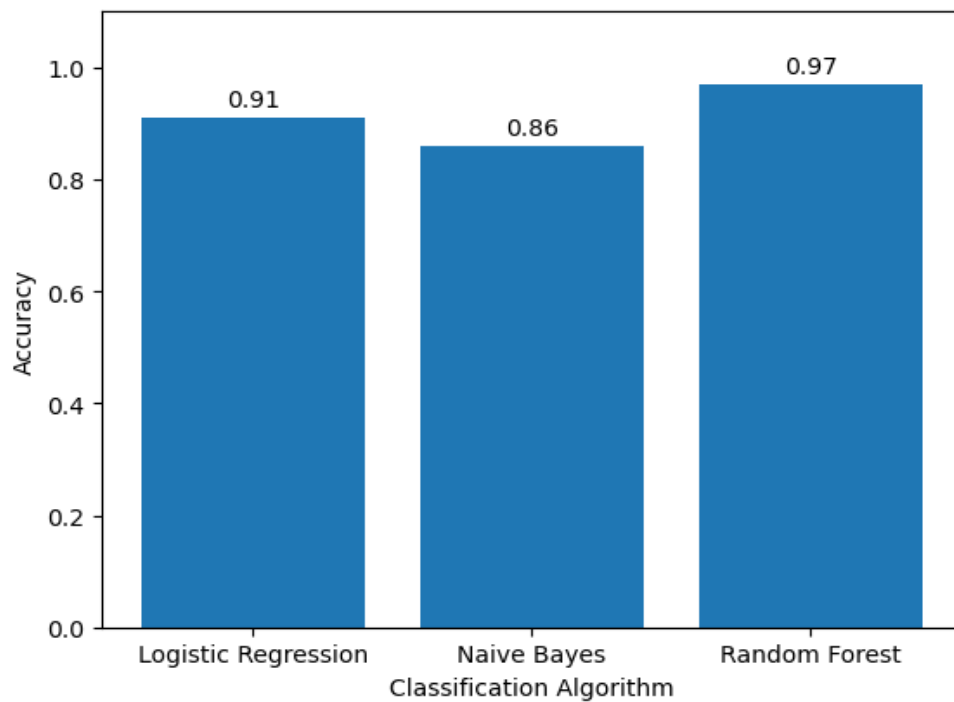


Figure 13. Accuracy Comparison of Machine Learning Classification Algorithms

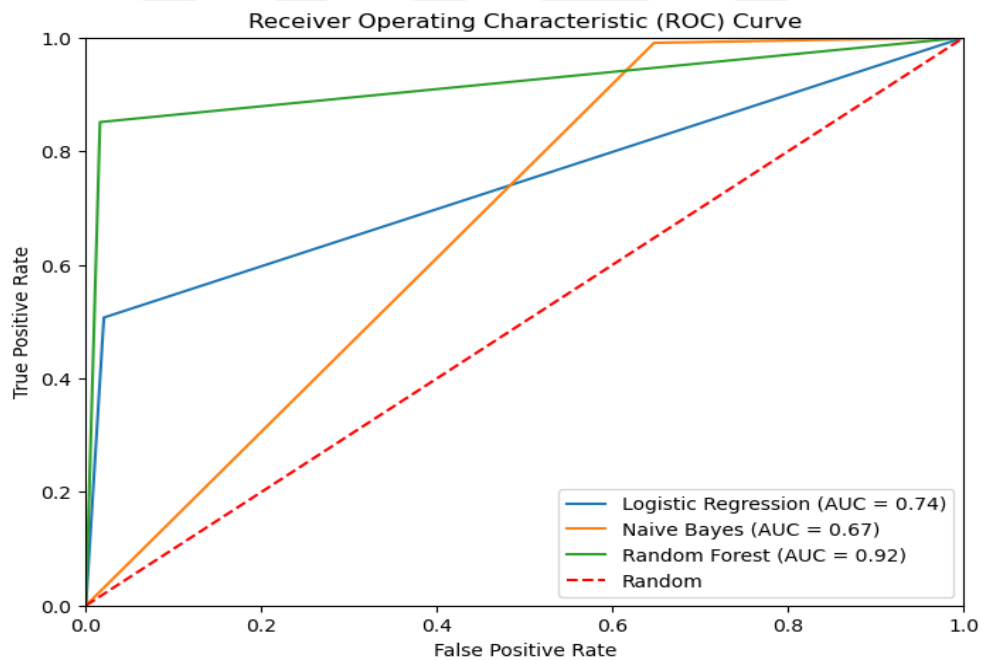


Figure 14. ROC Curve of Machine Learning Classification Algorithms

Three alternative machine learning models—Logistic Regression, Naive Bayes, and Random Forest—are compared in Table 16 to see how well they work, and numerous important metrics offer useful information. Figure 13. shows that Random Forest emerges as the most effective model, achieving an impressive accuracy of 0.97, outperforming both

Logistic Regression (0.91) and Naive Bayes (0.86). Random Forest also exhibits the highest precision (0.9), indicating that when it predicts a positive class, it is correct 90% of the time. Furthermore, Random Forest demonstrates superior recall (0.87) compared to Logistic Regression (0.57) and Naive Bayes (0.55), correctly identifying 87% of positive instances. The F1 Score reinforces Random Forest's performance, measuring an impressive 0.89, indicating a favorable balance between precision and recall. Moreover, demonstrating its capacity to rank predicted probabilities, Figure 14. shows that Random Forest produces the greatest AUC-ROC score of 0.92. In contrast, Logistic Regression performs admirably, with respectable values for F1 Score, Accuracy, Precision, Recall, and AUC-ROC. Naive Bayes, on the other hand, falls short with worse precision, recall, F1 Score, and AUC-ROC. In summary, in this study, three classification methods for predicting child labor are thoroughly compared. The statistical study, coupled with the table and graphics, show how well Random Forest performs in correctly identifying cases of child labor. The study emphasizes Random Forest's potential as a powerful prediction tool for preventing child labor.

CHAPTER 4 - CONCLUSION

In this study, we undertake a thorough analysis and comparison of three well-known machine learning classification algorithms in the fields of big data analysis and artificial intelligence: logistic regression, Naive Bayes, and random forest. Our goal is to create precise predictive models and assess how well they forecast cases of child labor in Turkey. We hope to gain insightful understanding of the intricate dynamics and root causes of child labor by utilizing the power of machine learning.

We examine the convergence of big data, artificial intelligence, and the use of machine learning throughout our inquiry, acknowledging their critical roles in addressing societal issues like child labor. The potential of machine learning in the public sphere is grounded in the tremendous data availability and policy prediction needs of this field (Pi, 2021). To create and assess the effectiveness of categorization algorithms, we make use of the enormous amount of data on child labor in Turkey that is now available. ML's predictive power allows the policy makers to anticipate a policy's impact before implementing, supporting the decisions of policy adoption (Pi, 2021). We are able to take use of these algorithms' prediction powers and shed light on many different aspects of child labor by using supervised learning approaches.

We concentrate primarily on classification algorithms, which are a key component of the machine learning framework and are essential for predicting categorical outcomes. In the field of machine learning, methods such as random forest, naive bayes, and logistic regression are well known and often used. These algorithms offer various ways of comprehending the links and patterns found in the data, and they may be able to reveal important information about the causes of child labor.

Through our comprehensive analysis and evaluation of these algorithms, we gain significant insights into the prediction of child labor instances in Turkey. The statistical correlations between various factors and the incidence of child labor can be better understood through the use of logistic regression. Naive Bayes efficiently predicts occurrences of child labor using probabilistic reasoning, demonstrating its effectiveness as a baseline technique for classification tasks. But when it comes to predicting child labor, the random forest algorithm shows the most promise, demonstrating its capacity to recognize complicated linkages and make precise predictions.

By leveraging the random forest algorithm, we achieve an impressive accuracy rate of 97% in identifying child labor cases. This outcome demonstrates the robustness and effectiveness of the random forest model in capturing intricate interactions among features and producing reliable predictions. Furthermore, the feature importance analysis within the random forest framework gender, education level, and parental occupation as the most influential factors contributing to child labor in Turkey. In order to establish effective initiatives and policies to combat child labor, policymakers and stakeholders can use these findings as important assistance.

The study has major consequences for society. It emphasizes the effects of age and limited educational options while drawing attention to gender differences, with male being more at risk of engaging in child labor. Parental education is identified as a critical barrier to child labor, and the role of working parents, particularly females, highlights the complexity of the economy. The investigation emphasizes how systematic child labor is, with the sector of the family leader having a part. Education stands out as a strong instrument to lower the risk of child labor, which is significant. In order to effectively address the pervasive problem of child labor, these insights demand for tailored initiatives that prioritize education, gender equality, and social support.

In conclusion, our research emphasizes the enormous potential of big data analysis, AI, and machine learning classification algorithms in addressing difficult societal problems like child labor. Through the evaluation of logistic regression, Naïve Bayes, and random forest algorithms, we identify the random forest algorithm as the most effective in accurately predicting child labor instances. Thanks to its powerful data processing capabilities, ML can help the public sector automate many highly labor-intensive data processing and analysis activities, thereby increasing the efficiency and speed of government services and actions (Pi, 2021). The insights gained from this study contribute to the existing body of knowledge and provide evidence-based guidance for policymakers, enabling them to implement targeted interventions and policies that effectively combat child labor in Turkey. According to Libaque-Saenz et al., (2018), this could be useful for governmental agencies to develop more effective strategies and to be more efficient in investing scarce resources to deal with this problem.

Future study should investigate more complex algorithms, include more pertinent features, and make use of cutting-edge methods for data pretreatment and model improvement as

machine learning techniques advance. Continued data gathering efforts and the use of cutting-edge machine learning techniques will improve the precision and application of prediction models, helping to eradicate child labor practices and safeguard children's rights in Turkey and around the world.



REFERENCES

- Abdallh, M. M. A., Bilal, K. H. ve Babiker, A. (2016), ‘‘Machine Learning (Pattern Recognition) Review’’, *International Journal of Engineering, Applied and Management Sciences Paradigms*, 36 /1: 17-27.
- Aktan, E. (2018), ‘‘B y k veri: Uygulama Alanları, Analitięi ve G venlik Boyutu’’, *Bilgi Y netimi Dergisi*, 1/1: 1-22.
- AlMamlook, R. E., Kwayu, K. M., Alkasisbeh, M. R., & Frefer, A. A. (2019). *Comparison of machine learning algorithms for predicting traffic accident severity*. 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT). <https://doi.org/10.1109/jeeit.2019.8717393>
- Arpacı, S. A. ve Kalıpsız, O. (2018), ‘‘Yazılım Hata Sınıflandırmasında Farklı Na ve Bayes Tekniklerin Kıyaslanması’’, * mer Halisdemir  niversitesi M hendislik Bilimleri Dergisi* 7/1: 1-13
- Atalay, M. ve  elik, E. (2017), ‘‘B y k Veri Analizinde Yapay Zeka ve Makine  ęrenmesi Uygulamaları’’, *Mehmet Akif Ersoy  niversitesi Sosyal Bilimler Enstit s  Dergisi*, 9/22: 155-172.
- Atasoy, D. (2001), ‘‘Lojistik Regresyon Analizinin İncelenmesi ve Bir Uygulaması, Y ksek Lisans Tezi, Cumhuriyet  niversitesi, Sosyal Bilimler Enstit s , Sivas.
- Athey, S. ve Imbens, Guido W. (2019), ‘‘Machine Learning Methods That Economists Should Know About’’, *Annual Review of Economics*, 11: 685-725.
- Aylak, B. L., Oral, O. ve Yazıcı, K. (2021), ‘‘Yapay Zeka ve Makine  ęrenmesi Tekniklerinin Lojistik Sekt r nde Kullanımı’’, *El-Cezeri Fen ve M hendislik Dergisi*, 8/1: 74-93.
- Bai, B., Tan, Y., Donchyts, G., Haag, A., Xu, B., Chen, G., & Weerts, A. H. (2023). Naive Bayes classification-based surface water gap-filling from partially contaminated optical remote sensing image. *Journal of Hydrology*, 616, 128791.

- Breiman. L. (2001), “Random Forests“, *Machine Learning*, 45:5–32.
- Ballestar, M. T., Doncel, L. M., Sainz, J. and Ortigosa-Blanch, A. (2019), “A novel machine learning approach for evaluation of public policies: An application in relation to the performance of university researchers”, *Technological Forecasting and Social Change*, 149.
- Barak, A., Karahan, S. ve Saraçbaşı, O. (2005). *Ordinal Lojistik Regresyon Modelleri*. 4. İstatistik Kongresi’nde Sunulan Bildiri, Belek-Antalya.
- Berry, M. J. and Linoff, G. (1997), *Data Mining Techniques: For Marketing, Sales, and Customer Support*, John Wiley & Sons, Inc., p.
- Biau, G. and Scornet, E. (2015), “A Random Forest Guided Tour”, *TEST* 25: 197–227
- Brownlee J. (2013), “How to Prepare Data for Machine Learning”, <https://machinelearningmastery.com/how-to-prepare-data-for-machine-learning/> adresinden alındı.
- Cackett, D. (2013), *Information Management and Big Data, a Reference Architecture*, White paper. Redwood Shores: Oracle Corporation.
- Cortes C. and Vapnik V. (1995), “Support-Vector Networks”, *Mach. Learn.* 20: 273–297.
- Cox, R. (1983), “Some Remarks on Overdispersion”, *Biometrika*, 70: 269-274.
- Çokluk, Ö., Şekercioğlu, G. ve Büyüköztürk, Ş. (2010), *Sosyal Bilimler için Çok Değişkenli İstatistik- Spss ve Lisrel Uygulamaları*. Ankara: Pegem Akademi Yayınları.
- Çokluk. Ö. (2010), Lojistik Regresyon Analizi: Kavram ve Uygulama. Kuram ve Uygulamada Eğitim Bilimleri, 10/3:1357-1407.
- Dean, C. B. (1992). “Testing for Overdispersion in Poisson and Binomial Regression Models”, *Journal of American Statistical Association*, 87/418: 451-457.
- Deng, H., Sun, Y., Chang, Y. ve Han, J. (2015), Probabilistic Models for Classification. C.C. AGGARWAL (Eds.), *Data Classification Algorithms and Applications* (pp. 67-70), CRC Press, New York, USA.
- Demir, S. ve Şahin, E.K. (2022), Evaluation of Oversampling Methods (OVER, SMOTE, and ROSE) in Classifying Soil Liquefaction Dataset based on SVM, RF, and Naïve Bayes. *Avrupa Bilim ve Teknoloji Dergisi*, 34: 142-147.

- Doğan, K. ve Arslantekin, S. (2016), “Büyük Veri: Önemi, Yapısı ve Günümüzdeki Durum”, *DTCF Dergisi*, 56/1: 15-36.
- Dornadula, V. N., & Geetha, S. (2019). Credit card fraud detection using machine learning algorithms. *Procedia Computer Science*, 165, 631–641. <https://doi.org/10.1016/j.procs.2020.01.057>
- Erl, T., Wajid K. and Paul B. (2016), *Big data Fundamentals. Concepts, Drivers & Techniques*. Indiana: Arcitura Education Inc.
- Field, A. (2009). *Discovering Statistics Using SPSS*. London: Sage.
- Friedberg R, Tibshirani J, Athey S. and Wager S. (2018), Local Linear Forests. arXiv:1807.11408 [stat.ML]
- Engelen, J. E. and Holger H. H. (2020), “A Survey on Semi-Supervised Learning”. *Machine Learning*, 109, 373–440.
- Jing B., Jian-Yun N. and François P. (2023), Using Language Models for Text Classification <http://rali.iro.umontreal.ca/rali/sites/default/files/publis/LMforTextClassification.pdf>, adresinden alındı.
- Jordan, M. I. and Mitchell, T. M. (2015). “Machine Learning: Trends, Perspectives, and Prospects”, *SCIENCE*, 349/6245: 255-261.
- Gaussian Naïve Bayes. scikit. (n.d.). https://scikit-learn.org/stable/modules/naive_bayes.html#gaussian-naive-bayes
- Gerşil, G. S. (2017). Çalışan Çocuk Gerçeği ve Sağlık Sorunları. Retrieved from https://www.academia.edu/16049792/percentC3percent87ALpercentC5percent9EAN_percentC3percent87OCUK_GERpercentC3percent87EpercentC4percent9EpercentC4percentB0_VE_SApercentC4percent9ELIK_SORUNLARI on 27 December 2021.
- Giuseppe C., Ignacio C., Kyle C., Laurent D., Maria S., Naftali T., Leslie Vogt-M., and Lenka Z. (2019), “Machine Learning and the Physical Sciences”. *CHORUS*, 9: 1-46.
- Gürsaka1, N. (2013), *Büyük Veri*. Bursa: Dora.

- Gürses, D. (2009), “An analysis of child poverty and social policies in Turkey”, *Journal of Developing Societies*, 25/2: 209-227.
<https://doi.org/10.1177/0169796X0902500204>
- Haghighat, A., & Belyadi, H. (2021). Machine Learning Guide for oil and gas using python: A step-by-step breakdown with data, algorithms, codes, and applications. Gulf Professional Publishing.
- Hair, J. F, Black, W. C., Babin, B., Anderson, R. E. and Tatham, R. L. (2006), *Multivariate Data Analysis*. Upper Saddle River, NJ: Prentice-Hall.
- Haltaş, A. and Alkan, A. (2013), İmmunohistokimyasal Boyalar İle Tiroid Tümörü Teşhisinde Naive Bayes Algoritması Kullanılması. *Ankara Bilişim Konferansı*.
- Han, J., Pei, J. ve Kamber, M., 2011, Data Mining: Concepts and Techniques, Elsevier.
- Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1), 100-108.
- Hazım, L.R. (2018). Four Classification Methods Naïve Bayesian, Support Vector Machine, K-Nearest Neighbors and Random Forest Are Tested For Credit Card Fraud Detection, Master Thesis, Altıbas University, Graduate School Of Sciences Engineering.
- Hosmer D.W. and Lemeshow, S. (2000), *Applied Logistic Regression*. John WileySons, Inc.:Canada.
- Hannah, W., Daniel, K. and Saskia, S. (2019), A literature review on machine learning in supply chain management. *International Conference of Logistics (HICL)*, 27: 413-441.
- Hussain, R. G., Ghazanfar, M. A., Azam, M. A., Naeem, U., & Ur Rehman, S. (2018). A performance comparison of machine learning classification approaches for robust activity of daily living recognition. *Artificial Intelligence Review*, 52(1), 357–379. <https://doi.org/10.1007/s10462-018-9623-5>
- İbrahim, M.H.I. (2019), Ayırıklaştırma ve Optimizasyon Yaklaşımları ile Sınıflandırma Algoritmalarının Performansının İyileştirilmesi, Doktora Tezi, Selçuk Üniversitesi, Fen Bilimleri Enstitüsü, Konya.

- ILO. (2022), Child labour. Child Labour (ILO-Ankara).
<https://www.ilo.org/ankara/projects/child-labour/lang--en/index.htm>
- ILO. (2021), What is child labour. Retrieved from <https://www.ilo.org/ipecc/facts/lang--en/index.htm> on 03 June 2022.
- Kahraman, F. ve Sallan Gül, S. (2015), “Türkiye’de Çocuk Yoksulluğu: Gaziantep Üzerine Bir Araştırma“, *Hitit Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 8/1: 339-366. <https://doi.org/10.17218/husbed.59837>
- Kalaycı, Ç. (Ed.). (2009), *SPSS Uygulamalı Çok Değişkenli İstatistik Teknikleri*. Ankara: Asil Yayın Dağıtım Ltd.Şti.
- Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *ICT Express*, 7(4), 432–439. <https://doi.org/10.1016/j.ict.2021.02.004>
- Koçali, K. (2022), “Perspective on Child Labor Situation in Terms of Occupational Health and Safety: The Case of Turkey”. *Journal of Human and Social Sciences*, 5/2: 126-151.
- Köseoğlu, Ö. and Demirci, Y. (2017), “Türkiye’de Büyük Veri ve Veri Madenciliğine İlişkin Politika ve Stratejiler: Ulusal Politika Belgelerinin İçerik Analizi”, *Süleyman Demirel Üniversitesi*, 22/15, 2223-2239.
- Libaque-Saenz, C. F., Lazo, J., Lopez-Yucra, K. G. and Bravo, E. R. (2018), “Could Machine Learning Improve the Prediction of Child Labor in Peru?” *Information Management and Big Data*, 15–30. https://doi.org/10.1007/978-3-319-90596-9_2
- Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2020), “Explainable AI: A Review of Machine Learning Interpretability Methods”, *Entropy*, 23/1: 18. <https://doi.org/10.3390/e23010018>
- Marsland, S. (2015), *Machine Learning An Algorithmic Perspective*. CRC Press, Boca Raton, Florida.
- Madhusri, V., Kesavkrishna, G., Marimuthu, R., & R., S. (2019). Performance comparison of machine learning algorithms to predict labor complications and birth defects based on stress. 2019 IEEE 10th International Conference on

Mertler, C. A. ve Vannatta, R. A. (2005), *Advanced and Multivariate Statistical Methods: Practical Application and Interpretation*. Glendale, CA: Pyrczak Publishing.

Mirza, T., & Hassan, M. M. (2020). Prediction of school drop outs with the help of machine learning algorithms.
https://www.researchgate.net/publication/342833162_PREDICTION_OF_SCHOOL_DROP_OUTS_WITH_THE_HELP_OF_MACHINE_LEARNING_ALGORITHMS

Mumtaz, Z. (2022), “Machine Learning-Based Approach for Exploring the Household Survey Data”. *Informal Social Protection and Poverty*, 141–200.
https://doi.org/10.1007/978-981-19-6474-9_7

O. Chapelle, B. Scholkopf and A. Zien, Eds., "Semi-Supervised Learning (Chapelle, O. et al., Eds.; 2006) [Book reviews]," in *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 542-542, March 2009, doi: 10.1109/TNN.2009.2015974.

Oğuzlar, A (2011), *Temel Metin Madenciliği*. Bursa: Dora.

Öngel, S. (2017), “Türkiye’de Çocuk İşçiliği Gerçeği Raporu”, *Mesleki Sağlık ve Güvenlik Dergisi*, 16/62-63: 83-85. Retrieved from <https://www.ttb.org.tr/MSG/images/files/dergi/62-63/62-63.pdf> on 12 May 2023.

Özdemir, S. (2018), “Random Forest Yöntemi Kullanılarak Potansiyel Dağılım Modellemesi ve Haritalaması: Yukarıgökdere Yöresi Örneği”, *Türkiye Ormancılık Dergisi*, 19/1: 51-56.

Pi, Y. (2021), “Machine learning in governments: Benefits, challenges and future directions”, *JeDEM - eJournal of eDemocracy and Open Government*, 13/1: 203–219. <https://doi.org/10.29379/jedem.v13i1.625>

Salur, M. U. (2016), Büyük Veri Araçlarından Hadoop Kullanarak Veri Madenciliği, Yüksek Lisans Tezi, Pamukkale Üniversitesi, Fen Bilimleri Enstitüsü, Denizli.

Sandhya N, Charanjeet KR (2016) A review on machine learning techniques. *Int J Recent Innov Trends Comput Commun* 395–399

- Savaş S. Topaloğlu N. and Yılmaz M. (2012), “Veri Madenciliği ve Türkiye'deki Uygulama Örnekleri”, *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*. 11/21: 1-23.
- Tabachnick, B.G. and Fidell, L.S. (2001) Using Multivariate Statistics. 4th Edition, Allyn and Bacon, Boston.
- Teklehaymanot, G. G., & Gandhi, N. (2018, July). Analyzing children’s data using Machine Learning: A case study in Ethiopia. https://www.researchgate.net/publication/328853488_Analyzing_Children’s_Data_Using_Machine_Learning_A_Case_Study_in_Ethiopia
- TURKSTAT. (2021). Turkish Household Labor Force Survey 2020 (Report No. 37484). Retrieved from <https://data.tuik.gov.tr/Bulten/Index?p=Labour-Force-Statistics-2020-37484&dil=2>
- Yavuz, Ö. (2021). An Optimization Focused Machine Learning Approach in Analysing Arts Participative Behavior with Fine Arts Education Considerations . International Scientific and Vocational Studies Journal , 5 (2) , 241-253 . DOI: 10.47897/bilmes.1029139
- Witten, I. H., Frank, E., Hall, M. A. and Pal, C. J. (2016), *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann.
- Zhu X (2008) Semi-supervised learning literature survey—TR1530. Technical report, University of Wisconsin, Madison