ANKARA YILDIRIM BEYAZIT UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES



MACHINE LEARNING-BASED FEATURE SELECTION APPROACH FOR NO-SHOW RATE PREDICTION: A CASE OF AVIATION INDUSTRY

M.Sc. Thesis by

Ahmet Süha HANCIOĞLU

Department of Industrial Engineering

July, 2024

ANKARA

MACHINE LEARNING-BASED FEATURE SELECTION APPROACH FOR NO-SHOW RATE PREDICTION: A CASE OF AVIATION INDUSTRY

A Thesis Submitted to

The Graduate School of Natural and Applied Sciences of

Ankara Yıldırım Beyazıt University

In Partial Fulfillment of the Requirements for the Degree of

Master of Department of Industrial Engineering

by

Ahmet Süha HANCIOĞLU

July, 2024

ANKARA

M.Sc. THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"MACHINE LEARNING-BASED FEATURE SELECTION APPROACH FOR NO-SHOW RATE PREDICTION: A CASE OF AVIATION INDUSTRY"** completed by **AHMET SÜHA HANCIOĞLU** under the supervision of **ASSOC. PROF. DR. İBRAHİM YILMAZ** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assoc. Prof. Dr. İbrahim YILMAZ

Supervisor

Prof. Dr. Ergün ERASLAN

Jury Member

Assist. Prof. Dr. Emre ÇALIŞKAN

Jury Member

Prof. Dr. Sadettin ORHAN

Director

Graduate School of Natural and Applied Sciences

ETHICAL DECLARATION

I hereby declare that, in this thesis which has been prepared in accordance with the Thesis Writing Manual of Graduate School of Natural and Applied Sciences,

- All data, information and documents are obtained in the framework of academic and ethical rules,
- All information, documents and assessments are presented in accordance with scientific ethics and morals,
- All the materials that have been utilized are fully cited and referenced,
- No change has been made on the utilized materials,
- All the works presented are original,

and in any contrary case of above statements, I accept to renounce all my legal rights.

Date: 08.07.2024 **Signature: ……..**

Name & Surname: Ahmet Süha HANCIOĞLU

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my supervisor, Assoc. Prof. Dr. İbrahim YILMAZ, for his invaluable guidance, continuous support, and patience throughout the course of this research. His expertise and insights have been indispensable, and his encouragement has kept me motivated during the most challenging times of this journey.

I am also immensely grateful to the faculty and staff of the Department of Industrial Engineering at Ankara Yıldırım Beyazıt University for their support and assistance throughout my studies. Their knowledge and dedication have greatly contributed to my academic growth.

A special note of thanks to my colleagues and friends who have provided not only technical assistance but also emotional support. Your camaraderie has made this experience much more enriching and enjoyable.

I would like to extend my heartfelt appreciation to my family. To my parents, for their unwavering support and belief in me; their encouragement has been my driving force. To my mother, for her endless love and for always believing in my potential. To my father, for his wise counsel and steadfast support. To my brother, for his constant encouragement and understanding. Their love and support have been crucial in achieving this milestone.

Finally, I want to express my deepest gratitude to my dear fiancé, whose support and motivation have been instrumental in keeping me focused and determined. Thank you for being a part of this journey and for helping me reach this significant point in my academic career.

2024, 08 July **Ahmet Süha HANCIOĞLU**

MACHINE LEARNING-BASED FEATURE SELECTION APPROACH FOR NO-SHOW RATE PREDICTION: A CASE OF AVIATION INDUSTRY

ABSTRACT

In today's environment of increasing competition, rapidly changing customer demands, and globalization, companies are compelled to operate with low profit margins and adapt swiftly to changing conditions to gain a competitive advantage. This thesis was conducted during a period when the importance of technologies such as artificial intelligence and big data is increasing in the highly competitive aviation industry. The aviation industry is known for its high capital requirements, low profit margins, intense competition, complex operational structures, and constant pressure from local or global conditions. Considering the low profit margins in the aviation industry, airlines are continually encouraged to enhance their efficiency, reduce operational costs while maintaining customer satisfaction, or explore areas that could generate additional revenue.

A significant proportion of passengers who purchase tickets do not show up for their scheduled flights. This situation allows airlines to practice overbooking, and accurately predicting the rate of no-shows can provide substantial additional revenue. However, if all passengers do show up, airlines would incur additional costs and suffer reputational damage. Therefore, accurately predicting no-show rates and optimizing operational planning processes are critically important for airlines. No-show rates are generally estimated by analyzing historical data using statistical methods. However, recently developed machine learning and big data analytics have the potential to play a significant role in solving such complex problems.

There are very few studies in the literature on predicting no-show rates using machine learning methods. These studies typically use only Passenger Name Record (PNR) data for no-show prediction, neglecting many external factors. Using only passenger information reduces the accuracy of predictions due to the unique characteristics of each passenger. Unlike other studies, this research investigates the impact of external factors such as weather conditions, traffic density, special occasions, and public

holidays on no-show rates without using PNR data, employing big data analysis and machine learning methods.

This study focuses on the unique and critical need to improve operational planning processes in the aviation industry. It aims to contribute to understanding and addressing the challenges in the aviation industry, particularly in accurately calculating the no-show rate. To achieve this goal, the importance and effects of accurately predicting the no-show rate have been comprehensively examined, and a model has been proposed using machine learning algorithms to address the problem at hand.

Keywords: No-show rate prediction, machine learning, big data analytics, aviation industry, operational efficiency

NO-SHOW ORANI TAHMİNİ İÇİN MAKİNE ÖĞRENİMİ TABANLI ÖZELLİK SEÇİMİ YAKLAŞIMI: HAVACILIK ENDÜSTRİSİNDE VAKA ÇALIŞMASI

ÖZ

Günümüzde artan rekabet koşulları, hızla değişen müşteri talepleri ve küreselleşme, şirketleri düşük kar marjlarıyla çalışmaya ve hızla değişen koşullara uyum sağlayarak, rekabet avantajı elde etmeye zorlamaktadır. Bu tez, günümüzün en yoğun rekabet ortamlarından birisi olan havacılık endüstrisinde, yapay zekâ ve büyük veri gibi teknolojilerin öneminin arttığı bir dönemde gerçekleştirilmiştir. Havacılık endüstrisi, yüksek sermaye gerektiren, düşük kar marjları, yoğun rekabet, karmaşık operasyonel yapıları ve yerel ya da küresel konjonktürün sürekli baskısı altında olmasıyla bilinmektedir. Bu durum, havacılık endüstrisindeki düşük kar marjları göz önüne alındığında, havayollarını sürekli olarak verimliliklerini artırmaya, müşteri memnuniyetlerini koruyarak operasyonel maliyetlerini azaltmaya veya ek gelir yaratabilecekleri alanları keşfetmeye teşvik etmektedir. Yolcuların bilet aldığı halde planlanan uçuşlarına gelmemesi azımsanamayacak bir oranda gerçekleşmektedir. Bu durum havayolları için overbooking imkânı tanımaktadır ve uçuşa gelmeyen yolcuların oranın doğruya yakın tahmin edilmesi durumunda ciddi miktarda ek gelir elde etme imkânı tanımaktadır. Ancak yolcuların tamamının uçuşa gelmesi durumda havayolları ek maliyetlere katlanmak zorunda kalacak ve itibar kaybı yaşayacağından no-show oranlarını doğru bir şekilde tahmin etmek ve operasyonel planlama süreçlerini optimize etmek havayolları için kritik öneme sahiptir. No-show oranları genel olarak tarihsel verilerin istatistiksel yöntemlerle analiziyle tahmin edilmektedir. Bununla beraber son dönemde hızla gelişen makine öğrenimi ve büyük veri analitiği, bu tür karmaşık sorunların çözümünde önemli bir rol oynama potansiyeline sahiptir.

Literatürde no-show oranlarının makine öğrenimi yöntemleriyle tahminine yönelik çok az sayıda çalışma bulunmaktadır. Bu çalışmalarda no-show tahmini için yalnızca Passanger-Name-Record (PNR) bilgileri kullanılmakta ve bu durum birçok çevresel-dışsal faktörü göz ardı etmektedir. Yalnızca yolcu bilgilerinin kullanılması ise, her yolcunun birbirinden benzersiz özellikle sahip olması sebebiyle tahmin başarısını düşürmektedir. Bu çalışmada ise diğer çalışmalardan farklı olarak PNR kullanılmadan, hava durumu, trafik yoğunluğu gibi çevresel faktörlerin yanı sıra özel günler ve resmî

tatiller gibi dışsal faktörlerin no-show üzerindeki etkisi büyük veri analizi ve makine öğrenimi yöntemleri kullanılarak araştırılmaktadır.

Bu çalışma, havacılık endüstrisindeki operasyonel planlama süreçlerini iyileştirme konusunda benzersiz ve kritik ihtiyacına odaklanmaktadır. Aynı zamanda havacılık endüstrisindeki zorlukları anlama ve no-show oranını hesaplamadaki mevcut sınırlamaları ele alma konusunda katkı sunması amaçlanmıştır. Bu amacı gerçekleştirmek için no-show oranını doğru bir şekilde tahmin etmenin önemi ve etkileri kapsamlı bir şekilde incelenmiş, ele alınan probleme yönelik makine öğrenimi algoritmaları ile bir model sunulmuştur.

Anahtar Kelimeler: No-show oranı tahmini, makine öğrenimi, büyük veri analitiği, havacılık endüstrisi, operasyonel verimlilik

CONTENTS

M.Sc. THESIS EXAMINATION RESULT FORM ii
ETHICAL DECLARATION iii
ACKNOWLEDGMENTS iv
ABSTRACT v
ÖZ vii
NOMENCLATURE xii
LIST OF TABLES xiv
LIST OF FIGURES xv

CHAPTER 1 - INTRODUCTION 1
1.1 Problem Definition of Thesis 6
1.2 Hypothesis 7
1.3 Thesis Overview 7

CHAPTER 2 – LITERATURE REVIEW 8
2.1 Review of Big Data and Machine Learning Studies Conducted in Different Branches 8
2.2 Review of Studies Conducted in the Field of Aviation 10
2.3 Review of Research on No-Show Rate Estimation in Aviation 16

CHAPTER 3 – BASICS OF BIG DATA & MACHINE LEARNING 18
3.1 Big Data 19
3.2 Data Mining 23
3.2.1 Big Data: Technologies and Tools 25
3.2.2 Data Preprocessing 26
3.2.2.1 Data Cleaning 26
3.2.2.2 Data Reduction 28
3.2.2.3 Feature Extraction 29
3.2.2.4 Instance Reduction (IR) 29
3.2.2.5 Discretization 30
3.3 Importance of Big Data 30
3.4 Overview of the Aviation Industry 32

3.5 No Show Rate Prediction on Airline Industry 34
3.6 Machine Learning 36
3.7 Types of Machine Learning 38
3.7.1 Supervised Learning 41
3.7.1.1 Simple Linear Regression 47
3.7.1.2 Multiple Linear Regression 48
3.7.1.3 Ridge Regression 49
3.7.1.4 Lasso Regression 49
3.7.1.5 Logistic Regression 50
3.7.1.6 K-Nearest Neighbors Algorithm 51
3.7.1.7 Support Vector Machines 51
3.7.1.8 Artificial Neural Networks 52
3.7.1.9 Decision Trees 53
3.7.1.10 Random Forests 54
3.7.1.11 Gradient Boosting Machines (GAM) 55
3.7.1.12 Light Gradient Boosting Machines 55
3.7.1.13 Extreme Gradient Boosting 56
3.7.1.14 Naive Bayes 56
3.7.2 Unsupervised Learning 56
3.7.2.1 K-means Clustering 58
3.7.2.2 Hierarchical Clustering 59
3.7.2.3 Principal Component Analysis (PCA) 61
3.7.2.4 Apriori Algorithm 62
3.7.3 Semi-supervised Learning 63
3.7.3.1 Self-Training Methods 64
3.7.3.2 Generative Mixture Models 64
3.7.3.3 Co-Training Algorithm 65
3.7.3.4 Semi-Supervised Support Vector Machines (S3VM) 65
3.8 Model Validation 67
3.8.1 k-fold Cross Validation 68
3.8.2 Stratified k-fold Cross Validation 68
3.8.3 Hold-out Validation 69
3.8.4 Leave-one-out Cross Validation (LOOCV) 69
3.9 Machine Learning Algorithms Evaluation Metrics 70
3.9.1 Evaluating the Performance of Classification Models 70

3.9.2 Assessment of Regression Model Performance 78

CHAPTER 4 - METHODOLOGY 81

4.1 Framework of Research 81

4.2 Data Acquisition 82

4.3 No-Show Prediction Model 83

4.3.1 Data Preparation 83

4.3.2 Descriptive Statistics 83

4.3.3 Data Preprocessing 92

4.3.3.1 Missing Values 93

4.3.3.2 Data Reduction & Data Balancing 93

4.3.3.3 Converting Categorical Covariates 105

4.3.4 Model Selection 106

4.3.5 Technical Specifications of the Computational Environment 107

4.4 Results and Sensitivy Analysis 107

4.4.1 Results 107

4.4.1.1 Random Forest Regression 107

4.4.1.2 Gradient Boosting 108

4.4.1.3 Linear Regression 108

4.4.1.4 Ridge Regression 109

4.4.1.5 Lasso Regression 109

4.4.1.6 Support Vector Regression (SVR) 110

4.4.2 Model Validation 111

4.4.3 Sensitivity Analysis 112

CHAPTER 5 - CONCLUSION 124

REFERENCES 131

CURRICULUM VITAE 149

NOMENCLATURE

Roman Letter Symbols

A	Area
C	Capacity
n	Number of observations
T	Temperature
U	Utility
V	Volume or velocity

Greek Letter Symbols

α	Significance level
β	Coefficient in regression analysis
γ	Rate parameter or learning rate
θ	Parameter
λ	Regularization parameter in Lasso/Ridge regression
μ	Mean
σ	Standard deviation or sum operator
Δ	Change or difference
Σ	Summation
Π	Product symbol
π	Mathematical constant (pi)
∇	Gradient
∂	Partial derivative
∞	Infinity

Abbreviations

AI	Artificial Intelligence
ADS-B	Automatic Dependent Surveillance-Broadcast
AUC	Area Under the Curve
ANN	Artificial Neural Network
BPCA	Bayesian Principal Component Analysis
CNN	Convolutional Neural Network
CRISP-DM	Cross Industry Standard Process for Data Mining
EM	Expectation-Maximization

FDR	Flight Data Recorder
FDM	Flight Data Monitoring
GAM	Gradyan Artırma Makineleri (Gradient Boosting Machines)
GBDT	Gradient Boosting Decision Tree
HDFS	Hadoop Distributed File System
IoT	Internet of Things
ITS	Intelligent Transportation Systems
kNN	k-Nearest Neighbors
LASSO	Least Absolute Shrinkage and Selection Operator
LDA	Latent Dirichlet Allocation
LOOCV	Leave-One-Out Cross Validation
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MI	Multiple Imputation
ML	Machine Learning
MSE	Mean Squared Error
MV	Missing Values
NLP	Natural Language Processing
PNR	Passenger Name Record
POS	Point of Sale
PCA	Principal Component Analysis
RF	Random Forest
RM	Revenue Management
RMSE	Root Mean Square Error
S3VM	Semi-Supervised Support Vector Machines
SVR	Support Vector Regression
SVM	Support Vector Machines
TF-IDF	Term Frequency-Inverse Document Frequency
VOD	Video on Demand
XGBoost	Extreme Gradient Boosting

LIST OF TABLES

Table 2.1 Review of big data and machine learning 9
Table 2.2 Aviation studies summary 14
Table 3.1 Confusion Matrix 73
Table 4.1 Descriptive statistics of variable**s** 83
Table 4.2 Five-Number summaries of variables 85
Table 4.3 Relations between continuous variables and target variable 94
Table 4.4 Relations between categorical variables and target variable 95
Table 4.5 Relations between continuous variables and target variable after convert 97
Table 4.6 Example matrix structure for an encoded variable 106
Table 4.7 Random Forest Regression performance metrics 107
Table 4.8 Gradient Boosting Regression performance metrics 108
Table 4.9 Linear Regression performance metrics 108
Table 4.10 Ridge Regression performance metrics 109
Table 4.11 Lasso Regression performance metrics 109
Table 4.12 Support Vector Regression performance metrics 110
Table 4.13 Model validation test results 111
Table 4.14 R^2 values according to models after feature extraction 113

LIST OF FIGURES

Figure 3.1 5V's of Big Data 21
Figure 3.2 CRISP-DM technique 24
Figure 3.3 Machine Learning algorithms 36
Figure 3.4 Classification and Regression models 38
Figure 3.5 Overfit, underfit and good fit models 41
Figure 3.6 Stages of Supervised Learning 46
Figure 3.7 Linear Regression 48
Figure 3.8 SVM hyperplane in dataset 52
Figure 3.9 ANN working principle 53
Figure 3.10 Decision Tree working principle 54
Figure 3.11 Clustering and Cassification difference on dataset 58
Figure 3.12 K-means and Hiearhical Clustering on datasets 61
Figure 3.13 Principles of Machine Learning algorithms 66
Figure 3.14 ROC curve 76
Figure 4.1 Daily Trip Count histogram 100
Figure 4.2 Montly Trip Count histogram 101
Figure 4.3 Capacity histogram 102
Figure 4.4 Temperature histogram 102
Figure 4.5 Wind Speed histogram 103
Figure 4.6 Hour frequency 104
Figure 4.7 Day frequency 104
Figure 4.8 Condition frequency 105
Figure 4.9 Random Forest R^2 value changes 117
Figure 4.10 Gradient Boosting R^2 value changes 118
Figure 4.11 Linear Regression R^2 value changes 119
Figure 4.12 Ridge Regression R^2 value changes 120
Figure 4.13 Lasso Regression R^2 value changes 121
Figure 4.14 SVM Regression R^2 value changes 122
Figure 4.15 Impact of feature removal on performance 123
Figure 4.16 Algorithms average R^2 results 123

CHAPTER 1

INTRODUCTION

From the Industrial Revolution to the present day, technology has rapidly advanced. Unlike the past, the current period is characterized by technology permeating every aspect of people's lives, work patterns, and social relationships, fundamentally changing daily life. With globalization, inter-company competition has mercilessly intensified, and thanks to companies achieving economies of scale, the unit costs of products have rapidly decreased. However, battles for market share among competitors have compelled companies to operate with low-profit margins. Technological advancements in the last 20 years have elevated competition to another level. Schwab defines the most intense and significant challenge we face today as understanding and shaping the new technological revolution [1]. The current era is undergoing fundamental changes following the widespread adoption of computers and the internet. The simplest example of this is the shift from the largest companies in 2000 to the largest companies in 2024 [2]. Many tasks once deemed impossible are now automated through machines and robots. Globally, an immense amount of data is generated every day, approximately 402.74 million terabytes [3]. Companies that grasp the significance of this data and utilize it through evolving algorithms gain a significant competitive advantage. Natural language processing and artificial intelligence are anticipated to replace two million manufacturing workers by 2025 [4]. All of this is made possible through applications that mimic human intelligence, developed under the leadership of companies striving to utilize generated data for their own operations without explicitly programming different data types and relationships, through self-learning algorithms.

Previously, only data defined in rows and columns could be analyzed with statistical methods to make predictions. However, today, predictions and analyses can be made using all kinds of data, including audio, images, tweets, and texts. Artificial intelligence-supported robots capable of analyzing human emotions, performing surgeries, and driving cars have the potential to successfully accomplish much more

complex tasks with the advancement of methods like natural language processing. Complex tasks such as teaching, human transportation like taxi or shuttle driving, or making diagnoses by providing clear prompts have the potential to be performed with ease.

The service sector is at the forefront of these areas. One of the greatest advantages of technological developments has been in the transportation sector, particularly in aviation. The speed of information dissemination increased with the development of aviation technologies before the discovery of the internet and computers. The advancement, facilitation, and acceleration of transportation technologies, especially aircraft, have increased both the speed of people and information circulation. Undoubtedly, this has played a crucial role in accelerating technological advancements in aviation.

However, despite technological advancements, they have not necessarily positively impacted profitability in the aviation industry. One of the main reasons for this is the intense regulations and challenging competitive conditions in the aviation sector. Moreover, past regulations that hindered competition in the aviation industry may have led companies of that time to focus on reducing costs and increasing profitability by being close to lawmakers, potentially causing structural damage to the industry.

Despite the increasing technology, developments in aviation have been limited to certain areas, although advancements in aircraft competition have led to significant improvements in both cost and comfort. Fierce competition in all service sectors worldwide drives companies to think innovatively and competitively, yet airlines, except for low-cost flights, remain stagnant in innovative approaches. The demand for flights by passengers, and thus predicting where and when they will travel to plan capacity and flights, aiming to fly with as many occupied seats as possible, and reducing delays and maintenance costs, provides a vast field of work for big data. Studies in the literature focus on Revenue Management, Capacity Allocation, and delay prediction. Efficiency improvements in aviation generally focus on operational enhancements and cost reduction while maintaining customer satisfaction, and opportunities to generate additional revenue are critical for gaining a competitive advantage. Passengers not showing up for their scheduled flights, despite purchasing

tickets, is termed as no-show. This occurs at a significant rate. To generate additional revenue, airlines often sell more tickets than their capacity, a practice known as overbooking. The primary purpose of overbooking is to create additional revenue through the extra tickets sold in case of no-shows. Although this may seem minimal for a single ticket, considering the approximately 130,000 commercial flights daily [5], it represents a tremendous revenue source. Overbooking is a legal practice for airlines, with specific regulations in place [6]. If all passengers show up for a flight, the excess ticket sales lead to additional costs. The costs vary based on the process, with providing an alternative flight ticket being the lowest cost. However, if this situation is unacceptable due to passenger rights, the minimum request of the person who has to be removed from the flight (cash compensation, hotel expenses, etc.) must be met. The additional cost incurred by the company, if all passengers show up, far exceeds the revenue gained from overbooking, requiring careful management of the entire process.

In the literature, the concept of overbooking is addressed in two fundamental ways. The first methodology involves determining the upper limit for extra tickets to be sold using statistical methods or developed mathematical models, based on information obtained from historical data, including revenue and costs associated with overbooking. The second methodology involves predicting no-shows directly on a passenger basis using machine learning methods with data containing passenger information, such as PNR numbers. Direct passenger-based predictions have some disadvantages. Firstly, using only passenger information overlooks environmental factors like weather, traffic density, special occasions, and public holidays. Additionally, every passenger has unique characteristics. Factors such as the reason for travel, necessity of travel, job status, income level, psychological condition, sleep status the previous night, or fatigue during the day vary for each passenger and can change momentarily. Predictions based on passenger information, which includes standard information like age, weight, gender, and ticket class, have limited generalization capability for no-show predictions.

In this thesis, however, the effects of external factors such as weather conditions, holidays, flight hours, and traffic density are examined. The primary reason is to mitigate the disadvantages caused by individual-based information and because these

factors are generalizable for all passengers. Calculating the no-show rate, predicting the likelihood of traveling passengers canceling their reservations or not showing up, can help more accurately predict it. This allows airlines to manage their aircraft capacities and revenues more effectively.

This study uniquely investigates the impact of external factors such as weather conditions, traffic density, flight time, flight day, and other contextual elements on no-show rates. Unlike previous research focused on historical data-based predictions, this study considers the entire flight, including anomalies like public holidays, weather conditions, and flight routes. Traditional methods, which assume similar flights will have similar no-show rates, fail to incorporate these external factors. By employing Machine Learning techniques, this study aims to analyze categorical data comprehensively and uncover previously unrecognized patterns. This approach is expected to enhance the accuracy of no-show predictions by accounting for variations among similar flights, thus providing sharper insights. Additionally, the limited information provided by PNR numbers is insufficient for making distinctive classifications for no-show predictions. The no-show behavior of a passenger in a particular age group and gender, even with additional information like marital status, cannot be distinctly explained compared to thousands of other passengers with similar profiles. This statistical approach is challenging because each person's unique characteristics and circumstances complicate the prediction of specific no-shows based on limited PNR information.

Accurately estimating the no-show rate offers various advantages for airlines. It enhances operational efficiency and reduces costs by allowing airlines to optimize flight planning and seat allocation processes. This results in more efficient flight management, minimized empty seats, and maximized occupancy rates, ultimately increasing revenues. Additionally, accurate no-show rate predictions provide greater flexibility in reservation management and flight planning. By considering no-show rates, airlines can quickly and effectively respond to demand fluctuations, optimize flight occupancy rates, and adjust seat allocation policies and reservation cancellation conditions, indirectly boosting revenue.

Accurate no-show rate predictions also improve customer satisfaction and reputation management. Optimized flight occupancy rates lead to higher customer satisfaction and a stronger airline reputation. Customers appreciate flexibility in changing or canceling their reservations, which enhances passenger comfort and service quality compared to competitors, thus strengthening the airline's competitive advantage. Moreover, accurate no-show rate predictions enable airlines to refine their revenue management and marketing strategies. By adjusting pricing strategies and developing special offers for different market segments based on no-show rates, airlines can better manage flexible demand and increase revenues. The ability to accurately predict no-show rates directly impacts stock prices and shareholder satisfaction by enhancing the company's value. Increased market share due to competitive advantage leads to higher revenues and company value in the medium to long term, benefiting investors.

However, accurately calculating no-show rates poses challenges. The accuracy of these calculations depends heavily on historical data and passenger behavior analysis, which can be time-consuming and costly to obtain. Collected data must be accurate and complete; otherwise, the calculations may be affected. No-show rates vary continuously based on factors such as individuals, travel dates, destinations, weather conditions, airline policies, and economic conditions, requiring constant updates and reassessments. Despite predictions, there can still be a margin of error due to unexpected events or unforeseen factors, potentially impacting the airline's operations and costs. If proper planning and actions are not taken, this could cause greater harm to the airline in the long run.

Furthermore, policies or pricing models based on no-show rates can lead to customer dissatisfaction or damage the company's reputation. Frequent travelers or loyal customers may develop a negative attitude towards the airline due to surprise fees or strict cancellation policies, creating challenges in managing customer relations and maintaining reputation.

Traditional methods for calculating no-show rates, which are conventional approaches, have their own advantages and disadvantages. They are generally simple and easy to understand, requiring low data and suitable for airlines with limited data access.

However, traditional methods often result in low prediction accuracy and lack flexibility, struggling to account for variable customer behaviors or external factors.

On the other hand, machine learning and big data technologies offer advanced approaches with various benefits. These technologies provide high prediction accuracy by detecting complex relationships and patterns, understanding hidden patterns among many parameters, and quickly adapting to changing conditions. Machine learning models can offer personalized solutions to airlines by modeling outlier behaviors and creating policies based on no-show rates for each passenger. However, machine learning models require large amounts of high-quality data, which is costly and time-consuming to obtain. They are also more complex and harder to understand compared to traditional methods, leading to reduced transparency and difficulty in identifying error sources. Additionally, the development and maintenance of machine learning models require significant computational power and expertise, creating a financial burden for small and medium-sized enterprises. Lastly, data security risks, such as breaches or cyberattacks, are associated with the need for access to large amounts of data in machine learning models.

1.1 Problem Definition of Thesis

The accurate calculation of the no-show rate in the aviation industry holds critical importance both operationally and financially. However, there is a limited number of studies in the literature on this subject, indicating that research in this area has not yet been fully explored. This thesis aims to fill this gap in the aviation industry by focusing on the impact of environmental and external factors on the no-show rate using machine learning methods. Within this framework, the importance and effects of calculating the no-show rate will be thoroughly examined, and existing studies in the literature will be reviewed. Subsequently, a model will be developed for calculating the no-show rate using machine learning algorithms, and the performance of this model will be evaluated. This problem statement provides a fundamental framework that will guide the understanding and exploration of the current limitations in calculating the no-show rate in the aviation industry.

1.2 Hypothesis

This thesis proposes the feasibility of developing a model utilizing machine learning techniques for the computation of the no-show rate. This hypothesis suggests that machine learning algorithms, through the analysis of environmental factors, can offer more accurate and reliable predictions compared to traditional methods. Furthermore, the thesis anticipates that precise prediction of the no-show rate will aid airline companies in optimizing their revenue management strategies and operational planning. To validate this hypothesis, a comparison will be drawn with prevailing methodologies for calculating the no-show rate found in the existing literature, and the efficacy of the machine learning model will be assessed. The outcomes of this thesis may furnish a significant contribution to scholars and practitioners engaged in the estimation of the no-show rate within the aviation sector.

1.3 Thesis Overview

The thesis is structured as follows: Chapter 1 outlines the background of this research, providing an introduction to the topic and the framework. It discusses the challenges faced by the aviation industry and operational efficiency, addressing current inefficiencies such as airline practices and the estimation of no-show rates worldwide. The importance, uniqueness, and contribution of this research to the existing literature are also highlighted.

Chapter 2 is dedicated to the literature review, offering an in-depth analysis of previous research conducted on the topic. Chapter 3 presents the methodology applied in this study, along with a detailed description of the data used. Chapter 4 focuses on the results, presenting and discussing the findings. Finally, Chapter 5 concludes the thesis and discusses potential directions for future research.

CHAPTER 2

LITERATURE REVIEW

2.1 Review of Big Data and Machine Learning Studies Conducted in Different Branches

Pang's study examines the combination of statistical and machine learning algorithms for big data analysis modeling [7]. It presents an analysis of the effectiveness of statistical machine learning methods in financial data analysis. The study utilizes the random forest algorithm in financial credit data analysis and proposes an enhanced random forest algorithm to identify key variables for more accurate credit risk assessment.

Sundarakani et al., examine the significance of Blockchain in Big Data for supply chain management in Industry 4.0 [8]. Combining action research and case study methods, the study tests the applicability of Blockchain architecture in global logistics environments. The results address the implementation of Blockchain in supply chain management by identifying successful and unsuccessful cases in various industry scenarios.

In the study by Nassat and Kamal, the importance of effectively detecting threats in cybersecurity for protecting digital assets and privacy is emphasized [9]. It suggests that a holistic approach integrating machine learning and big data analytics plays a pivotal role in identifying and mitigating cyber threats. While machine learning techniques are examined, big data analytics is highlighted as a critical component for managing the volumes of data generated in the digital age to obtain actionable insights.

López-Martínez et al., underscores the importance of big data and artificial intelligence in innovation and predictive analytics in the healthcare sector [10]. It discusses Keralty organization's efforts in developing an intelligent big data analytics platform based on machine learning and data integration principles.

The study by Sohangir et al., investigates the significance of deep learning and big data analytics in the field of data science [11]. It mentions notable results achieved by deep learning models in speech recognition and computer vision areas and explores the use of deep learning models to enhance sentiment analysis on financial social networks.

The research by Talón-Ballestero et al., addresses the importance of customer data in the hospitality sector and the role of Big Data techniques in this area [12]. The findings suggest that Big Data technologies could play a critical role in better understanding customers' travel habits and profiles. A summary of the studies conducted on Big Data and Machine Learning in different branches is presented in Table 2.1.

The study conducted by Shi and Abdel-Aty addresses the impact of big data in the fields of science and engineering [13]. Specifically, it examines the changes in traffic management and safety, particularly with the growing popularity of Intelligent Transportation Systems (ITS). Emphasizing the use of proactive approaches such as real-time traffic monitoring strategies and accident prediction models, the study highlights the potential of big data in traffic management. The results indicate that the large data generated by ITS systems plays a significant role in enhancing traffic operations and safety.

In the research by Luckow et al., the utilization of Hadoop in the automotive industry is investigated and analyzed [14]. Particularly, it explores how Hadoop's capabilities in processing big data can be integrated into automotive applications and identifies effective usage scenarios. The research discusses how the Hadoop ecosystem can be utilized to meet advanced data discovery, integration, exploration, and analysis requirements in the automotive industry.

Table 2.1 Review of big data and machine learning

Author-Year	Title	Category
Pang, (2023)	Big Data Analysis Method based on Statistical Machine Learning: A Case Study of Financial Data Modeling	Finance

Table 2.1 (continued) Review of big data and machine learning

Author-Year	Title	Category
Pang, (2023)	Big Data Analysis Method based on Statistical Machine Learning: A Case Study of Financial Data Modeling	Finance
Sundarakani et al., (2021)	Big data driven supply chain design and applications for blockchain: An action research using case study approach	BlockChain
Nassar and Kamal, (2021)	Machine Learning and Big Data analytics for Cybersecurity Threat Detection: A Holistic review of techniques and case studies	Cybersecurity
López-Martínez et al., (2020)	A case study for a big data and machine learning platform to improve medical decision support in population health management	Health care
Sohangir et al., (2018)	Big Data: Deep Learning for financial sentiment analysis	Finance
Talón-Ballestero, (2018)	Using big data from customer relationship management information systems to determine the client profile in the hotel sector	Hotel (Service)
Shi & Abdel-Aty, (2015)	Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways	Transportation
Luckow et al., (2015)	Automotive big data: Applications, workloads and infrastructures	Automotive

2.2 Review of Studies Conducted in the Field of Aviation

Helgo, in the study, examines the use of machine learning and deep learning algorithms in the aviation industry, emphasizing their potential to enhance effectiveness in areas such as aircraft diagnosis/prognosis, predictive maintenance, feature selection, and flight data monitoring (FDM) [15].

Adamopoulou and Daskalakis' research aims to academically review significant publications surrounding the application of Big Data in aviation [16]. The findings indicate that Big Data provides various benefits in the aviation sector and extends to other sectors such as health, environment, and humanitarian operations.

Xu et al., explore how model selection using the Least Absolute Shrinkage and Selection Operator (LASSO) method could better predict stock prices of three major international airlines using daily closing prices [17].

Huang and Cheng emphasize the importance of accurately and economically predicting aircraft fuel consumption [18]. They investigate a strategy using machine learning methods, leveraging flight data recorder (FDR) and automatic dependent surveillance-broadcast (ADS-B) data.

Gao and Mavris recommend leveraging big data to mitigate the adverse environmental impact of global aviation operations by utilizing representative studies to classify widespread research trends and suggest research opportunities for the feature [19].

Park et al., emphasize the importance of customer satisfaction in the airline industry [320]. Findings indicate the effectiveness of deep learning techniques in analyzing airline customer trends.

Brunton et al., suggest utilizing big data and machine learning to solve multi-objective, constrained optimization problems arising in aircraft design and production in the aviation industry [21]. This review will explore the opportunities and challenges of integrating data-driven science and engineering into the aviation industry, emphasizing the importance of interpretable, generalizable, explainable, and certified machine learning techniques for critical safety applicationsIn the study by Ning et al., it is noted that big data analytics and deep learning approaches hold significant potential in operations, maintenance, and security areas, particularly in the commercial aviation sector [22]. The research demonstrates how deep learning techniques can be utilized, especially in predictive maintenance, for big data analytics.

Jiang et al., investigate the use of machine learning models in predicting flight delays [23]. High success rates are achieved in predicting flight delays through machine learning models.

Huo et al.,'s study aims to predict flight delays using machine learning methods to address congestion and overload issues at Hong Kong International Airport [24].

Hwang et al., aim to predict customers' return visits using a machine learning approach on feedback and satisfaction ratings regarding airline services [25]. In a model considering emotional features, machine learning demonstrates high accuracy in predicting customers' return visits.

The study conducted by Gui et al., discusses the use of machine learning methods to predict flight delays [26]. It is noted that a random forest-based model stands out in predicting flight delays, following the examination of comprehensive factors and comparison of different machine learning models. Oehling and Barry explore the availability of machine learning (ML) methods to generate new safety-related information from existing flight data [27]. A scalable method capable of working with large data clusters is developed, demonstrating effective analysis of 1.2 million flights.

Korfiatis et al., provide a flexible expansion to numerical ratings for understanding customer satisfaction and service quality using data obtained from online reviews [28]. Latent Dirichlet Allocation (LDA) are applied to airline passenger online reviews, indicating the significance of this information for companies and its potential use in expert systems.

Kumar and Zymbler highlight the importance of customer experience for the airline industry [29]. They present a machine learning approach for analyzing passenger reviews shared on social media platforms, identifying interesting relationships that could help improve customer experience in the airline industry.

The study presented by Ma et al., focuses on the concept of denied boarding, which arises when all passengers show up under the frequently used overbooking policy in aviation [30]. The study addresses the complaints caused by denied boarding through the concept of crisis management. It investigates the reputation loss of companies due to passenger complaints on widely used social media platforms today. The study is significant as it is the first to aim to use social big data in the airline industry for managing the risk associated with overbooking and crisis communication. Additionally, it plays a pioneering role in researching the ratio of bad crisis management and social media's impact on reputation and potential feature revenue loss compared to the revenue obtained from overbooking.

Sternberg et al., research how Turkish Airlines could utilize social media platforms to enhance performance metrics [31]. Big Data and machine learning methods are employed to explore concepts of customer satisfaction and sentiment analysis. Although a weak relationship between business data and social media data is found, insights into customer behaviors are provided.

In the study by Burmester et al., the application areas and technology of Big Data in aviation are examined [32]. It is highlighted that Big Data technology in the aviation context can optimize aspects such as safety, fuel consumption, maintenance processes, and flight scheduling by leveraging continuously increasing data sources such as sensors, radars, cameras, weather stations, and airports. The study explains the Big Data application process and summarizes the relevant existing Big Data methods in aviation.

Tong et al., propose a precise aircraft landing speed prediction model supported by flight sensor data. The dataset is analyzed using statistical methods, and features are designed with the random forest algorithm [33]. The model is expected to reduce aircraft landing accidents and positively contribute to air traffic controllers' air management.

Li et al.,'s study proposes a civilian aircraft big data platform by analyzing the large data generated by the aviation industry Daily [34]. The suggested platform is envisioned to function as a decision support system for aircraft maintenance planning, real-time alarms, health management, fuel savings, and airline scheduling.

Zheng et al., recommend a deep learning approach when classical methods are insufficient for passenger profiling [35]. The proposed method, utilizing a Pythagorean fuzzy deep Boltzmann machine (PFDBM) combined with a hybrid algorithm, develops deep neural networks to classify normal passengers from potential aggressors, showing higher learning capability and classification accuracy than current profilers based on experiments on Air China datasets.

In the study conducted by Garrow and Koppelman, an attempt is made to explain the No-Show behavior [36]. In this context, modeling using Multinomial and nested logit

models revealed that No-Show behavior is best explained by passenger attributes, including e-ticket indicators, booking class, frequent-flyer status, and group size. A summary of the studies conducted in the field of aviation is presented in Table 2.2.

Table 2.2 Aviation studies summary

Author-Year	Title	Category	Methodology
Helgo, (2023)	Deep Learning and Machine Learning Algorithms for Enhanced Aircraft Maintenance and Flight Data Analysis	Aircraft Maintenance and Flight Data Analysis	Review
Adamopoulou & Daskalakis, (2023)	Applications and Technologies of Big Data in the Aerospace Domain	Review of Big Data Applications in Aviation	Review
Xu et al., (2023)	Forecasting stock closing prices with an application to airline company data	Finance	LASSO, Ridge, Decision Tree, SVM, Neural Network
Huang & Cheng, (2022)	Estimation of aircraft fuel consumption by modeling flight data from avionics systems	Fuel Consumption Estimation	Regression Tree, Neural Networks
Gao & Mavris, (2022)	Statistics and machine learning in aviation environmental impact analysis: A survey of recent progress	Aviation Environmental Impact Analysis	Review
Park et al., (2022)	A deep learning approach to analyze airline customer propensities: the case of South Korea	Service Systems	kNN, Decision Tree, RF, XGBoost, CNN, CNN-LSTM
Brunton et al., (2021)	Data-driven aerospace engineering: reframing the industry with machine learning	Review of Big Data Applications in Aviation	Review
Ning et al., (2021)	Applications of deep learning in big data analytics for aircraft complex system anomaly detection	Predictive Maintenance	A Long Short-Term Memory Network Based Auto-Encoders
Jiang et al., (2020)	Applying machine learning to aviation big data for flight delay prediction	Flight Delay Prediction	SVM, Decision Tree, RF, Multilayer Perceptron

Table 2.2 (continued) Aviation studies summary

Author-Year	Title	Category	Methodology
Huo et al., (2020)	The prediction of flight delay: Big data-driven machine learning approach	Flight Delay Prediction	RF, Logistic Regression, kNN Decision Tree, Naïve Bayes Decision Tree, Gaussian Naive Bayes, Logistic Regression, RF, kNN, SVM, XGBoost
Hwang et al., (2020)	Who will be your next customer: A machine learning approach to customer return visits in airline services	Service Systems	
Gui et al., (2019)	Flight delay prediction based on aviation big data and machine learning	Flight Delay Prediction	Random Forest
Oehling & Barry, (2019)	Using machine learning methods in airline flight data monitoring to generate new operational safety knowledge from existing data	Flight Data Monitoring	Local Outlier Probability
Korfiatis et al., (2019)	Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews	Service Systems	Latent Dirichlet Allocation
Kumar & Zymbler, (2019)	A machine learning approach to analyze customer satisfaction from airline tweets	Service Systems	SVM, Artifcial Neural neNwork, Convolutional Neural Network
Ma et al., (2019)	Examining customer perception and behaviour through social media research–An empirical study of the United Airlines overbooking crisis.	Service Systems	SentiStrength Classifier, TF-IDF
Sternberg et al., (2018)	Analysing customer engagement of Turkish airlines using big social data	Service Systems	Naive Bayes
Burmester et al., (2018)	Big data and data analytics in aviation	Review of Big Data Applications in Aviation	Review
Tong et al., (2018)	A novel deep learning method for aircraft landing speed prediction based on cloud-based sensor data	Aircraft Landing Speed Prediction	Longshort Term Memory

Table 2.2 (continued) Aviation studies summary

Author-Year	Title	Category	Methodology
Li et al., (2017)	Civil aircraft big data platform	Big Data platform for civil aviation	Review
Zheng et al., (2016)	Airline passenger profiling based on fuzzy deep machine learning	Service Systems	Pythagorean Fuzzy Deep Boltzmann Machine
Garrow & Koppelman, (2004)	Multinomial and nested logit models of airline passengers' no-show and standby behaviour.	Revenue Management	Multinomial and Nested Logit Models

2.3 Review of Research on No-Show Rate Estimation in Aviation

In the literature, the concept of no-show in the aviation industry is examined based on its impact on overbooking and revenues. The main aim is to determine the optimal overbooking amount (upper limit amount) using statistical models such as Markov chains with the help of historical data. These studies are conducted within the scope of revenue management, taking into account the costs of denied boarding.

There are 3 studies focused on calculating the no-show amount directly by using Machine Learning methods. In the study by Petrović and Milošević, Boolean algebra (IBA) and soft computing techniques were employed [37]. The prediction model for no-show combines case-based reasoning and IBA approaches. In the study, the model is compared with Machine Learning techniques such as ANN. It is explained that the model yields more successful results.

In the study conducted by Zenkert, a No-Show Forecast was made using Passenger Booking Data. Passenger Booking Data includes various data such as ticket purchase time, date, PNR number, Point of Sale (POS), airport, and booking class [38]. Preprocessing processes were followed to make the data compatible for the study, and then Machine Learning methods including CART-Recursive Partitioning, Extreme Gradient Boosting, and Neural Network were applied. While the results of the Neural Network indicate significantly low performance, the other methods yielded the same results as the base model. Zenkert attributes the poor performance of the model to the

randomness of the problem, where a passenger simply doesn't show up sometimes. Although the correlation among these data has been examined, the success of the model remains low due to the fact that the No-Show rate is related to passenger behaviors and external factors, resulting in many statistically insignificant data within the Passenger Booking Data.

In the study conducted by Neuling et al., Passenger Name Records (PNR)-based no-show forecasting was performed. PNR information typically includes passenger name, gender, age, origin and destination details, flight date and time, contact number, and Point of Sale (POS) data [39]. Successful results were achieved using the C4.5 algorithm based on PNR numbers. Furthermore, in the second model, performance significantly improved when exponential smoothing and the PNRPRO model were used together.

CHAPTER 3

BASICS OF BIG DATA & MACHINE LEARNING

Since the Industrial Revolution, rapid technological advancements, the concentration of production tools in specific sectors through capital accumulation, leading to increased mass production, rapid global population growth, and demand for consumer goods, coupled with the globalization and cost reduction of raw materials due to increased production, have collectively elevated living standards for individuals across all segments of society. The acceleration of technological developments, coupled with the globalization of trade, has initiated fierce local and international competition for businesses, necessitating a survival struggle under these intense competitive conditions. The widespread adoption of globalization can be attributed to the removal of barriers to international trade, the shift from global-scale armed conflicts to economic (and technological) supremacy races post-World War II, and the influence of advancing communication technologies and transportation networks.

During this period, the significance of information and technology has swiftly escalated, leading to an exponential increase in the pace of scientific and technological advancements. As a natural consequence, the accessibility of affordable consumer goods has become widespread throughout society, instigating profound shifts in consumption habits, particularly in both developed and developing nations. Simultaneously, advancing and cost-effective transportation options, technological developments facilitating the transport of products to distant regions, international tax and customs facilitations, and the abolition of import and export restrictions have empowered companies to operate on an international scale. This, in turn, has intensified competitive conditions, emphasizing customer relations, product quality, and customer satisfaction in terms of features and quality, thereby heightening competition. Product-centric competitive conditions have compelled companies to constantly pursue innovation, consequently shortening product life cycles. In the relentless competitive environment characterized by local and international competitors, as well as efforts by capital owners to increase low profit margins,

companies have engaged in endeavors to explore alternative markets, business branches, and differentiated product manufacturing. The reduction of global-scale conflicts, coupled with the widespread adoption of new technologies, industries, products, and technologies such as the internet, has increased risk appetite for capital owners globally. Relatively low interest rates have stimulated consumption and investment, thereby creating a mutually reinforcing cycle that accelerates technological advancements.

3.1 Big Data

Throughout human history, the endeavor to produce tools and devices that facilitate life has been persistent, contributing to the advancement of science and technology as knowledge (and technology) spreads at a pace commensurate with communication among diverse cultures and communities. In recent years, the rapid proliferation of communication, particularly through the internet, and advancements in transportation technologies have facilitated access to information, leading to an unprecedented global dissemination of knowledge.

Until the late 20th century, data production was limited and stored in items such as books, newspapers, and magazines, which were both restricted in production and costly. The massive size and limited processing power of early computers made them expensive and difficult to use. Additionally, data storage was limited to small volumes and was very costly. Today, personal computers found in nearly every home have shrunk to sizes that can fit into a pocket, and due to advances in production technologies and economies of scale, they have become accessible to everyone at affordable prices. The annual cost of storing 1 GB of data is less than $0.03, whereas 20 years ago, this cost was over $10,000 [1]. The invention and widespread adoption of computers brought about new tools for data production and storage, followed by the advent of the internet. The widespread availability of broadband internet and various services emerging in recent times, such as Video on Demand (VOD), cloud storage and services, and data sets, have resulted in the daily generation of substantial amounts of digital data [40].

The digitization of all aspects of life, facilitated by Industry 4.0 and the Internet of Things (IoT), continuously amplifies data production. Presently, smart devices (such as mobile phones, computers, smart home appliances, and automobiles—essentially, all devices containing computers) generate, collect, and store vast amounts of data every day. Approximately 402.74 million terabytes of data are generated in a single day [3]. The security and privacy of the data produced have become paramount concerns in today's context, extending from the moment of generation to its destruction [41-43].

A short while ago, data analysis was confined to limited tables and databases. However, thanks to today's technology, it is now possible to collect, store, and work with different types of data that cannot be organized within rows and columns [44]. The immense volume of data generated and stored provides a colossal resource that can be utilized by scientists, experts, researchers, and companies. Aktan's account indicates that the concept of Big Data was first coined by Michael Cox and David Ellsworth, who highlighted the overwhelming size of data sets that filled all storage devices [45]. Subsequently, in the literature, Big Data has been frequently defined as massive and complex data sets [46]. Sagiroglu and Sinanc define Big Data as “Big data is a term for massive datasets having large, more varied and complex structures with the difficulties of storing, analyzing, and visualizing for further processes or results” [47].

In the literature, the fundamental characteristics of Big Data are defined as Volume, Variety, and Velocity, collectively referred to as the '3Vs' [46-59, 80]. To this definition, Veracity has been added to form the '4Vs' [60-62], and further, with the inclusion of Value, it has been described as the '5Vs' [63-74], as shown in Figure 3.1.

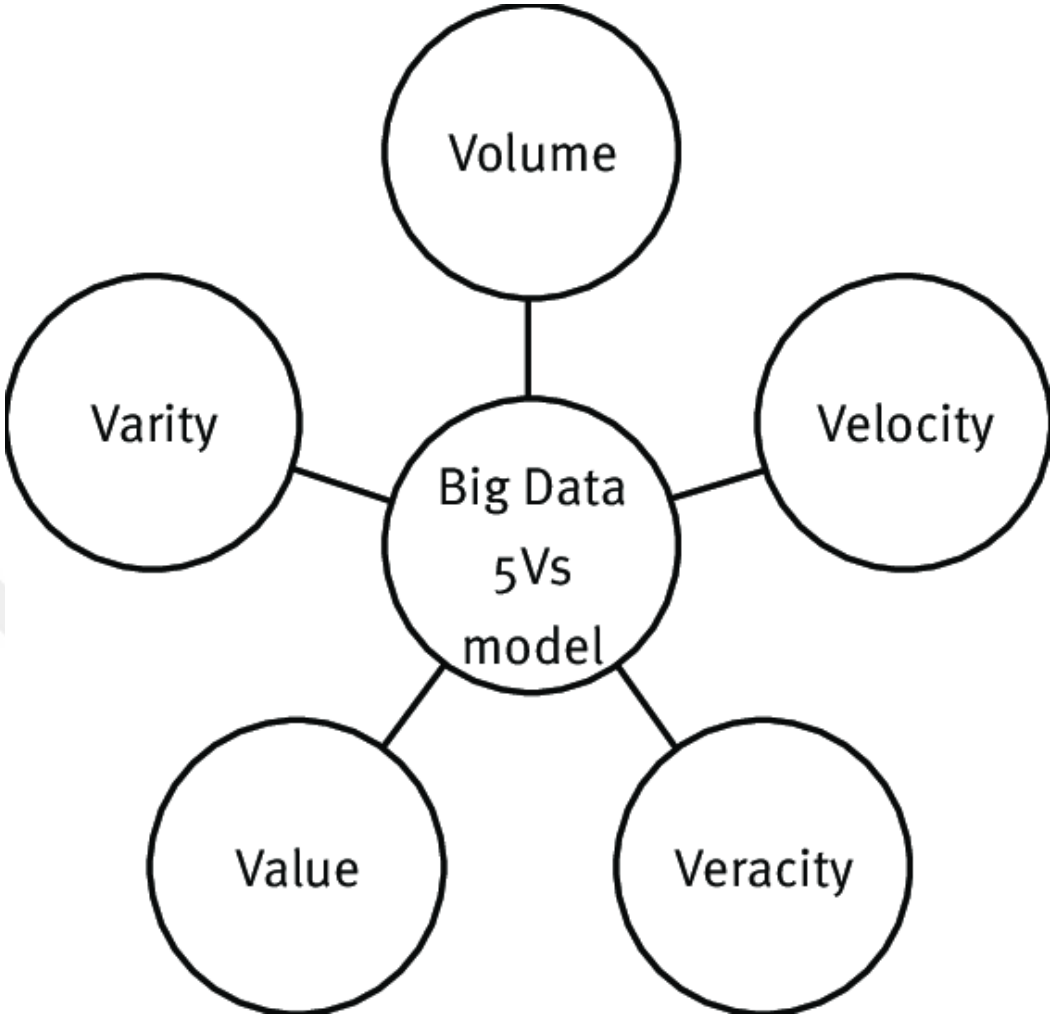

Figure 3.1 5V's of Big Data

Variety: Refers to the variety of data types, including images, text, videos, audio, etc. [45, 59, 63-64]. Data types can be categorized into three types: structural, semi-structural, and non-structural [45]. Structural data denotes information that can be expressed in databases through rows and columns. Semi-structural data is defined as information that can be placed into data tables without adhering to a specific format and can be categorized or labeled according to specific orders. Non-structural data refers to information that cannot be organized into rows and columns in databases and lacks specific formats and arrangements (examples include social media data, images, and audio).

Volume: The magnitude of globally generated and stored data is quantified in zettabytes. Internet usage spans across various domains, from social media

engagement to browser searches, image and video sharing, and communication. This implies that billions of individuals worldwide continually produce data through smart devices connected to the internet. Additionally, corporations, governments, scientific organizations, and others store all accessible data for their respective purposes. This situation leads to the formation of an enormous data heap, necessitating expertise, advanced technologies, and strategies for its storage, utilization, and processing.

Velocity: The technological advancements that lead to the generation of a vast amount of data are simultaneously both the cause and the consequence of the speed of data creation. With the widespread adoption of technology and the internet, more individuals contribute to data generation, and the pace of data creation accelerates. The proliferation of the internet, coupled with the increased speed through fiber optic cables, has heightened social media usage, online shopping, and all forms of internet-based communication, consequently increasing the velocity of data creation. The culture of remote work, facilitated by the internet, has further contributed to the surge in data production, particularly during the pandemic. This rapid and extraordinary process witnessed over the last two decades demands advanced technology for the processing of the substantial volume of generated data.

Veracity: Big Data veracity addresses the accuracy of data through two fundamental dimensions: consistency and reliability [74]. Ensuring the value of data and facilitating the analysis and utilization of big data relies on the presence of data of superior quality [64]. The quality of data is a crucial prerequisite for establishing models, conducting statistical analyses, discovering patterns, and engaging in machine learning and artificial intelligence research, ensuring consistent and meaningful results from these models. Multinational corporations, governments, and organizations of all sizes make decisions based on the outcomes derived from Big Data. Therefore, it is imperative for the data to be consistent and reliable. The reliability of data entails ensuring its integrity from the source of acquisition through storage until its utilization, safeguarding it against corruption. Data consistency refers to the data's accurate representation of the pertinent information related to the subject of the study. If data is incorrect, incomplete, nonsensical, or unreliable, the results will be untrustworthy. In situations where data accuracy is compromised, decisions made by companies,

institutions, or organizations can lead to both financial losses and damage to their reputations.

Value: The data itself is not widely accepted as a singular value or holds very limited significance. For Big Data, the collective gathering, processing, and deriving a result that directly or indirectly adds value or benefit from numerous pieces of data are essential. This is a significant characteristic that distinguishes Big Data from any ordinary data. The value of Big Data is not inherent in individual data points but emerges through the collective analysis of events or processes, whether they exhibit stochastic, probabilistic, regular, or random patterns [74].

Big Data is increasingly being utilized by companies, particularly those with highly competitive prowess, to create value. Offering significant opportunities in areas such as customer experience, operational efficiency, innovative product and service development, marketing and sales, risk management, supply chain management, financial analysis and forecasting, and social media analytics, Big Data has become a significant asset and opportunity for decision-makers through established models. For these models to yield accurate and meaningful results, data must be processed correctly. Therefore, proper data preparation is essential before modeling. Data preprocessing is the most crucial aspect for obtaining results from Big Data.

3.2 Data Mining

Data mining, also known as Knowledge Discovery in Databases, is the process of extracting meaningful information from large data sets in databases or data warehouses. This process aims to generate value by leveraging various disciplines such as machine learning, deep learning, artificial neural networks, statistics, computer applications, and visualization to uncover previously unnoticed patterns and relationships [75]. In data science projects, the most commonly used technique for project processes is the Cross Industry Standard Process for Data Mining (CRISP-DM), published by Chapman et al. [76]. According to a survey, 49% of participants in data science projects prefer the CRISP-DM process, 18% choose the Scrum process, 12% opt for the Kanban process, and 12% follow their own process (MyOwn), while 4% prefer the Team Data Science Process (TDSP) [77].

In the CRISP-DM framework, Data Processing starts with Business Understanding and Data Understanding [75]. It continues with the collection of data from various sources, involving high variety, volume, and velocity. The initial challenge encountered in the context of big data pertains to the inadequacy of traditional data management systems to efficiently transfer and particularly store data of such magnitude [81,82,86]. Achieving quality results is contingent on high-quality data; using inaccurate or incomplete data in model construction may yield incorrect results and compromise the accuracy of the analysis. The raw data, as it is initially collected, is frequently inappropriate for direct input into models, requiring preprocessing for suitability, commonly referred to as Data Cleaning [75,79]. This stage involves various operations such as missing value imputation, noise data handling, dimensionality reduction, and feature selection [80]. After the Data Preparation phase, the subsequent stages include Modeling (where various techniques are applied and calibrated), Evaluation (focused on thoroughly assessing the model's quality and alignment with business objectives), and Deployment (involving the organization and presentation of gained knowledge for practical use, potentially in real-time decision-making processes or repeatable data mining procedures across the enterprise) [75]. A visualization of the CRISP-DM technique is shown in Figure 3.2.

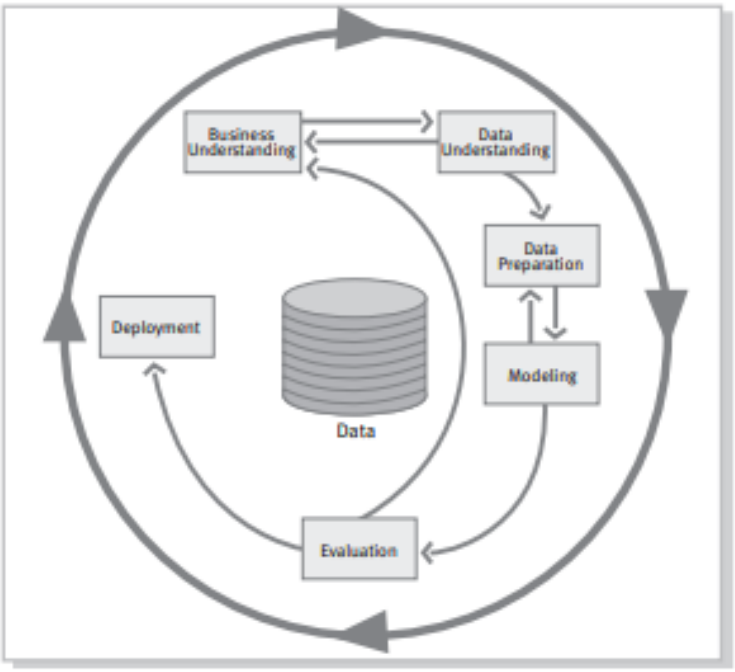

Figure 3.2 CRISP-DM technique [76]

3.2.1 Big Data: Technologies and Tools

The original definition of big data emphasized structured data, but the evolving understanding acknowledges that a substantial amount of contemporary data lacks innate structured formats [81]. For instance, tweets and blogs represent weakly structured textual content, while images and videos are structured for storage and display but not designed for semantic content and search. Transforming such diverse content into a structured format for subsequent analysis poses a significant challenge [82].

With the widespread adoption of internet and computer technologies today, vast amounts of raw data are generated daily. According to research by Kaisler et al. in 2013, the disk technology limits at that time were about 4 terabytes per disk [81]. As of 2024, despite increasing advancements, the daily data production has reached approximately 402.74 million terabytes. However, storage capacities still lag behind the pace of data production. Although the cost per unit of storage has significantly decreased over the years, the rate of data generation has increased even more rapidly, maintaining its status as one of the most critical contemporary issues.

The initial framework facilitating the processing of extensive datasets was MapReduce [83]. Later, Apache introduced Hadoop, a pivotal technology in the realm of big data, offering an encompassing solution for systematic data management, including storage, processing, system administration, and module integration. Comprising HDFS (Hadoop Distributed File System) and the MR framework (MapReduce Framework), Hadoop effectively addresses challenges associated with massive datasets [75]. The emergence of new technologies and services, such as cloud computing, and the reduction in hardware prices are leading to an ever-growing rate of information on the Internet [84].

1. Understanding the Business
2. Data Understanding
3. Data Preparation

4. Modeling
5. Evaluation
6. Deployment, Positioning

3.2.2 Data Preprocessing

Data preprocessing is an integral stage in the data mining process, involving tasks such as data preparation, integration, cleaning, and transformation, essential for refining the quality and representation of data to enhance the accuracy and effectiveness of subsequent data mining algorithms. This essential step involves mitigating issues like out-of-range values, missing data, and irrelevant information, ensuring that the resulting dataset is reliable and suitable for efficient data mining algorithms. By addressing these issues, data preprocessing contributes to improved predictive accuracy and facilitates various learning tasks, such as classification, regression, and unsupervised learning [85].

In the literature, methods applied during the data preprocessing stage are generally grouped in similar ways and categorized under three main headings: Integration, Cleaning, and Redundancy Elimination [75, 82, 84-86], Also, there are studies that specifically address each process in the data preprocessing stage separately [87].

3.2.2.1 Data Cleaning

The majority of techniques in data mining assume the availability of datasets that are considered either complete or free from noise [84]. However, in everyday life, most raw data contains many errors. Data cleaning can be divided into two categories: handling missing data and managing noisy data.

3.2.2.1.1 Handling Missing Data

Missing values (MV) pose a common challenge in data mining, contradicting the assumption of dataset completeness. These values can result from faulty sampling, cost constraints, or acquisition limitations, leading to poor knowledge extraction and erroneous conclusions. Handling MV is intricate, with discarding instances often

proving suboptimal as it risks biasing the learning process and discarding crucial information [85]. Statistical approaches, rooted in probability function modeling and maximum likelihood procedures for imputation, have been foundational. Machine learning techniques, known for their flexibility, are increasingly favored for MV handling, offering effective solutions without requiring prior information [75].

Missing values (MV) are a common challenge in data mining and are vital for maintaining data consistency [75]. These values can be due to incomplete data recording, loss during transfer, human or software errors, faulty sampling, or cost constraints, leading to the absence of valuable information in the dataset. Handling MVs in a dataset is complex. The initial approach for dealing with MVs is to discard instances containing MVs [85]. However, this can result in the loss of valuable data and introduce strong biases if MVs exhibit a pattern [80]. Aside from discarding MVs, another approach is to impute the missing values, which fundamentally divides into statistical methods and machine learning (ML) methods. Statistical approaches are based on probability function modeling and maximum likelihood procedures for imputation [85]. Some methods for MV imputation include Maximum Likelihood Imputation Methods, Expectation-Maximization (EM) Algorithm, Multiple Imputation (MI) approach, and the MV estimation method based on Bayesian Principal Component Analysis (BPCA). Machine learning techniques, known for their flexibility, are another method for MV imputation. These techniques are increasingly preferred for MV imputation due to their ability to provide effective solutions without requiring prior information [80].

3.2.2.1.2. Noisy Data

The prevalence of noise in data poses significant challenges in classification tasks, affecting data collection and preparation in Data Mining applications, ultimately impacting model performance. Accurate solutions to noise-laden problems often require specialized techniques, especially for noise-sensitive scenarios [85].

In supervised scenarios, noise can impact input features, output values, or both. When noise affects input attributes, it is termed as attribute noise, while its impact on output attributes, known as class noise, can introduce significant bias, especially in

classification tasks [80]. Two main approaches are commonly employed to address noise in data mining. The first involves correcting noise using data polishing methods, particularly if it impacts instance labeling, albeit challenging and typically restricted to small noise levels. The second method entails utilizing noise filters to identify and eliminate noisy instances from training data without modifying the data mining technique [80, 84, 85].

3.2.2.2 Data Reduction

Unnecessary repetitions or excess data can lead to decreased data reliability or inconsistency of results, cause time loss for algorithms, and increase costs during storage by occupying unnecessary space. Especially, unstructured data occupies a large volume in terms of size and increases the cost of storage, while processing it leads to both time loss and increased costs. Data reduction techniques aim to condense large datasets to make data mining more feasible and efficient. These methods involve tasks like removing or grouping data based on attributes and examples, simplifying data domains, and addressing the "curse of dimensionality." By providing a reduced representation of the dataset while preserving its integrity, data reduction enables more efficient mining processes without sacrificing accuracy. The techniques are broadly categorized into dimensionality reduction, sample numerosity reduction, and cardinality reduction [85].

3.2.2.2.1 Dimensionality Reduction

The effectiveness, applicability, and reliability of models decrease when the number of data points, or in other words, the number of independent variables, is relatively high compared to the size of the data. The increase in input dimensionality exponentially expands the search space and concurrently heightens the likelihood of obtaining erroneous models. It is known in the context of achieving reliable and high-quality outcomes that there exists a linear relationship between the training set and dimensionality. In the case of non-parametric learners, however, as the number of independent variables increases, it is anticipated that the sample size will exponentially grow [85].

3.2.2.2.2 Feature Selection

In data mining processes, having more features does not guarantee obtaining a higher quality model and results. Firstly, having a larger input dimension increases both costs and time spent on searching for training data. This not only leads to additional time and cost in the data collection processes but also increases storage-related expenses. Feature selection is the process of selecting an optimal subset of features. Avoiding unnecessary features from being included in the model, starting from the data collection stage, contributes to reducing costs, decreasing the search space of the model, enhancing the reliability and quality of results, preventing the risk of overfitting in algorithms, reducing model complexity, and increasing the accuracy of interpretations of results.

3.2.2.3 Feature Extraction

Apart from feature selection, another method applied for dimensionality reduction is feature extraction. It is also referred to as space transformations [80]. Instead of feature selection, the focus is on creating a new dataset by combining some features from the original dataset [84]. Feature extraction can be applied under suitable conditions depending on the subject of data mining, the dataset being worked on, and the relationship between independent variables. However, the process of combining data can adversely affect the quality of the model and the accuracy of the results in most data mining processes.

3.2.2.4 Instance Reduction (IR)

Sample reduction is considered a complementary task to feature selection. It focuses on reducing the input size by reducing the amount of data in large datasets or by creating new data samples without compromising the quality of the model and the reliability of the results. The selection of a subset of instances to represent all instances in data mining applications is called Instance Selection. Instance Generation, on the other hand, refers to the artificial data generation methods applied when there is a lack of representation or insufficient representation in a field [84, 85].

3.2.2.5 Discretization

Different datasets work with different data mining algorithms. There are learning techniques that work with numerical datasets, such as statistical learning techniques, as well as learning techniques that work with categorical datasets. Similarly, there are learning techniques that work with continuous variables or discrete variables. Therefore, to obtain accurate results with high-quality models, the model selection should be compatible with the dataset used. Discretization is the process of converting numerical data into categorical data. Thus, a relationship is established between each numerical value and a certain range. As discretization reduces data from a large numerical value domain to a subset of categorical values, it is considered a data reduction mechanism [88].

3.3 Importance of Big Data

Although data itself is not a recent discovery, advancements in statistics and computer science have revolutionized our lives by enabling the use, processing, and valuation of data, giving rise to the concept of big data. In various sectors such as retail, healthcare, sports, information technology, finance, agriculture, automotive, space, and beyond, big data is utilized in areas such as supply chain, after-sales service, customer relations, engineering, production, R&D, marketing, quality, logistics, maintenance, and repair, as well as various forecasting and analytical domains. For instance, during the 2012 elections, President Obama leveraged big data to analyze voters, conduct specific campaigns to increase voter turnout in his favor, and influence undecided voters [44]. The global trade volume, which was 383 million dollars in 1970, has surged to 31.37 trillion dollars in 2022 [89]. Examples like these emphasize the growing importance of digitization and leveraging big data for companies striving to gain a competitive advantage in the globalized and challenging business environment. Leading organizations invest in innovation to continuously exploit the growing opportunities to collect new data, integrate external and internal data, and apply big data and analytics to rapidly adapt to changing conditions and outperform competitors [90]. In 2017, The Economist stated that 'The world’s most valuable resource is no longer oil,

but data' [91]. At present, the big data market is worth more than 55 billion dollars of revenues and it is expected to grow to 103 billion dollars by 2027 [92].

Big corporations, scientists, and experts are turning to the use of big data to respond in real-time to changing conditions, formulate accurate strategies, predictions, and forecasts, optimize business processes, innovate with new products and solutions, manage and improve human, machine, and raw material resources more efficiently, make more efficient decisions, and gain different perspectives on surrounding events with previously undiscovered patterns. For businesses, the generated data is crucial in three areas: improving decision-making processes, enhancing operations, and monetizing data [44]. To handle the generated or digitized data effectively, it is essential to develop strategies that can accurately store, process, and create value from it. This has led to the emergence of a specialized field known as 'data science' in today's world. A data scientist is defined as "At a high level, data science is a set of fundamental principles that support and guide the principled extraction of information and knowledge from data" [93]. Data scientists extract meaningful information from large datasets using complex algorithms and artificial intelligence models. Algorithms analyze data piles, allowing for more accurate decision-making and more efficient use of resources. This improves the quality of life for both individuals and society, and increases productivity and economic efficiency. In this context, the value and importance of data scientists and the algorithms they develop, and the value they add to human life, are increasing day by day.

All of this has been made possible by technologies that enable the rapid generation of enormous heterogeneous data and the production of value from it. Thanks to exponential increases in computing power and big data, many developments have been realized, from software used in drug discovery to AI-equipped robots that automatically perform chemical analyses, smart assistants and algorithms that predict our interests and make personalized recommendations, gene sequencing, and robots. Many of these algorithms utilize data like digital footprints [1]. Obtaining useful information by drawing meaningful conclusions from big data and discovering patterns statistically, as well as ensuring that machine learning or artificial intelligence

models yield accurate results, are as important as defining the right strategy. A reliable dataset and data preparation suitable for the models to be established are also crucial.

3.4 Overview of the Aviation Industry

The development of technology has had striking effects in many areas, including transportation. Following advancements in security, the first civil aviation initiatives began. Compared to land, sea, and rail transportation, which took a long time for all types of transportation and travel activities, air transportation took much less time, and the demand for it became almost unlimited during this period. In response to the increasing demand, entrepreneurs in other countries also tried to engage in airline transportation activities, sometimes with state support and sometimes entirely with private companies.

Throughout history, the airline transportation industry has experienced many ups and downs. Since the aviation industry is a capital-intensive sector, it has naturally been a sector with competitive moats. Moreover, security concerns from the earliest civil aviation companies to the present have necessitated strict audits and regulations.

Regulations have been implemented in many countries worldwide in the aviation industry. For example, after the Great Depression, legal regulations were introduced in the US to form a government-regulated cartel in the airline industry, aiming to limit competition and prevent bankruptcies. These regulations, allowing airlines to reflect cost increases in prices, removed the incentive to seek cost-reduction measures and caused airlines to focus on satisfying regulators rather than having sufficient flexibility to meet customer demands [94]. Similar regulations existed in regions such as the European Union and Southeast Asia. These regulations, which limited competition and led to inefficiencies in airlines, were first repealed in the US by Congress with the Airline Deregulation Act of 1978 [95]. Similar steps towards liberalization in aviation were taken in Turkey with the Turkish Civil Aviation Law [96] in 1983 and at various times in the 1980s in the European Union [97,98]. With the increase in deregulation, competition increased worldwide, leading to a decrease in transportation costs and a skyrocketing demand. However, high growth often resulted in low profits and did not translate into financial success [99].

While there are many reasons for this, one of the main reasons is that the service sector, particularly the aviation sector, is more sensitive to micro and macro factors than other sectors. The 1973 oil crisis, known in history as the crisis that dramatically affected global oil prices, was the oil embargo declared by the Organization of Arab Petroleum Exporting Countries (OAPEC). The resulting cost increases caused a recession worldwide. The aircraft orders advanced due to the increasing demand before the crisis later led to idle capacity problems. In subsequent periods, the emergence of low-cost carriers and increasing competition reduced airline profitability. Furthermore, the aviation industry, being much more capital-intensive than other sectors, is significantly affected by macro variables for working capital and fixed capital (aircraft) investments needed for growth. However, one of the most significant issues for airlines is the impact of micro variables in addition to macro variables with global effects. The first hijacking in history occurred in 1961 in Peru [100]. Later, in events like the September 11 attacks, which would have significant global impacts, planes were used, affecting flights and air transportation periodically. Apart from terrorist attacks, factors such as the inability to use airspaces due to regional impacts, disruptions in countries affecting tourism activities, exchange rates, jet fuel prices, geopolitical risks in regions with oil refineries, local economic crises, global pandemics (Swine Flu, Bird Flu, Covid-19), restrictions on international transportation activities, and many other factors affect the profitability of the aviation industry much more than other industries.

Within the framework of these risks, to counter the increasing competition, aviation companies are also trying to protect their profitability by developing many new marketing activities and cost-reducing methods. Big data is used in many stages of aviation. Rolls-Royce benefits from big data connected to IoT. With the help of sensors embedded in the engines produced, it regularly monitors the performance of over 3,700 jet engines worldwide, tracking issues before they arise. This provides a significant cost advantage by preventing the need for expensive equipment replacement every 3 or 6 months with a dynamic maintenance option, thus preventing unexpected parts or engine failures, contributing to increased flight safety [44]. General Electric estimates that a 1% fuel savings from using big data in aircraft engines would save the commercial airline industry $30 billion over 15 years [101].

Another area where big data is used in the aviation industry is Revenue Management (RM). Assigning seats and managing overbooking are critical components of yield management, with profitability contingent on effectively controlling both of these factors [102, 103]. Airlines use RM systems for optimal fare classes (or booking classes) [104]. Airlines try to maximize their revenues through different ticket pricing, whether for discount or business seats [105]. They use big data to predict potential demand with the help of ML algorithms by more accurately analyzing passenger behaviors and flight statistics [106-108] and determining the optimal number of seats and prices to offer in response to demand [109-112]. Therefore, the accuracy of demand prediction is crucial for flight profitability. Forecasting models can be quickly simulated and compared with today's computational power, but critical steps in data preprocessing such as data structuring, unconstraining, and aggregation play a crucial role in generating complex predictions [113]. The negative impact of service system disruptions on customer satisfaction and company profitability is also being mitigated with the help of big data. By predicting potential flight delays within this scope [26, 114-117], airlines try to minimize possible costs and maintain customer loyalty by increasing passenger satisfaction throughout the flight.

3.5 No-Show Rate Prediction on Airline Industry

Aviation faces profitability issues due to increasing competition, price pressures from low-cost carriers, the inability to immediately reflect cost fluctuations in prices, and problems created by international flight restrictions due to recent pandemics. These problems necessitate more comprehensive marketing and revenue management strategies for airlines. Many different methods are used to reduce flight costs and increase revenues. One of these methods is the Over-booking strategy used against the no-show probability of passengers.

Overbooking is defined as selling more tickets than the maximum capacity to mitigate risks of empty seats due to situations such as passengers canceling their flights (free of charge or at a very low cost) or failing to show up for their scheduled flights without prior notice. Airlines aim to minimize the risk of empty seats against the demand generated for flights, thereby avoiding missed revenue opportunities and maximizing

their earnings. The overbooking method is also frequently employed in hotels [118-122], restaurants [123], and medical clinics [124-126].

Finding the optimal level of overbooking in the aviation industry is highly crucial. The primary reason for this is the potential occurrence of denied boarding when all passengers holding tickets arrive, exceeding the aircraft's capacity, which can lead to substantial costs. Denied boarding is the responsibility of the airline, and passengers' rights for their flights are fully protected. Rothstein proposed a model in his study to balance the potential loss of revenue due to empty seats against the costs incurred from denied boarding (such as refunds, passengers opting for rival airlines, accommodation and dining expenses, and goodwill loss), aiming to enhance profitability through optimal booking policies [127]. Moreover, in today's context, along with all these costs, there is increasing attention to additional costs that companies may have to bear due to heightened social media usage, whose impacts have not yet been fully modeled [36]. Furthermore, research on passengers' no-show behaviors describes factors influencing no-shows and standby behaviors, which can provide guiding principles for airlines [128].

In the study conducted by Klophaus and Pölt, it is noted that Lufthansa German Airlines experienced 4.9 million no-show passengers in 2005 [129]. According to the study, through overbooking, Lufthansa was able to carry more than 570,000 additional passengers in 2005, resulting in €105 million in additional revenue (after deducting denied boarding costs). Considering the increase in passenger numbers and flights since 2005, these figures can be estimated to have increased by 3-4 times in today's context.

Despite being a valuable source of revenue for airlines, there are relatively few studies on overbooking in the literature. The primary reason for this scarcity is attributed to airlines' reluctance to share data due to its high confidentiality (marketing value). The few existing studies in the literature are divided into two categories: those calculating no-show probabilities to forecast passenger behavior for flights [37-39], and those setting upper limits for optimal overbooking under Revenue Management (RM) to maximize profitability [130-137].

3.6 Machine Learning

Machine learning is a subfield of artificial intelligence that focuses on the development of algorithms and statistical models enabling computers to perform tasks without explicit instructions. These algorithms utilize patterns and inference to progressively improve their performance on a given task through experience, thereby allowing them to make predictions or decisions based on data rather than relying solely on pre-programmed rules. The various machine learning algorithms and their classifications are illustrated in Figure 3.3.

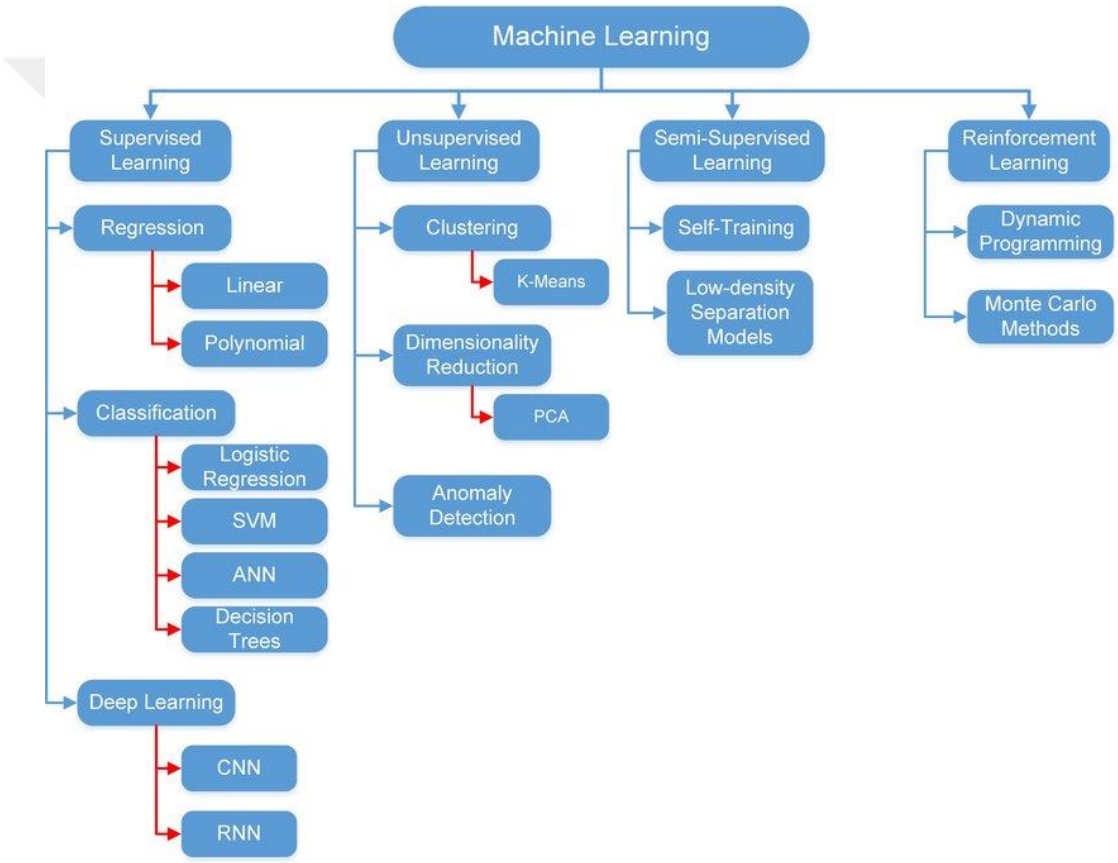

Figure 3.3 Machine Learning algorithms [225]

An algorithm is defined as a sequence of commands executed to transform input into output [138]. Algorithms can range from simple tasks like sorting data according to a specific criterion using a computer to solving extremely complex problems such as winning a game like Go, which involves numerous options and incomplete information [139]. Classical statistical models and deterministic mathematical models

rely on precise numerical data. However, ready-made algorithms do not exist for many modern-day issues crucial to businesses, such as predicting customer behavior or identifying unwanted emails [138].

In classical statistical problems, the information and data available are used to generate inputs, which produce outputs within a certain confidence interval. Today, researchers work on algorithms that operate under constraints such as incomplete descriptions of input relationships, processes without clear causal calculations, and scenarios with numerous missing data, assumptions, and observations. These algorithms strive to produce results under conditions where full computation is either impossible or very difficult, adapting to rapidly changing conditions and addressing both numerically describable and indescribable constraints [140].

When the output is categorical, it is termed classification; when numerical, it is regression. Clustering, an explanatory modeling technique, involves assigning similar observations to the same clusters. Association rules uncover interesting connections between observations. While classification and regression methods are considered supervised learning as they require training data, clustering and association rules fall under unsupervised learning as they do not need labeled training data. Semi-supervised learning combines labeled and unlabeled data, bridging supervised and unsupervised methods in classification and clustering problems. Reinforcement learning, on the other hand, is a machine learning algorithm outside these categories, enabling software agents or machines to evaluate optimal behavior automatically based on feedback in the form of rewards or penalties [140]. The working principle of classification and regression on the data set is visually analyzed and presented in Figure 3.4.

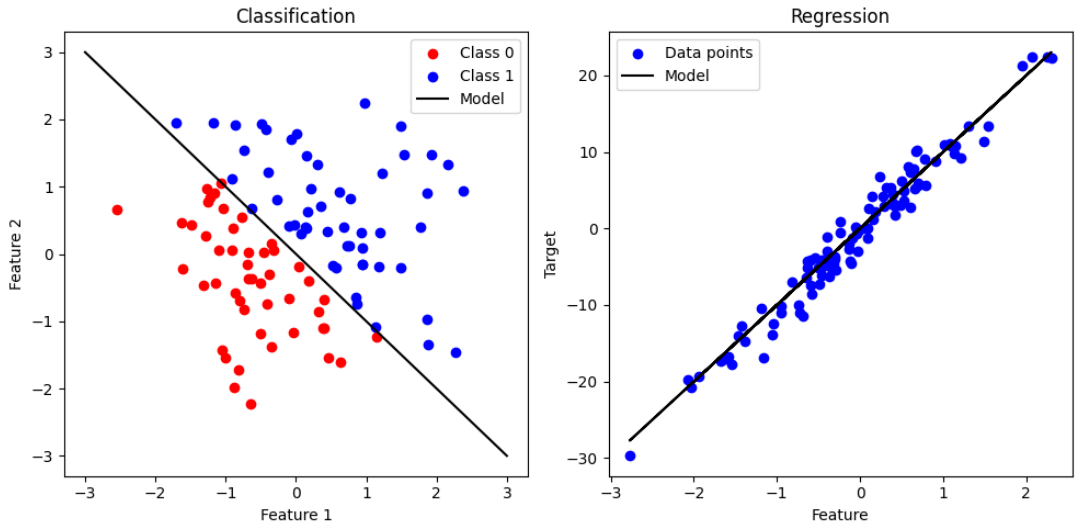

Figure 3.4 Classification and Regression models

3.7 Types of Machine Learning

Machine Learning is the general term for computational algorithms that mimic human intelligence by learning from surrounding events (data) to derive meaningful conclusions [141]. These algorithms can perform complex and challenging calculations more effectively than traditional statistical methods. Statistical methods typically require precise data and numerous calculations to discover relationships between variables, whereas machine learning models can identify patterns (relationships) quickly without the need for explicit programming [142]. In simple terms, while statistics derive insights about a population by analyzing a sample, machine learning identifies generalizable predictive patterns.

Machine learning is successfully applied in various fields such as pattern recognition, computer vision, medicine, finance, biotechnology, manufacturing processes, decision-making based on predictions, analyzing customer behaviors, personalized product development, logistics, natural language processing, and speech recognition [144].

Machine learning methods are broadly categorized into three types: supervised learning, unsupervised learning, and semi-supervised learning. Recent advancements also include deep learning and reinforcement learning [145]. Supervised learning

involves learning from labeled data, while unsupervised learning aims to explore the inherent structure within unlabeled data. Semi-supervised learning utilizes both labeled and unlabeled data to enhance the learning process.

Deep learning, as an advanced form of artificial neural networks (ANNs), operates with more complex structures and typically performs better with closed data sets compared to other techniques [147]. Classification and clustering are fundamental tasks in data mining. Classification is used in supervised learning methods, whereas clustering is applied in unsupervised learning. Clustering aims to identify groups of inputs (variables), while classification is predictive in nature. In clustering methods, new groups are formed, and the number of groups defines a descriptive quality for each variable within a group. In classification, the evaluation is based on a predefined class set, aiming not to create new descriptive groups as in clustering.

Regression, on the other hand, is a statistical tool used to explore relationships between variables (dependent and independent). In regression analysis, the aim is to discover causal relationships between the dependent variable and the independent variables, predicting the effects of independent variables on the target variable [149]. In clustering and classification models, the target values are categorical, while in regression models, they are numerical (continuous) values.

In supervised machine learning, the effectiveness of algorithms can be measured by testing the model's compatibility with the basic logic of the algorithms. It is important to test the performance of the model with new data that has not been seen before, to measure the effectiveness of the model. Generalizability in machine learning refers to the ability of the model to perform well on new, unseen data sets, in addition to the performance shown in the training data set [151]. Overfitting and underfitting in machine learning describe two basic problems encountered when evaluating the performance of a model.

Overfitting or excessive fitting means that the model memorizes the data rather than understanding the relationship (causality) between the target variable and the independent variables [152]. In this case, the model may fail to make accurate predictions and perform poorly on test data, even though it performs well with training

data. The problem of overfitting in machine learning models increases when the training set is uniform. In general, even if the model performs well with the training data, it will have a high error rate in the test data, which has not been memorized. Overfitting is a problem in machine learning models with high variance and low bias.

Several strategies are recommended to mitigate the issue of overfitting in machine learning models:

i. **Removing Noisy Data:** To avoid overfitting caused by noisy data, it is advisable to exclude noisy data from the dataset.

ii. **Adding More Training Data:** Increasing the amount of training data helps improve the model's ability to generalize. This strategy enhances the model's exposure to diverse examples, reducing overfitting tendencies.

iii. **Generating Synthetic Data:** Generating new data based on the distribution of existing data can enhance the diversity of the dataset. This approach helps in training the model on a more comprehensive range of scenarios, thereby improving its robustness.

iv. **Regularization:** By reducing the weight of excessively influential variables in the model, their impact can be minimized [153].

Incomplete learning (underfitting) occurs when the model fails to fit the training data adequately, unable to capture the variability in the dataset [154]. This inadequacy stems from using a overly simplistic model to describe complexities in the dataset effectively. In other words, insufficient independent variables lead to decreased model generalizability. Models suffering from incomplete learning exhibit high error rates on both training and test datasets, resulting in poor overall performance. These models have low variance and high bias. They are unable to learn the causality or patterns between independent and dependent variables in the training data.

In machine learning, poor model performance, or alternatively overfitting or underfitting, occurs due to the model's high bias or high variance. Bias and variance are critical metrics used to assess a model's generalization ability and fit to the training

dataset. Bias refers to how much the predicted values deviate from the true values, indicating the model's inability to learn the relationships between variables. Biased models typically have fewer parameters or lower complexity, making them more likely to experience incomplete learning in the training data. Variance, on the other hand, expresses how much the predicted values vary across different datasets. High variance, which indicates the model's sensitivity to minor variations in the training data, leads to low generalization ability. It shows that a model with high variance fits the training data too closely, potentially resulting in high errors (poor performance) when predicting on new datasets [155]. Bias relates to the simplicity or inadequacy of the model, while variance relates to its complexity and overfitting to the training dataset. An ideal model should have both low bias and low variance. This ensures that the model fits the training data well while making accurate predictions on new data. A visual representation of the models resulting in underfit, good fit, and overfit is presented in Figure 3.5, each plot includes data points (black dots) and the corresponding model fit (red dashed line).

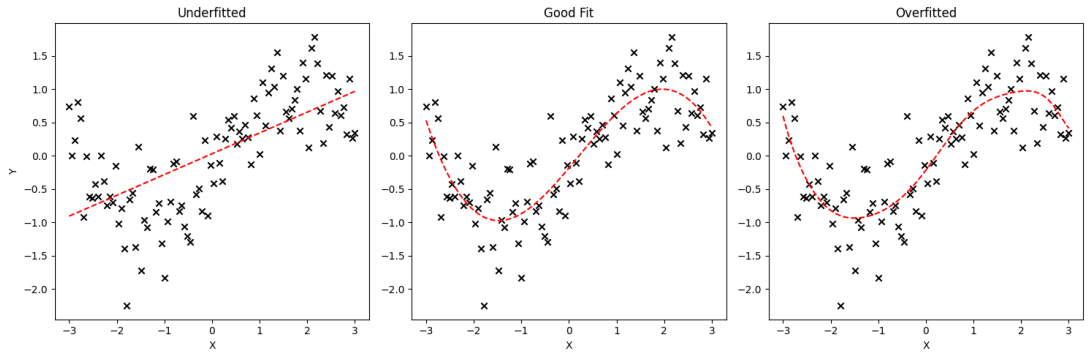

Figure 3.5 Overfit, underfit and good fit models

3.7.1 Supervised Learning

Supervised learning is a significant subfield of machine learning, where the learning process of a model is conducted using labeled training data, consisting of measured features of data and their corresponding labels. Supervised machine learning can be considered a powerful tool in the age of data for extracting information from large datasets, making predictions using this information, automating learning processes, and reducing human errors in decision-making processes. As a result, it enables

businesses and organizations to adopt more data-driven and informed approaches in decision-making, leading to more accurate and consistent decisions. Many businesses currently use supervised machine learning models to understand customer needs and offer personalized product and service recommendations, achieving competitive advantages through successful outcomes. In supervised learning, each example in the dataset is paired with a label that represents its correct output. Commonly used for tasks such as classification or regression, supervised learning aims to predict specific outputs. During the learning process, the model learns a certain relationship between input data using the training data, typically establishing a function that maps data to labels. Therefore, the dataset and the expected outputs from this data are known [156].

Supervised machine learning methodology can be academically examined in three main stages: foundational theoretical development, discovery of practical algorithms, and increase in computational capacity.

Supervised machine learning emerged at the intersection of computer science and statistics. The foundations of this field can be traced back to the 1950s, with Alan Turing's work on "machine learning capabilities" [157]. Early algorithms were also shaped by researchers like McCarthy and Arthur Samuel [158].

Basic statistical concepts (such as regression and classification) and probability theory formed the basis of supervised learning models during this period. For instance, techniques like linear regression and the method of least squares have roots dating back to before the 1900s, but gained significant popularity in the field of computer science during this era [159].

The 1980s and 1990s witnessed a rapid increase in data availability due to the widespread adoption of the internet and computers. Technological advancements, coupled with the increase in computational power, facilitated the development of supervised learning algorithms. While early studies on new algorithms like artificial neural networks, support vector machines, decision trees, and Bayesian classifiers date further back, their practical development and widespread adoption aligned with this period.

Since the early 2000s, there has been a revolution in the collection and storage of large datasets, greatly influenced by the rapid proliferation of technology. During this period, particularly with the explosion of data on the internet and the widespread adoption of cloud computing technologies, the computational capacity required for training machine learning models has significantly increased.

Deep learning, facilitated by the growth in GPU-based computing power, has notably advanced during this era, achieving successful results in neural network-based analysis of large and complex datasets. Deep learning stands out for its ability to automatically extract features from multi-layered neural networks, especially in large-scale data scenarios.

Supervised learning is a crucial subfield of machine learning, where a model's learning process is conducted using labeled training data, consisting of measured features of data and their corresponding labels. Therefore, the foundation of supervised learning is defined as learning information or patterns from the data used to train the model. Each example in this data is paired with a label that indicates its correct output. Consequently, in established models, data is divided into training and testing sets to verify the effectiveness of learning and training to a certain extent [160].

The general stages of data analysis with supervised learning algorithms are as follows: Data Collection and Preparation, Data Preprocessing, Model Selection, Splitting Data into Training and Testing Sets, Feature Selection, Model Training, Model Testing and Evaluation, Model Adjustment and Improvement, and Model Deployment or Application.

Data Collection and Preparation:

The first step is to collect appropriate data for the problem at hand and prepare this data to be ready for processing. During the data collection process, the objective is determined first, followed by the creation of a dataset containing input variables (inputs) and target outputs (outputs) that are suitable for the purpose. This dataset will be used to train the learning algorithm.

Data Preprocessing:

Raw data collected is often not suitable for application in a model. The collected raw data is processed by analyzing, cleaning, and preparing the data in the dataset to be suitable for the model. This step includes handling missing or incorrect data, performing feature engineering (if applicable), and operations such as data normalization or standardization. The data preprocessing step is important to improve the model's performance.

Model Selection and Training:

At this stage, a supervised learning algorithm suitable for the available data and the defined objectives is selected. Various algorithms such as regression, support vector machines, decision trees, or artificial neural networks are available options.

Splitting Data into Training and Test Sets:

After selecting the model, a portion of the data (typically 20%) is separated for testing to measure the algorithm's generalization ability. The remaining data is used for the model's training, which involves learning the relationship between the target variable and the independent variables.

Feature Selection:

Feature selection involves identifying the most effective and meaningful independent variables that directly influence the model's generalization ability. This process aims to enhance the model's performance and reduce computational costs by removing unnecessary features (if any).

Model Training:

The selected algorithm is trained on the prepared dataset. During the training process, the model acquires the ability to generalize by learning patterns and relationships between input variables and target outputs.

Testing and Evaluation of the Model:

After training, a separate validation set (typically 20% of the training data) is used to evaluate the model's performance. This validation set assesses the model's ability to generalize to independent scenarios it has not encountered before. Metrics such as accuracy, precision, and recall are used to evaluate the model's performance. If the model's performance is inadequate, adjustments may involve revisiting model settings or data preprocessing steps.

Model Tuning and Improvement:

Various methods can be applied to improve the model's performance. This phase includes tuning hyperparameters, comparing different algorithms, reviewing feature selection, or collecting more data to further train the model.

Model Deployment or Application:

Once training and evaluation are completed and the model is deemed suitable, it can be deployed or applied to real-world problems. This process involves integrating the model, monitoring its performance, and updating it as necessary.

These stages generally constitute the fundamental steps of the supervised learning process, as shown in Figure 3.6. Each stage is crucial for a successful machine learning application and should be managed carefully.

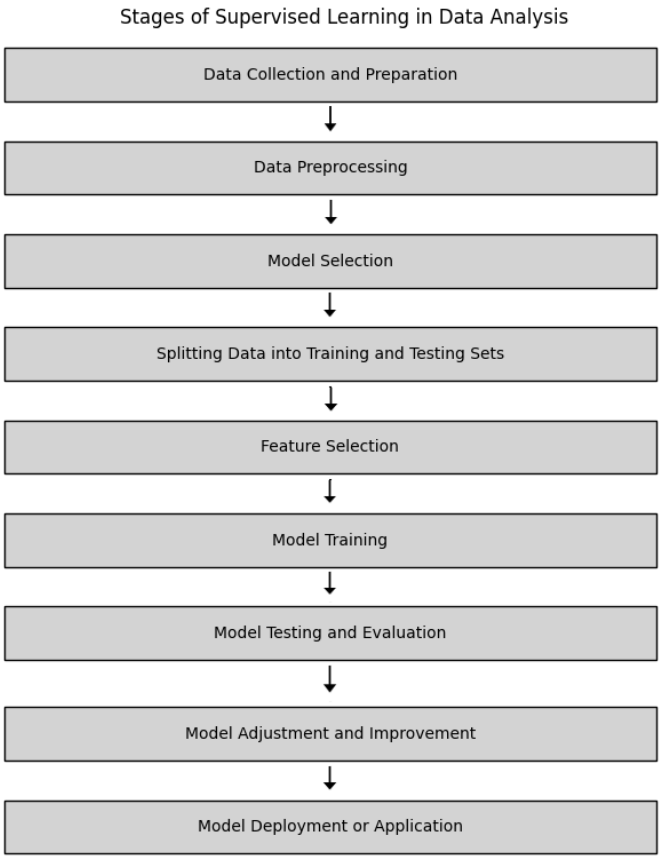

Figure 3.6 Stages of Supervised Learning

Supervised learning problems are divided into regression and classification. Classification is referred to as such when the output is categorical, while regression applies when it is continuous (numeric). Based on data, supervised machine learning finds extensive application in various fields such as image and object recognition, discovering latent patterns, sales and demand forecasting, natural language processing, financial predictions, travel and logistics, education, healthcare applications, marketing, and customer service, among others, yielding successful outcomes. Supervised machine learning has become a significant component of modern technological advancements, impacting societal economic development and human welfare in numerous ways.

To effectively utilize these algorithms with high performance, consistency, and efficiency, critical steps include proper data preparation, model selection, training,

validation, and evaluation of results. Additionally, techniques such as hyperparameter tuning and feature engineering are employed to enhance algorithm performance, thereby improving the model's ability to generalize.

3.7.1.1 Simple Linear Regression

Simple linear regression is a fundamental model used in both statistics and machine learning. This model is employed to model the relationship between a dependent variable and an independent variable used to predict it. It seeks to explain how the dependent variable changes with the value of the independent variable. The primary objective of the model is to predict the dependent variable with minimal error given the value of the independent variable. The model's application and interpretation are straightforward, and it performs well in well-established models. Since simple linear regression explains the relationship between a dependent variable and a single independent variable, multiple regression models are generally preferred for real-life data analysis. The mathematical formulation of the simple linear regression model is presented in Eq. (3.1), the error term is shown in Eq. (3.2). The visualization of linear regression is illustrated in Figure 3.7.

$$\hat{y}_i = \beta_o + \beta_1 X + \varepsilon \quad \textbf{(3.1)}$$

$$\varepsilon = y_i - \hat{y}_i \quad \textbf{(3.2)}$$

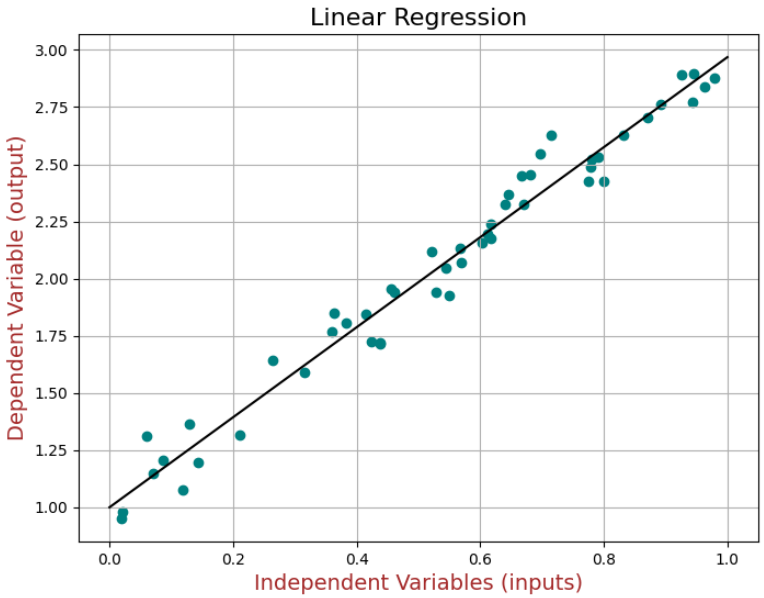

Figure 3.7 Linear Regression

The symbol $\hat{y}_i$, represents the predicted value of the dependent variable for an observation, while X denotes the independent variable used for the prediction. The parameters β_0 and β_1 are model coefficients, referred to as the intercept and slope, respectively. The actual value of the dependent variable is represented by y_i. The term ϵ denotes the error term, which is the difference between the observed and predicted values. The goal of the model is to minimize this error term.

3.7.1.2 Multiple Linear Regression

Multiple linear regression extends simple linear regression by incorporating multiple independent variables into the model. This approach aims to model the relationship between a dependent variable and multiple independent variables. The mathematical formulation is presented in Eq. (3.3). In well-established models, it is generally assumed that each independent variable has an independent effect on the dependent variable. Due to its ability to examine the relationship between multiple independent variables and the dependent variable, multiple linear regression is more suitable for real-life problems.

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \quad \textbf{(3.3)}$$

The symbol Y represents the predicted value of the dependent variable for an observation, while $X = 1,2, \ldots, n$ denotes the independent variables used for the prediction. The parameters $\beta = 0,1,2, \ldots, n$ are the model coefficients. The term ϵ denotes the error term, which is the difference between the observed and predicted values. The goal of the model is to minimize this error term.

3.7.1.3 Ridge Regression

Ridge regression is essentially derived from multiple linear regression and is defined as a method used to reduce overfitting by addressing the problem of multicollinearity. In other words, its primary goal is to reduce the error rate in multiple linear regression models. It is particularly effective when there is multicollinearity among the independent variables or when there are more independent variables than observations. Ridge regression, with its regularization parameter λ, achieves this by shrinking the coefficients of irrelevant independent variables towards zero, allowing for comparison of variable importance levels. As a result, ridge regression mitigates overfitting issues and enhances the model's ability to generalize. The mathematical formulation of ridge regression is presented in Eq. (3.4).

$$\hat{\beta}^{ridge} = argmin_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \quad \textbf{(3.4)}$$

The symbol y_i represents the i-th observation of the dependent variable, and x_{ij} denotes the j-th component of the independent variables in the i-th observation. The parameter β_0 is referred to as the model parameter or intercept. The parameter β_j represents the coefficients of the independent variables, also known as the slope coefficients. The term λ is the regularization parameter in Ridge Regression, also known as the penalty term

3.7.1.4 Lasso Regression

Lasso Regression (Least Absolute Shrinkage and Selection Operator) is defined as a regression model that, like Ridge Regression, is considered a regularization method.

While Ridge Regression shrinks the coefficients of irrelevant variables towards zero using the regularization parameter λ, Lasso Regression aims to both shrink and set coefficients of independent variables to zero with the same parameter λ (Eq. 5). This approach reduces the impact of unnecessary or ineffective variables in the model, thereby enhancing overall performance. Like Ridge Regression, Lasso Regression also reduces overfitting and improves the model's generalization ability. It is more resilient against noise and complexity in the dataset.

Especially in cases with a high number of variables, Lasso Regression increases model interpretability, reduces prediction variance, and can automatically select variables. This makes it particularly useful for handling datasets with many variables while improving explanatory power and model efficiency. The mathematical formulation of Lasso regression is presented in Eq. (3.5).

$$\hat{\beta}^{lasso} = argmin_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \quad \textbf{(3.5)}$$

3.7.1.5 Logistic Regression

Lojistik regresyon is defined as a statistical modeling technique used when the dependent variable is categorical. Using linear regression for categorical variables is inappropriate because the target values cannot be measured on a ratio scale, and the error terms are not normally distributed [161]. Additionally, while the solution space of regression models ranges from negative to positive infinity, categorical variables are limited within a specific range. The method primarily employs maximum likelihood estimation to determine the coefficients of the linear model [162].

Logistic regression is commonly used in classification problems, where it predicts the probability of the dependent variable belonging to a particular class using a linear combination of independent variables. It restricts the outcome to the range [0, 1], making it suitable for calculating probabilities. Logistic regression is robust against noise and outliers in the dataset.

3.7.1.6 K-Nearest Neighbors Algorithm

KNN (K-Nearest Neighbors) algorithm is essentially defined as a classification and regression model, much like logistic regression. In classification problems, it relies on the similarities, or in other words, the class averages of a specific number *k* of nearest data points—referred to as neighbors—to classify a new data point. In regression problems, similarly, it uses the average value of neighboring data points to predict the value of a data point.

In this algorithm, the parameter "neighbors" or *k* specifies how many neighboring observations will be considered, directly influencing model performance. Therefore, selecting an appropriate value for *k* is crucial for the success of the model.

For classification tasks, the KNN algorithm first identifies the *k* training data points closest to the new data point, and then predicts the most frequently occurring class among these neighbors as the class of the new data point. When used for regression, it predicts the average value of the neighboring points as the estimated value for the new data point.

3.7.1.7 Support Vector Machines

Support Vector Machines (SVM) are defined as a powerful learning algorithm used for classification and regression problems in machine learning. These algorithms are used to determine a line or curve that can include the maximum number of observations with the least error. In simpler terms, the goal of SVM is to find a hyperplane that maximizes the separation between classes. This hyperplane is determined to best separate the classes.

During the training process of SVM, there is a concept called margin, which attempts to maximize the gap between two classes. This enhances the model's ability to generalize. SVM's capability to perform multi-dimensional classification is achieved through "kernel functions," which can also be described as a mathematical trick [163]. In simpler terms, kernel functions allow SVM to extend its single-dimensional classification ability to two or more dimensions. Hence, SVMs can effectively classify not only linearly separable datasets but also nonlinearly separable datasets. SVM is

resilient against outliers and generally less dependent on data compared to other classification algorithms.

$$minimize\ \left(\frac{1}{2}\ \|w\|^2\right) \tag{3.6}$$

Subject to:

$$y_i(w\ \cdot\ x_i\ +\ b)\ \geq 1\ \ for\ i = 1\ ,\ldots,n \tag{3.7}$$

The term $\|w\|$ represents the norm of the weight vector, y_i denotes the class of the data point (either +1 or -1), w ,s the normal vector to the hyperplane (weight vector), and b is the bias term. The visualization of SVM showing the hyperplane and margin is illustrated in Figure 3.8.

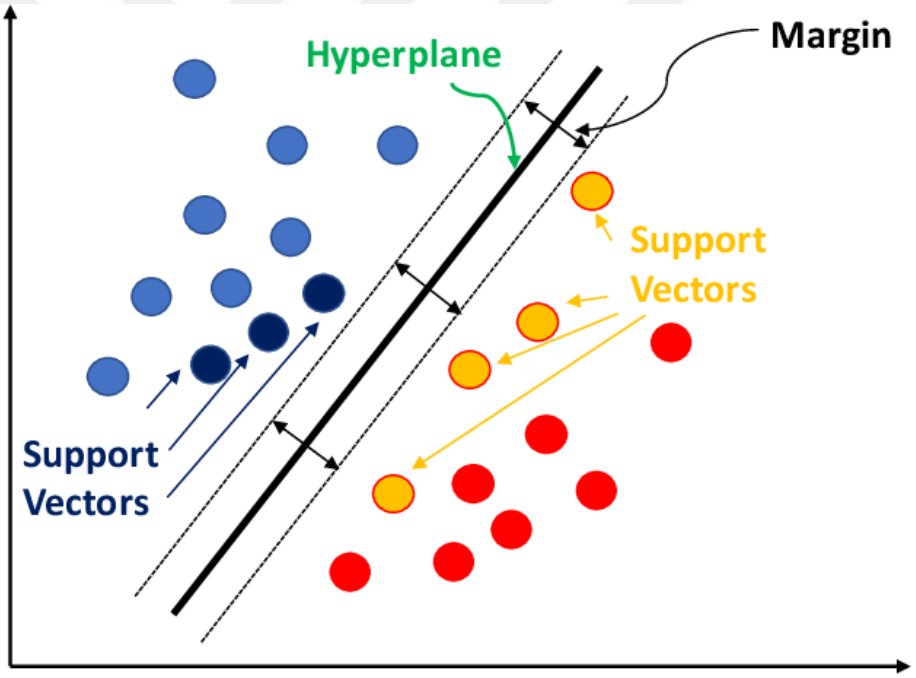

Figure 3.8 SVM hyperplane in dataset [226]

3.7.1.8 Artificial Neural Networks

Artificial Neural Networks (ANN) are defined as a model used in machine learning and artificial intelligence fields, inspired by the structure of the human brain, or more specifically, the working principles of neurons (nerve cells). Artificial Neural

Networks consist of simple processing units called neurons and their connections. In other words, neurons in artificial neural networks represent computational components or processing units. Each neuron receives inputs, processes them through an activation function, and then produces an output.

An artificial neural network model typically consists of an input layer, one or more hidden layers, and an output layer. The learning process of ANN is often carried out using the backpropagation algorithm. This algorithm propagates errors backward through the network to update weights. In simpler terms, it adjusts the weights of inputs to minimize the difference between the network's output and the actual output. The structure of an artificial neural network is illustrated in Figure 3.9.

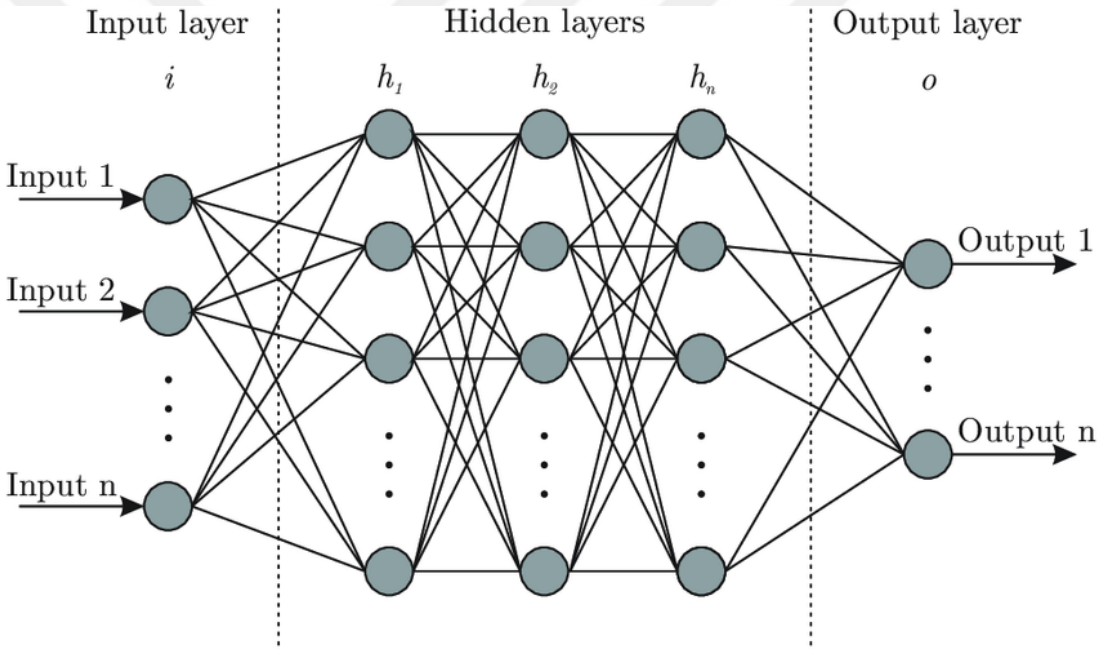

Figure 3.9 ANN working principle [227]

3.7.1.9 Decision Trees

Decision Trees are models that classify data points based on the values of features in the dataset using a series of decision rules, resembling a tree structure. Each feature test is selected to make the best possible split in the dataset. Decision Trees start with a root node and end with leaf nodes that provide the outcomes. Each internal node

represents a feature test, and each edge represents a branching based on the outcome of a test.

When input data related to an observation enters the decision tree, it branches according to the rule sets corresponding to the nodes, and the predicted class is determined at the end of the tree. There is no generalized formula for decision trees, but a logical query can also be used to represent branching in decision trees. The process of a decision tree is illustrated in Figure 3.10.

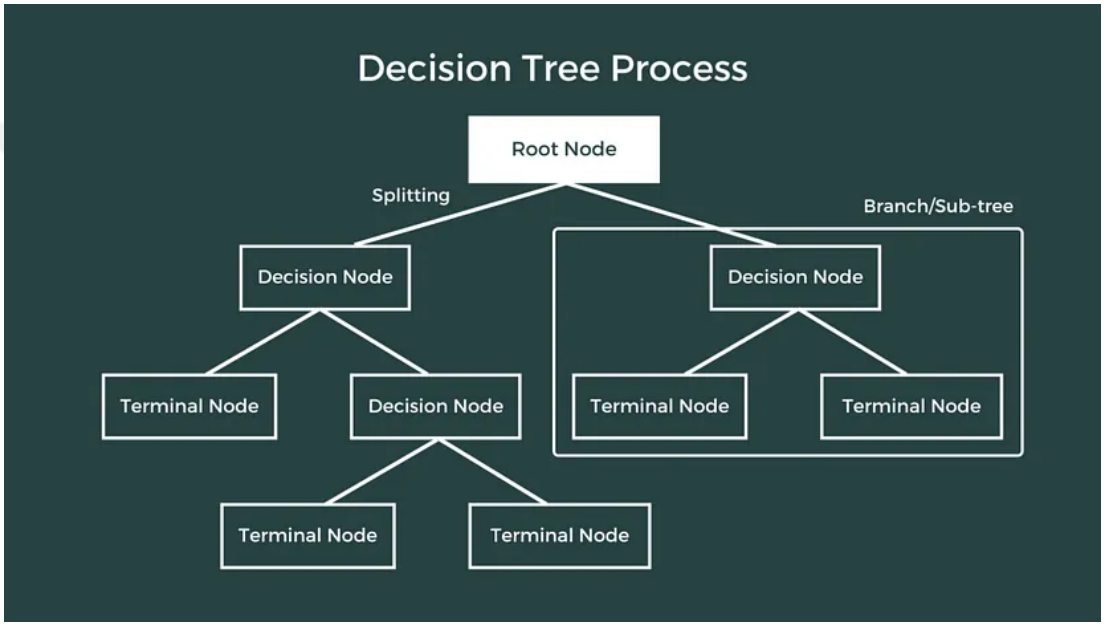

Figure 3.10 Decision Tree working principle [228]

3.7.1.10 Random Forests

Random Forests are defined as an ensemble learning technique where multiple decision trees are combined to form the model, and the result is obtained by averaging or taking the mode of predictions from each tree. Each decision tree in a Random Forest is trained on different subsets of features and random samples from the dataset. Initially, for *t* trees, *m* observations are selected from *n* observations using the bootstrap method, resulting in predictions from *t* trees.

In the Random Forest algorithm, both observation selection and variable selection are done randomly. Particularly, adjusting parameters like the number of trees

(n_estimators) and the number of variables to consider for splitting at each node (max_features) is crucial for the model's generalization ability. Unlike a single tree, Random Forests are less prone to overfitting due to the ensemble of trees and the randomness in both sample and feature selection.

3.7.1.11 Gradient Boosting Machines (GAM)

Generalized Additive Models (GAM) are described as a method that allows simple models, often weak learners like decision trees, to come together to form a powerful model used in regression and classification problems. Successive models are trained to focus on reducing the errors of the previous model.

In GAM techniques, a group of weak decision tree classifiers is iteratively trained, aiming to optimize the loss function to achieve a strong classifier. GAM also provides variable importance rankings by combining predictions from each model. Due to its nature, GAM is less prone to overfitting compared to many models used independently.

3.7.1.12 Light Gradient Boosting Machines

Light Gradient Boosting Machines (LGMB or LGBM) are defined as a machine learning technique based on traditional Gradient Boosting Machines (GBM) and derived from it. Similar to GBM, LGMB combines multiple weak learner models to create a powerful model. Essentially, LGMB is an enhanced version of GBM with faster training times and lower memory usage. This allows it to provide faster training on large datasets and operate with lower system requirements.

LGMB is particularly effective when dealing with large-scale datasets that contain a high number of variables. It improves efficiency and scalability compared to traditional GAM, making it more effective in terms of performance and scalability.

In decision trees, there are two strategies that can be used: level-wise (depth-wise) or leaf-wise strategies. In the level-wise strategy, the tree grows while maintaining balance. In the leaf-wise strategy, the splitting process continues from leaves that reduce the loss. LGMB distinguishes itself from other boosting algorithms by using

the leaf-wise strategy, which leads to lower error rates and faster learning. This is because the training time of decision trees is directly proportional to the number of computations and thus the number of splits.

3.7.1.13 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is defined as a scalable machine learning technique known as a part of the Gradient Boosting Machines (GBM) family. XGBoost is an enhanced and optimized version of GBM techniques. Similar to GBM methods, XGBoost builds a powerful predictive model by combining weak learner models.

3.7.1.14 Naive Bayes

The Naive Bayes algorithm is defined as a probability-based classification algorithm used in machine learning and statistical classification. As implied by its name, the algorithm is based on Bayes' theorem. The reason it is termed "naive" is because it assumes independence among features. In other words, it assumes that the relationship between class labels and features is independent. This independence assumption simplifies the computation and complexity of the model.

The algorithm calculates conditional probabilities for each class under the given features and selects the class with the highest probability.

3.7.2 Unsupervised Learning

Unsupervised learning is defined as a technique used initially when there is no label information available for the data set, aiming to discover hidden structures and relationships within the data and to create new clusters. This method utilizes pre-existing biases about which aspects of the input structure should be captured in the output [164]. Consequently, it automatically identifies hidden patterns and groupings, allowing for deeper exploration of the data and obtaining meaningful insights [165]. Unsupervised learning algorithms ascertain information such as similarities, differences, densities, and patterns within the data, providing significant clues for feature decisions. Therefore, it plays a crucial role in data exploration and

understanding processes. Given that the primary objective of unsupervised learning is the discovery of new clusters, it does not require the separation of the dataset for training purposes.

Despite the core aim of unsupervised learning being to explore relationships in unlabeled data, in some cases, reasons such as labeling costs, difficulties, or changes in data patterns over time (e.g., seasonal transitions) may lead to a preference for working directly with unlabeled data instead of labeling it [166]. The wide array of algorithms available in the literature complicates the selection of an appropriate algorithm for a given problem, which can result in inconsistent results from the data [167]. Depending on the type of algorithm used, selecting an algorithm that suits the characteristics of the dataset and the problem being analyzed is crucial for the model's performance. Clustering algorithms, which form the foundation of unsupervised machine learning, often contain implicit assumptions based on cluster shape or similarity measures, making it challenging for every clustering algorithm to perform well with every dataset [165,168]. Furthermore, selecting an appropriate subset of features from the raw data enhances model performance [169].

The general stages of data analysis with unsupervised machine learning algorithms are as follows: Data Collection and Preparation, Data Preprocessing, Modeling, Feature Selection, Model Evaluation and Interpretation, Review and Implementation of Results. The historical development of unsupervised machine learning algorithms has occurred during periods similar to supervised machine learning algorithms. Increased computing power with computers, the pursuit of solutions to emerging problems, and variations in the datasets have led to the development of different algorithms. The evolution of unsupervised learning methodology has been shaped by the consolidation of theoretical foundations, the widespread availability of computers and the internet, the generation of large volumes of data, and subsequent significant advances in practical applications.

Currently, unsupervised machine learning is used in clustering for tasks such as data segmentation [170], market segmentation [171], customer profiling [172], social network analysis [173], clustering and classification of web pages [174], visualization of high-dimensional datasets [175], data compression [176], health monitoring

systems with anomaly detection [177], cybersecurity [177,178], and financial fraud detection [177,179]. The difference between clustering and classification on a dataset is illustrated in Figure 3.11 [229].

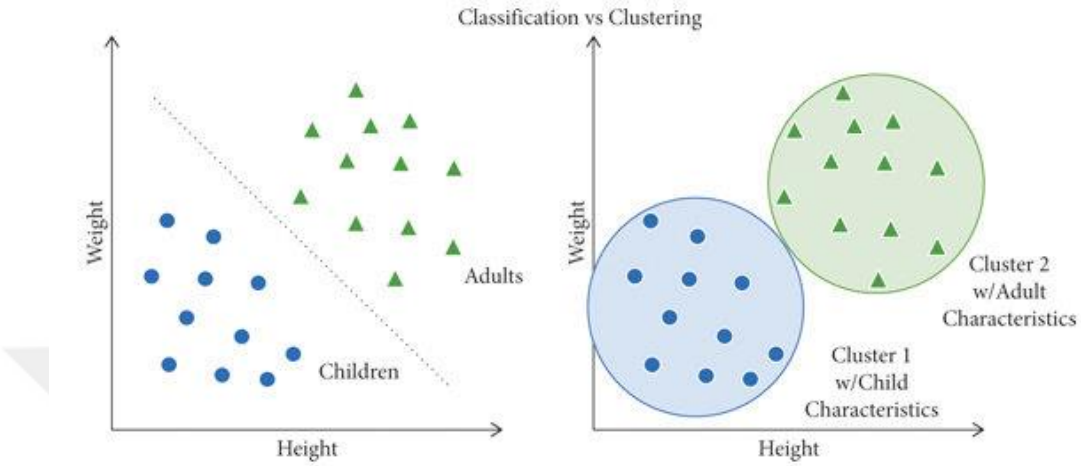

Figure 3.11 Clustering and Cassification difference on dataset [229]

3.7.2.1 K-means Clustering

Clustering is defined as one of the most common exploratory data analysis techniques used to gain insights into the structure of data. It is fundamental for separating, understanding, and learning meaningful groups within data, making clustering algorithms crucial for generating value from data [180]. The k-means Clustering algorithm aims to partition n observations into k clusters where each observation belongs to the cluster with the nearest mean (cluster center) [181]. Cluster centers represent the centroids or mean points of the n observations. In simpler terms, the K-means Clustering algorithm is defined as the process of categorizing unlabeled data into clusters alongside the closest data points. The algorithm creates clusters based on similarities (distances) between data points. Initially, the number of clusters k is determined using methods like Elbow or Silhouette [182], and cluster centers are initialized randomly or by a specific method (e.g., k-means^{++}, which selects initial cluster centers strategically) [183]. Subsequently, each data point is assigned to the nearest cluster, typically based on measures like Euclidean distance or Manhattan distance [184]. New cluster centers are computed by averaging the data points assigned

to each cluster. This process repeats until cluster centers and assignments stabilize, indicating convergence. At the final stage, when cluster centers no longer change after iterations, the unlabeled data is effectively divided into k clusters.

3.7.2.2 Hierarchical Clustering

Hierarchical clustering is a clustering method that arranges objects into a hierarchical structure and merges them based on similarity criteria. Unlike the k-means algorithm, which requires selecting the number of clusters to search for and initial configuration assignments, Hierarchical Clustering aims to overcome these disadvantages [185]. Instead, it requires specifying a similarity measure based on differences between pairs of observations (discrete) groups. The fundamental idea of the algorithm is based on grouping similar attributes together or vice versa. There are two basic approaches: agglomerative (bottom-up) and divisive (top-down). Agglomerative Hierarchical Clustering, also known as bottom-up clustering, initially considers each data point as its own cluster. Then, it finds the two closest clusters and merges them. This process is repeated until only one cluster remains. As a result, a hierarchical tree structure (dendrogram) containing all data points is obtained. An advantage of Agglomerative Hierarchical Clustering over the k-means clustering algorithm is that the hierarchical structure created by Hierarchical Clustering is more informative than the irregular collection returned by k-means [140]. Divisive Hierarchical Clustering, also known as top-down clustering, starts by considering all data points as a single large cluster. At each step, objects are separated from the main cluster based on the distance/similarity matrix, creating different sub-clusters. This process continues until each sub-cluster contains only one data point, resulting in a hierarchical structure of the dataset from top to bottom. In both methods, the calculated distance/similarity matrix forms the basis for selecting the linkage method used. Algorithms start with the assumption of n individuals and n clusters. Then, the two closest clusters are merged, reducing the number of clusters by one, and the process of iterated distance matrices is carried out (n-1) times.

Hierarchical clustering utilizes various metrics to measure similarity or distance between data points. Here are different techniques categorized by their approach:

Linkage-Based Techniques:

i. Single Linkage (Nearest Neighbor Method): Considers the smallest distance between two clusters.

ii. Complete Linkage (Furthest Neighbor Method): Considers the largest distance between two clusters.

iii. Average Linkage: Considers the average distance between clusters.

Variance-Based Techniques:

i. Ward's Linkage: Minimizes the differences in total squares within clusters.

Centroid-Based Techniques:

ii. Median Linkage: Uses the median value of each cluster to compute the distance between two clusters.

iii. Centroid Linkage: Considers the distance between the centroids (centers) of two clusters.

These techniques help define the hierarchical structure in clustering algorithms, influencing how clusters are formed and organized based on their similarity or distance metrics. Each method has its strengths and is chosen based on the specific characteristics of the data and the goals of analysis. The application of k-means and hierarchical clustering on datasets is illustrated in Figure 3.12.

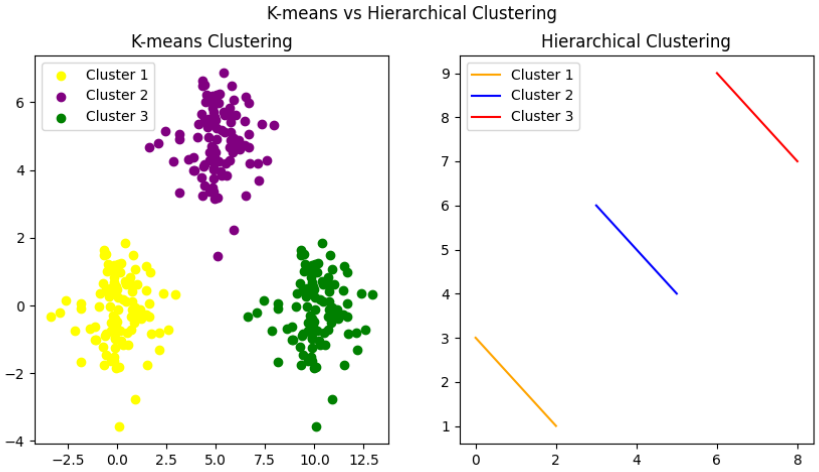

Figure 3.12 K-means and Hiearhical Clustering on datasets

3.7.2.3 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is defined as a dimensionality reduction technique used to explore fundamental patterns and variances in high-dimensional data, and it serves as a technique for classification and image compression [186]. Also known as Karhunen-Loève transform [187], its primary objective is to retain the dataset with the highest variance in high-dimensional data while reducing dimensionality. In clearer terms, its goal is to create a new coordinate system that best represents the variance in data consisting of numerous interrelated variables, thereby transforming the original dataset into a lower-dimensional form. In doing so, it constructs a set of new orthogonal variables called principal components, showcasing patterns of similarity between observations and the positioning of variables on maps [188,189]. This new foundation is expected to reveal hidden structures within the dataset and filter out noise [186], thereby facilitating easier data analysis and visualization by reducing the number of dimensions in the dataset. By reducing dimensionality, PCA can serve as a feature extraction technique that decreases the number of dimensions in datasets, helping to construct an effective machine learning model [140].

While reducing dimensions, some features may be lost. However, the intended purpose is to retain those features that contain minimal information about the population or

preserve the fundamental structure and crucial information of the dataset. Mathematically, PCA relies on the eigenvalue decomposition of positive semi-definite matrices and singular value decomposition (SVD) of rectangular matrices [188]. In other words, PCA aims to determine the directions in which the dataset exhibits the highest variance using the covariance matrix of the dataset. Initially, the data is standardized by subtracting the mean of each feature to obtain a mean-centered dataset. The features are then divided by their standard deviations to achieve unit variance. Subsequently, the covariance matrix of the standardized data matrix is computed. In the next step, the eigenvalues and eigenvectors of the covariance matrix are calculated. The eigenvalues are then sorted by magnitude, and the eigenvectors corresponding to the largest eigenvalues are selected. Finally, the *k* largest eigenvalues and their corresponding eigenvectors are used to construct a new lower-dimensional space for the data.

3.7.2.4 Apriori Algorithm

The Apriori algorithm is an unsupervised machine learning algorithm used for association rule learning. Methods of data mining that analyze the occurrence of events together are referred to as association rule methods. Despite the realization by more managers that these methods could give them a competitive edge in markets, the association rule mining algorithms developed to date have not found enough usage due to their low efficiency and complexity [190]. The name "Apriori" comes from its ability to derive information from the previous step. This algorithm aims to discover frequent itemsets in large databases and the association rules that exist among these itemsets. It is defined as an approach that supports feature-oriented work by analyzing the association behaviors in past data. In simpler terms, its goal is to analyze how often items co-occur in databases and make predictions based on these associations. This algorithm is inherently iterative, used to discover frequent itemsets in databases containing transactional data. Features and data in this algorithm are evaluated using boolean association rules [191].

Traditional Apriori algorithms face two major drawbacks: frequent scanning of the database and generating a large number of candidate itemsets [190]. The first step of the algorithm involves transforming the data in the database into a format that can

analyze relationships between items. Typically, this step involves converting the data into binary format. Next, the process moves to generating candidate itemsets, where a minimum support threshold is set. An itemset must have support above this threshold to be considered frequent. Then, all individual itemsets are generated, and their support values are calculated. Itemsets with support values above the minimum threshold are marked as frequent. Finding frequent itemsets is straightforward but costly [192].

Following this, frequent *k*-itemsets are used to generate *k+1* itemsets. This process continues until no larger itemsets can be found. Once frequent itemsets are identified, association rules are derived from these sets. The objective is to uncover relationships among frequent items through association rules. These rules are evaluated based on whether they meet minimum confidence values. Confidence is a measure in association rule analysis that indicates how accurate a rule is. It expresses the probability of the right-hand side itemset given the presence of the left-hand side itemset [193].

3.7.3 Semi-supervised Learning

Semi-supervised learning (SSL) is a machine learning approach that utilizes both labeled and unlabeled data to enhance model performance. This method addresses the challenge of acquiring labeled data, which is often costly and time-consuming, by allowing models to learn from a combination of labeled examples and a larger pool of unlabeled data. SSL aims to maximize learning effectiveness while minimizing the need for extensive human annotation effort. Like much of machine learning research, a significant portion of SSL studies focuses on classification [194].

SSL has found application in diverse fields such as computer-aided medical diagnosis and remote sensing image classification, where large volumes of unlabeled data are available but labeling is resource-intensive [194,195]. By effectively incorporating unlabeled data, SSL algorithms strive to improve model accuracy and efficiency in real-world scenarios. Unlike supervised learning, which relies solely on labeled data, and unsupervised learning, which operates exclusively on unlabeled data, SSL strikes a balance by leveraging both types to achieve better learning outcomes. These

algorithms are particularly valuable in situations where labeled data is limited or expensive to obtain.

Generative semi-supervised machine learning, however, relies on the assumption that the input data distribution matches the task of classification. When this assumption does not hold, these methods can struggle to perform effectively [196].

3.7.3.1 Self-Training Methods

Self-Training is a machine learning approach aimed at improving model performance by leveraging both labeled and unlabeled data. This algorithm is typically used in classification or regression problems where there is a limited amount of labeled data and a large amount of unlabeled data. Without making additional assumptions about the data distribution, models use confidence measures to identify decision boundaries in low-density areas commonly encountered in data distributions where classification models often face uncertainty [197].

In this technique, a classifier is initially trained with a small labeled dataset. Subsequently, this trained model makes predictions on unlabeled data. Predictions made with high confidence are then expanded with new labels. The classifier continues to self-train using its own predictions [195]. Introduced by Yarowsky in 1995 and later popularized by Abney's work [198], Self-Training reduces the cost of additional data labeling in scenarios with limited labeled data, enabling models to train on more data. However, there is potential for early errors to reinforce themselves in the model's learning process.

3.7.3.2 Generative Mixture Models

Generative Mixture Models are probabilistic models that assume observed data comes from multiple different distributions and attempt to explain the dataset by modeling these different distributions. They can identify mixture components with a large amount of unlabeled data, requiring only one labeled sample per component to precisely determine the ideal mixture distribution. These models operate under the assumption that data points are instances of specific distributions, treating each distribution as a sub-group or cluster. Gaussian Mixture Models (GMMs) are among

the most commonly used generative mixture models, but their success has been limited due to the requirement for many mixture components or states to perform well, leading to the development of more advanced models [199]. While a well-established model can be highly effective, incorrect assumptions by the generative model can adversely affect unlabeled data. Furthermore, validating the accuracy of these models is often described as challenging [195].

3.7.3.3 Co-Training Algorithm

Co-Training, a semi-supervised learning algorithm proposed by Blum and Mitchell in 1998 [200], is an approach that expands on the self-training method. This algorithm iteratively trains two or more supervised classifiers using two different and conditionally independent views of the data [201]. These views typically correspond to different subsets of the feature set, aimed at promoting classifier diversity [195]. The fundamental idea behind Co-Training is that sometimes features describing the data may be unnecessary, and each feature alone might be sufficient for correct classification [202].

Classifiers initially use a small amount of labeled data and a large amount of unlabeled data, training independently on each view. Throughout the training process, each classifier labels unlabeled examples that are deemed most reliable by the other classifier, incorporating them into the training set.

3.7.3.4 Semi-Supervised Support Vector Machines (S3VM)

Semi-Supervised Support Vector Machines (S3VM) are defined as a type of support vector machine (SVM) developed for semi-supervised learning based on the cluster assumption [203]. While traditional SVMs are trained solely on fully labeled datasets, S3VM aims to build a classifier using both labeled and unlabeled data concurrently. This approach seeks to enhance classification performance by leveraging more data, including unlabeled examples, which could potentially improve performance by using more likely data instead of unlabeled data that may degrade classifier performance [204]. The principles of machine learning algorithms are illustrated in Figure 3.13.

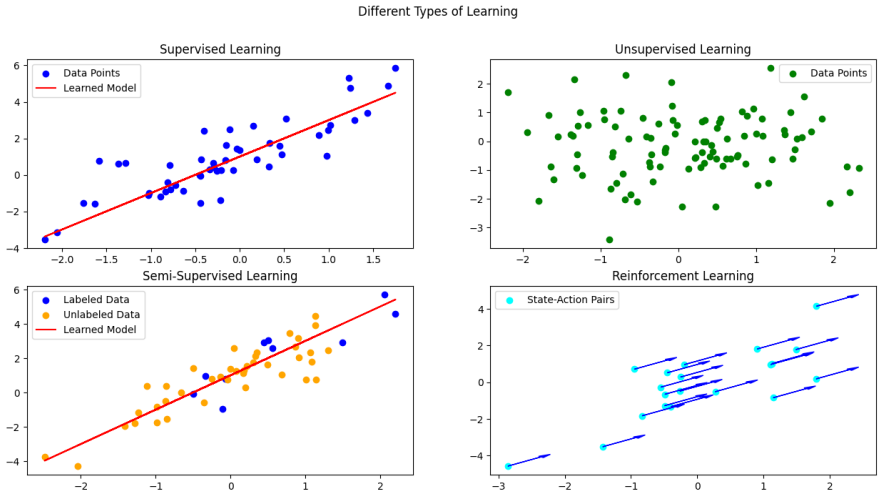

Figure 3.13 Principles of Machine Learning algorithms

Reinforcement learning, which utilizes reward and penalty mechanisms to guide optimal behavior in complex environments, stands as a potent machine learning paradigm with wide-ranging applications in robotics, autonomous systems, manufacturing, and logistics, offering unparalleled potential for enhancing efficiency and automation in sophisticated systems [140]. It offers the prospect of programming agents to learn behavior through reward and punishment, without explicit task specifications, although formidable computational challenges remain. Essentially, agents in reinforcement learning must learn behavior through trial and error in dynamic environments [205]. This approach, akin to trial-and-error learning in psychology, represents a class of problems rather than just a set of techniques, with two primary strategies for solving reinforcement learning problems: behavioral search and statistical estimation. The former explores the space of behaviors to find optimal actions, while the latter employs statistical techniques and dynamic programming methods to estimate action utilities in various states of the environment [206].

In reinforcement learning, the goals and rewards determined by the trainer or software are crucial for the agent to achieve its objectives. Unlike supervised learning, where input-output pairs are presented, reinforcement learning tasks involve learning from rewards and subsequent states, necessitating active exploration of the environment for optimal behavior, while unsupervised learning focuses on discovering patterns and

structures within unlabeled data, without explicit guidance from labeled examples. Reinforcement learning generally consists of the stages of Observation, Action, Reward, Strategy Development, Learning, Update, and Exploration. Due to its capacity for decision-making and learning in complex and dynamic environments, reinforcement learning is widely used in fields such as autonomous vehicles, robotics, finance, healthcare, energy, disease diagnosis, transportation, and more [207].

3.8 Model Validation

Model validation, a process used in machine learning, aims to evaluate and verify the performance of developed models. This process assesses the model's ability to generalize and make predictions, identifying models that may suffer from issues like overfitting or underfitting, which can lead to poor performance. To test the accuracy of a model, it is evaluated using data different from what was used for training—data the model has not seen before. Model accuracy is typically tested in supervised learning. Usually, 80% of the dataset is used for training, while the remaining 20% is reserved for validation. The model's performance is tested on this separate validation set, and the results are evaluated to determine how well the model fits real-world data.

However, splitting data into training and testing sets can pose challenges in some cases. When the dataset is small, it's preferable to use as much data as possible for training, leaving less for testing. Another issue arises when splitting the data, known as distributional differences, where data in training or testing sets may not be evenly representative. This can lead to some data being overrepresented or underrepresented in either set, thereby compromising model performance.

For such cases, the literature recommends techniques like k-fold cross-validation, stratified k-fold cross-validation, and hold-out validation. These methods help mitigate the issues of data scarcity and distributional differences by systematically partitioning the data into multiple subsets for training and validation, ensuring more robust model evaluation and performance estimation.

3.8.1 k-fold Cross Validation

Cross-validation is defined as a data resampling method used to evaluate the generalization ability of prediction models and to prevent overfitting. Cross-validation belongs to the family of Monte Carlo methods [208]. In k-fold Random Subsampling method, essentially a holdout method is repeated *k* times. Put simply, the dataset is divided into *k* random and equally sized parts. One part out of *k* is selected as the test set, and the remaining *k-1* parts are used as the training set to train the model. The model is then evaluated on the test set. This process is repeated *k* times, each time selecting a different part as the test set. The evaluation (performance) of the model is measured by averaging the results across all k test sets.

On the other hand, k-fold Cross-Validation works similarly by dividing the data into *k* parts, but with a different approach. In this method, the first *k-1* parts are used as the training set, and the last part is used as the test set. The model is trained and tested accordingly. In the next step, the previous test set is included in the training set, and a different part is chosen as the new test set. This process is repeated *k* times, each time selecting a different part as the test set. Similarly, the evaluation (performance) of the model is measured by averaging the results across all *k* test sets.

In k-fold Random Subsampling, each selection of parts is completely random, potentially leading to overlaps between training sets or test sets in different iterations. In contrast, k-fold Cross-Validation divides the dataset into *k* parts in a fixed manner, ensuring that each data point appears exactly once in the test set. This ensures more consistent and repeatable results, as each dataset point is tested exactly once.

3.8.2 Stratified k-fold Cross Validation

Stratified k-fold cross-validation is defined as a data resampling method used to evaluate the generalization ability of prediction models and to prevent overfitting or underfitting. This method is a variation of traditional k-fold cross-validation and is particularly recommended for datasets with class imbalance. In traditional k-fold cross-validation, the dataset is randomly divided into k equal subsets (folds), with each

fold used once as the test set and the remaining *k-1* folds used as the training set. Thus, this method involves randomness.

In stratified k-fold cross-validation, however, the class distribution within each fold is preserved similar to the original dataset. This means that each subset reflects the class ratios present in the original dataset. This method allows for a more accurate evaluation of model performance, especially in cases of high-class imbalance [209]. The fundamental difference from k-fold cross-validation lies in preserving the class distribution of the dataset when dividing it into *k* equal parts. This ensures that each fold maintains similar class ratios.

3.8.3 Hold-out Validation

Hold-out validation is described as the fundamental and most commonly used method for evaluating the performance of machine learning models. Despite appearing to have a different objective within various data resampling strategies, hold-out validation fundamentally operates on the same logic. It involves setting aside a portion of the data for testing to eliminate overfitting, with the remaining portion used for training [210]. It is defined as one of the simplest data resampling methods. The single holdout method, in simpler terms, refers to splitting the dataset into training and testing data only once (80%-20%), evaluating the model based solely on this split. While the training data is used for learning and adjusting parameters, the test data serves to assess the overall performance of the model.

3.8.4 Leave-one-out Cross Validation (LOOCV)

Leave-One-Out Cross-Validation (LOOCV) is a special case of cross-validation used to evaluate the accuracy of a model. It is a cross-validation method where the number of folds equals the number of samples, especially useful when the dataset is small [211]. Since the dataset is small, each example in the dataset is sequentially separated as the test set, and the model is trained on the remaining dataset. Therefore, the process is repeated for each example in the dataset.

3.9 Machine Learning Algorithms Evaluation Metrics

The evaluation of machine learning methods is defined as a critical component of planning. To assess the success of a model, it must be tested on data it has never seen before. If the model performs well on this unseen data, it is referred to as having good generalization ability [212]. When evaluated using metrics such as accuracy, a model can provide satisfactory results; however, when evaluated using other metrics such as logarithmic loss, the results may be weak. Sometimes, classification accuracy is attempted to be used to evaluate model success; however, it is not sufficient for model selection [213].

In machine learning, bias represents how far the model's predictions are from the true value. High bias indicates that the model does not encompass the solution to the problem. Variance, on the other hand, indicates how much the model's predictions vary with different datasets. In machine learning, the bias-variance trade-off is a concept that determines the balance between the complexity of the model and the predictions, meaning a balance must be struck between low variance with high bias or high variance with low bias [214].

There are various metrics for evaluating machine learning algorithms. These include accuracy, precision, recall, F1 score, ROC curve, and other performance measures. Each metric evaluates a different aspect and provides a broad perspective to understand the model's performance. The use of these metrics helps in better understanding the model's fit with real-world data and its generalization ability. In unsupervised learning (clustering algorithms), since there is no specific class value information, the aim is to form new clusters; therefore, there are no reference values for the model's accuracy. The success of classification and regression models is measured based on their performance on the allocated test data.

3.9.1 Evaluating the Performance of Classification Models

The performance of classification models is a measure of how successfully a machine learning algorithm distinguishes between categories. It is expected that the constructed model performs well according to its purpose and contributes to generating meaningful

values from the data for the practitioner. Therefore, assessment metrics are needed to evaluate the effectiveness of classifiers and to solve the problem effectively. There are many different metrics available to measure the performance of classifiers, each evaluating different aspects provided by the classification algorithm [215]. Thus, the selection of performance metrics used to measure the performance of the constructed model is crucial. Choosing metrics that are suitable for the topic and dataset at hand is one of the most important factors in accurately measuring the success of the model. In many cases, using a single evaluation measure is not sufficient to properly assess the model's performance. Evaluating based on multiple performance metrics for measuring the generalization ability of the model allows for a more consistent evaluation of the problem at hand. Evaluation metrics are used for three different purposes besides evaluating generalization ability: Model Selection, Distinguishing Optimal Solutions [216].

Classification models are fundamentally different from regression models due to their dependent variable nature. Some performance metrics used for classification problems lack a clear theoretical basis, and whether recent methods have introduced a new perspective has not yet been comprehensively examined [217].

Performance metrics are classified into 3 categories [216, 217].

i. Metrics based on error threshold and qualitative error understanding

ii. Metrics based on the probabilistic understanding of error probability

iii. Metrics that measure the performance of the model based on how it ranks examples

Threshold Metrics and Qualitative Error Metrics: Metrics based on threshold and qualitative error understanding evaluate classification results based on a decision threshold. For example, accuracy and error rate are threshold metrics. The decision threshold is typically set to 0.5 by default, and the classifier's probability estimates are compared against this threshold [215]. The threshold can be adjusted to obtain different performance metrics and to account for class imbalance or unequal costs of misclassification.

Probability Metrics: These metrics evaluate the classifier's estimated probability distribution for each class. The classifier provides probability predictions for each class, and probability metrics are used to measure the accuracy and consistency of these predictions. For instance, multi-class logarithmic loss is used as a probability metric in multi-class problems.

Ranking Metrics: These metrics evaluate the ranking of predictions generated by the classifier. They are useful for assessing the classifier's accuracy, precision, and specificity. For example, metrics such as Receiver Operating Characteristic (ROC) curve and Area Under the ROC Curve (AUC-ROC) are used to evaluate the classifier's performance at different threshold values.

Other metrics: There are additional graphic-based metrics not included in other metric categories and specific metrics designed for evaluating the performance of particular classification algorithms [216]. Metrics like SAUC and PAUC fall between probability metrics and ranking metrics [215].

For binary classification problems, a 2x2 confusion matrix is a table visualizing model predictions against actual class labels, showing the counts of predicted samples for each of the four possible outcomes [215,216]. Each row of the table represents a predicted class, while columns represent actual classes. Although the confusion matrix is not a direct performance measure, it serves as a foundation for calculating other metrics, and many performance metrics are derived from it. Values in the confusion matrix are termed as true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). A sample is counted as TP if its actual class is positive and the model predicts it as positive, TN if the actual class is negative and the model predicts it as negative, FP if the actual class is negative but the model predicts it as positive, and FN if the actual class is positive but the model predicts it as negative [215,216]. A classification algorithm uses a decision threshold, a parameter between [0, 1], to determine class membership for a specific example. The 2x2 confusion matrix is presented in Table 3.1.

Table 3.1 Confusion Matrix

	Actual Positive Class	**Actual Negative Class**
Predicted Positive Class	True positive (TP)	False negative (FN)
Predicted Negative Class	False positive (FP)	True negative (TN)

Accuracy, sensitivity, and precision are commonly used and straightforward metrics to evaluate the effectiveness of classification models [215,216]. Accuracy represents the ratio of correctly predicted observations to the total number of observations. Many classifiers aim to maximize accuracy [215,216]. Sensitivity measures how well observations belonging to a specific class are correctly predicted, while precision evaluates how many of the predictions for a specific class are actually correct. The use of these metrics is crucial for analyzing the performance of classification models in detail and assessing their reliability. Particularly in real-world applications, accuracy alone may not be sufficient; therefore, additional metrics such as sensitivity, precision, and F1 score are recommended [215,216]. These metrics help understand how the model performs in different scenarios and can guide improvements to the model.

Accuracy: Accuracy measures the ratio of correctly predicted observations by a classification model to the total number of observations (Eq. 3.8). In simpler terms, it assesses how accurately the model predicts. An increase in accuracy indicates improved overall model performance, while a decrease suggests less accurate predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \textbf{(3.8)}$$

Error Rate: Error Rate also known as the misclassification rate, measures the ratio of incorrectly predicted observations by a model to the total number of observations (Eq. 3.9). In simpler terms, it directly quantifies how much error the model makes. It is used as a complementary metric to accuracy, providing insight into the proportion of observations that the model fails to classify correctly.

$$Error\ Rate = \frac{FP + FN}{TP + TN + FP + FN} \quad \textbf{(3.9)}$$

Sensitivity: Sensitivity measures how accurately the observations belonging to a particular class are predicted (Eq. 3.10). In other words, it indicates the proportion of actual positive examples that are correctly identified as positive. An increase in sensitivity indicates an improvement in the model's ability to correctly recognize the positive class, while a decrease indicates a decline in this ability.

$$Sensitivity = \frac{TP}{TP + FN} \quad \textbf{(3.10)}$$

Specificity: Specificity, also known as recall or true positive rate, measures the proportion of actual positive instances that are correctly identified by a classification model (Eq. 3.11). In simpler terms, it shows how well the model identifies positive examples out of all actual positives. An increase in sensitivity indicates improved ability of the model to correctly identify the positive class, while a decrease indicates a reduction in this ability.

$$Specificity = \frac{TN}{TN + FP} \quad \textbf{(3.11)}$$

Precision: Precision, a measure in classification, evaluates the proportion of predicted positive instances that are actually positive (Eq. 3.12). In simpler terms, it shows how many of the predicted positives are truly positive. An increase in precision indicates that the model's positive predictions are more reliable, while a decrease suggests that the model is prone to producing false positives.

$$Precision = \frac{TP}{TP + FP} \quad \textbf{(3.12)}$$

Recall: Recall, also known as sensitivity, measures the proportion of actual positive instances that are correctly predicted by the model (Eq. 3.13). It uses the same formula as sensitivity. An increase in recall indicates that the model is capturing more true positives, while a decrease indicates a reduction in this capability.

$$Recall = \frac{TP}{TP + FN} \quad \textbf{(3.13)}$$

F-Measure (F1 Score): F-Measure is defined as a metric calculated by taking the harmonic mean of Precision and Recall. In other words, it evaluates the balance between precision and recall performance of a classification model, measuring the trade-off between sensitivity and positive predictive value (Eq. 3.14). An increase in F-Measure indicates improvement in both precision and recall performance, while a decrease suggests a deterioration in this balance. It is important because it assesses the model's ability to balance between making correct positive predictions and minimizing false alarms.

$$F\ Measure = \frac{2 * Precision * Recall}{Precision + Recall} \quad \textbf{(3.14)}$$

Geometric-mean: Geometric-mean, is defined as a performance metric calculated by taking the geometric mean of Sensitivity and Specificity (Eq. 15). It demonstrates the model's ability to correctly classify both positive and negative classes. An increase in Geometric Mean indicates improvement in the model's ability to classify both positive and negative classes accurately, while a decrease suggests a decline in this ability. It is particularly useful in evaluating overall model performance in situations that require balanced sensitivity and specificity.

$$Geometric\ Mean = \sqrt{Sensiticity * Specificity} \quad \textbf{(3.15)}$$

The importance of each of these criteria varies depending on the type of problem being addressed. For example, in a model used to diagnose a pandemic disease, high sensitivity (correctly identifying all patients) is crucial. Conversely, in a model used to diagnose whether a person has heart disease, high specificity (minimizing false positives among healthy individuals) may be more important. Therefore, the evaluation should be conducted with the appropriate performance metric suitable for the specific problem at hand.

ROC Curve (AUC): The original name is Receiver Operating Characteristic [218]. The ROC curve plots False Positive Rate $(1 - Specificity)$ on the X-axis and True Positive Rate $(Recall/Sensitivity)$ on the Y-axis (Eq. 3.16). The area under the curve (AUC) increases as the discrimination performance between classes improves. To

obtain a successful ROC curve, it is necessary to find the optimal threshold value that distinguishes between classes the best. The AUC value ranges between 0 and 1. A model that makes all predictions wrong would have an AUC of 0.0. The AUC also demonstrates the performance of the classification model as the classification threshold changes. The higher the value, the better the model's classification ability. The ROC curve is illustrated in Figure 3.14 [230].

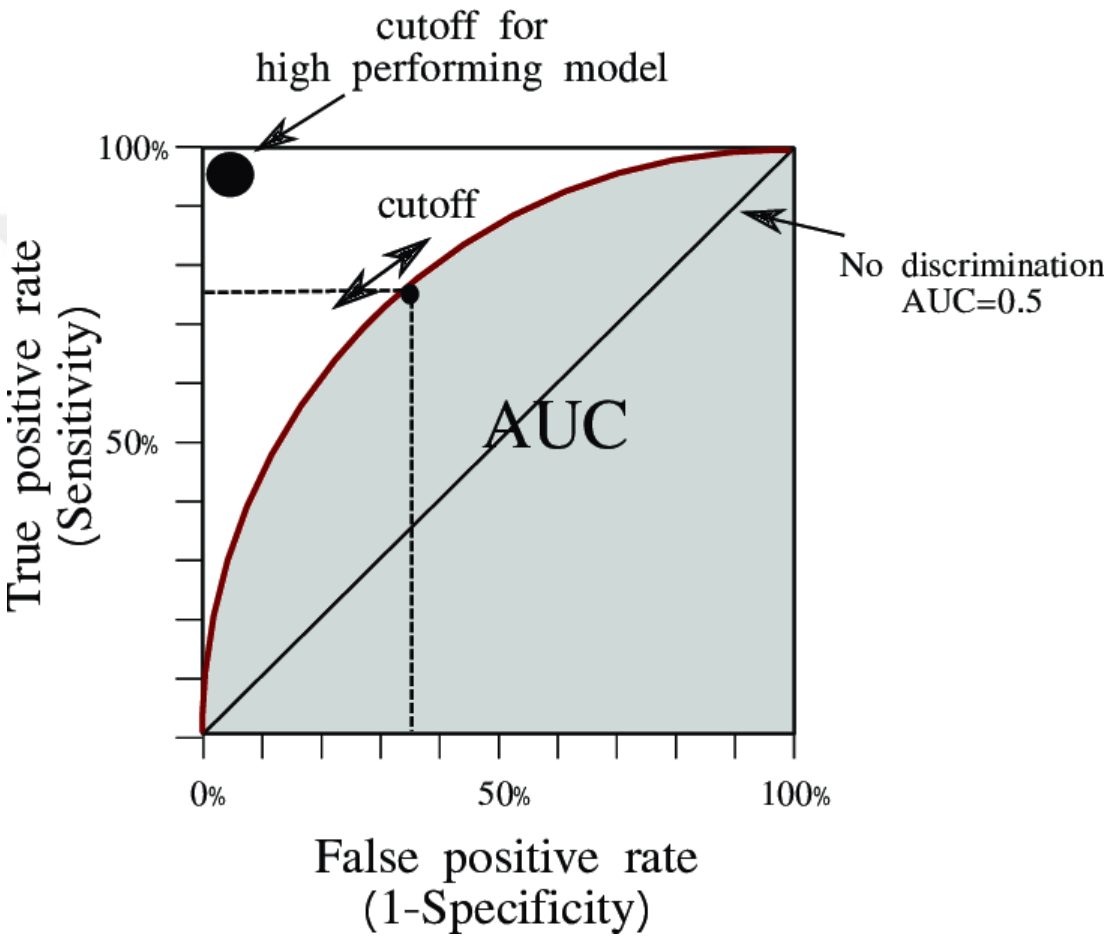

Figure 3.14 ROC curve [230]

$$AUC = \int_0^1 TPR(fpr)d(fpr) \tag{3.16}$$

Area Under the Precision-Recall Curve: This performance metric illustrates the balance between precision and recall for all possible threshold values of the model, similar to ROC curves (Eq. 3.17). Unlike single-value metrics such as accuracy or F1 score, precision-recall curves provide a graphical representation of classifier

performance across multiple threshold values. AUC-PR is obtained by calculating the area under the precision-recall curve. This area ranges between 0 and 1. Values closer to 1 indicate better performance of the model across all threshold values. It is particularly considered useful in imbalanced classification problems [219].

$$AUC = \int_0^1 Precission\ (recall)\ d(recall) \quad \textbf{(3.17)}$$

Logarithmic Score (LogLoss): Logarithmic score indicates how well the predictions of a model align with the actual values [220]. It is sensitive to the accuracy of predictions, showing not only whether the model makes correct predictions but also how confident it is in those predictions. Logarithmic score is used for probability-based classification models to calibrate the model's predictions and to prevent biases towards different classes (Eq. 3.18). A lower Log Loss value indicates better alignment between the model's predicted probabilities and the actual values. As Log Loss increases, it indicates a greater deviation between the predicted probabilities and the actual values.

$$Logarithmic\ Score\ = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(p_i) + (1-y_i)\log(1-p_i)) \quad \textbf{(3.18)}$$

Brier Inaccuracy: The Brier error metric is used to evaluate the accuracy of probability predictions and is calculated directly from the squared error of probability predictions (Eq. 3.19). The Brier error score assesses the quality of probability predictions. In simpler terms, it evaluates the success of predictions based on the probability estimate. A lower score indicates higher prediction quality.

$$Brier\ Inaccuracy = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2 \quad \textbf{(3.19)}$$

Kolmogorov-Smirnov Statistic: Originally designed as a hypothesis test to assess the fit of data to a distribution, the Kolmogorov-Smirnov test has also been used as a distance metric to measure the discriminatory power of classifiers in binary classification problems [221]. It is a non-parametric test and statistical metric used to measure how similar or different the cumulative distribution functions (CDFs) of two

probability distributions (or datasets) are. Because it does not depend on the scale or range of measured values, it is defined as suitable for comparing classifiers. It is used to determine differences in medians between two groups or the shapes of their distributions, assessing how much the model's learned distribution differs from the actual distribution (testing model accuracy).

3.9.2 Assessment of Regression Model Performance

Measurement of the success of regression models can be easily done with statistical methods due to the target variable being continuous (numerical). Performance metrics used for evaluating regression models include MAE, MSE, RMSE, and the coefficient of determination (R^2), which represents the percentage of variation in the dependent variable explained by the independent variables.

Mean Absolute Error (MAE): MAE is a metric that measures the average absolute difference between predicted and actual values (Eq. 3.20). It is more robust against outliers, and a low MAE value indicates that the model has high generalization ability.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad \textbf{(3.20)}$$

Mean Squared Error (MSE): MSE gives the average of the squared errors between predicted and actual values (Eq. 3.21). It is commonly used in regression problems and indicates the magnitude of errors. A low MSE value indicates that the model fits the data well.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \textbf{(3.21)}$$

Root Mean Squared Error (RMSE): RMSE is a metric calculated by taking the square root of MSE (Eq. 3.22). It is easier to interpret compared to MSE and is commonly preferred for understanding error rates in regression models

$$RMSE = \sqrt{MAE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad \textbf{(3.22)}$$

Coefficient of Determination (R^2): R^2, measures the percentage of the variance in the dependent variable that is explained by the independent variables (Eq. 3.23). In simpler terms, it shows how much of the variance in the dependent variable can be explained by the model. It ranges between 0 and 1, where higher values (approaching 1) indicate a better fit of the model.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad \textbf{(3.23)}$$

Adjusted R-squared: Adjusted R^2 is similar to R^2 but penalizes for the addition of extra independent variables in the model (Eq. 3.24). In other words, it adjusts the R^2 value considering the complexity of the model. It helps prevent overfitting problems by evaluating whether the increase in R^2 due to added variables truly improves the model. Especially in multiple regression models, adding each new independent variable can increase R^2, but Adjusted R^2 corrects for this increase to assess whether the model has genuinely improved.

$$Adjusted\ R^2 = 1 - \left(\frac{(1 - R^2)\,(n-1)}{n-k-1}\right) \quad \textbf{(3.24)}$$

Mean Absolute Percentage Error (MAPE): Mean Absolute Percentage Error measures the prediction error as a percentage of the actual values, making it useful for comparing datasets with different scales (Eq. 3.25). In other words, MAPE quantifies how much the prediction deviates from the actual value as a percentage. However, MAPE is sensitive to zero values because of the division involved. If a model predicts zero where the actual value is zero, MAPE becomes infinite. Therefore, it is not widely used among performance evaluation metrics.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| * 100 \quad \textbf{(3.25)}$$

These performance metrics are used to evaluate the predictive capabilities of regression models, each measuring different aspects of model performance. Metrics like MAE and RMSE indicate the magnitude of errors, while R-squared (R^2) and adjusted R-squared assess how well the model explains the data.

CHAPTER 4

METHODOLOGY

Section 1 provides an overview of the research methodology used in this study and a general description of each phase. Section 2 covers the process of data acquisition and compilation. Section 3 includes the prediction model, detailing the data acquisition process, dataset size, selection of candidate no-show factors, description of attributes, the approach utilized in data preprocessing, the reasoning behind actions used in this step, the techniques utilized to overcome the data imbalance problem, and the prediction algorithms used in this research. Section 4 presents the obtained results and includes sensitivity analysis findings.

4.1 Framework of Research

In this phase, the scope and basic framework of the study are briefly outlined. There are two primary categories of studies that forecast passenger no-show rates. The first category involves using statistical methods to determine the upper limit of tickets that can be sold for no-shows. The first category's studies utilize exponential adjustment methods to find the optimal profit point for extra tickets sold and balance potential losses if all passengers arrive, integrating these two methods. The second category utilizes machine learning methods to predict passenger-specific no-show rates. The second category's studies compare passenger information obtained from PNR numbers with patterns from available datasets to identify similarities and aim to make specific no-show predictions for each passenger.

This proposed study explores the impact of environmental factors on no-show rates, a topic not previously addressed in the literature. It examines how external factors such as daily (and hourly) weather conditions, traffic density, flight time, special events, which can vary at any time, influence no-show rates. Therefore, the data needed for this study were selected and examined for the first time to fit the research topic, differentiating it from previous studies in the literature that predict no-show rates.

PNR-based studies focusing on passenger information in predicting no-show rates are found to be insufficient. These studies typically make individual predictions for each passenger and aggregate these predictions to estimate total no-shows for a flight. However, due to each passenger being unique, these methods struggle to identify patterns effectively. The fundamental reason lies in the complexity of Machine Learning methods used to find these patterns, as each individual traveler is inherently unique. Factors such as a passenger's sleep state, mental condition, fatigue level, reason for ticket purchase, employment status, among others, can significantly influence their decision to show up for a flight. Therefore, the limited personal information available and its insufficiency to identify patterns make PNR-based no-show predictions less effective.

In contrast, this study investigates the impact of external factors (such as weather conditions, traffic density, flight time, flight day, and other contextual factors) on no-show rates. This research stands apart from existing literature on aviation no-show predictions by being the first to focus on these external factors. The study considers the flight as a whole in its predictions and examines the effects of anomalies like public holidays, flight times, weather conditions, and flight routes on no-show rates. Historical data-based methods like exponential adjustments assume similar flights will have similar no-show rates, which align with a statistical normal distribution. However, a significant drawback is the inability to incorporate environmental factors into these predictions. Hence, Machine Learning methods are preferred in this study, allowing for the simultaneous examination of categorical data and the discovery of previously unrecognized patterns. This approach aims to account for variations among similar flights (same departure-arrival points, same time, same aircraft capacity) in predicting no-shows more accurately and obtaining sharper insights.

4.2 Data Acquisition

Data collection refers to the process of gathering data for the purpose of predicting no-show rates. The data for no-show prediction has been categorized into two groups: flight information and weather data. The flight information dataset includes one month's worth of both domestic and international flight data for five major cities in

Turkey, provided by public data from an airline company [222]. This data was obtained from publicly available sources. Weather data, on the other hand, was obtained from a web-based source [223]. Personal data of passengers has not been included in the scope of this study.

4.3 No-Show Prediction Model

4.3.1 Data Preparation

In this research, each flight in the datasets has been considered uniquely. It means that the flight data does not have the stopover information for a flight that is starting from one location and connecting to another. In addition to weather information, traffic congestion segmented by hours has also been included within the scope of the study. Furthermore, any official or religious holidays on the flight day have been identified and their impact considered in the study. Within the scope of the study, Origin and Destination (O&R) data has been used for the five major airports in Turkey and the flights departing from these airports to ensure effective results in data analysis and the application of models. Consequently, a dataset excluding the dependent variable (no-show rate) but including 25 columns (features) and 6302 rows has been obtained. Descriptive Statistics for the independent variables are presented in Table 4.1.

4.3.2 Descriptive Statistics

The descriptive statistics of variables and summary statistics are presented in Tables 4.1 and 4.2, respectively.

Table 4.1 Descriptive statistics of variables

Variable	Type	Levels	Description
City	Categorical	6	The city where the journey begins
Flight Number	Categorical	6302	A unique identifier number for each flight
Direction	Categorical	2	Indicates whether the flight is outbound or return
Departure Airport	Categorical	72	The airport where passengers board the plane

Table 4.1 (continued) Descriptive statistics of variables

Variable	Type	Levels	Description
Destination Airport	Categorical	72	The airport where passengers disembark
Date	Categorical	30	The day the flight takes place
Day	Categorical	7	The day of the week the flight takes place (Monday, Tuesday, ...)
Week	Categorical	2	The part of the week the flight falls into (weekdays, weekends)
Hour	Categorical	72	The time the flight occurs (departure time)
Line	Categorical	2	The travel direction (domestic, international)
Number of Daily Trips	Numerical	5	The total number of flights from the departure airport to the arrival airport on that day
Number of Monlty Trips	Numerical	39	The total number of flights from the departure airport to the arrival airport in that month
Plane	Categorical	14	The type of aircraft used for the flight
Capacity	Numerical	12	The maximum number of passengers the aircraft can carry (maximum number of seats offered)
Special Day (for Departure Location)	Categorical	22	Whether there is a holiday at the departure airport (country)
Special Day (for Destination Location)	Categorical	23	Whether there is a holiday at the arrival airport (country)
Temperature	Numerical	76	The temperature at the departure airport
Dew Point	Numerical	72	The dew point at the departure airport
Humidity	Numerical	100	The humidity level at the departure airport

Table 4.1 (continued) Descriptive statistics of variables

Variable	Type	Levels	Description
Wind	Categorical	18	The wind direction at the departure airport
Wind Speed	Numerical	31	The wind speed at the departure airport
Wind Gust	Numerical	26	The wind gusts at the departure airport
Pressure	Numerical	396	The atmospheric pressure at the departure airport
Precipitation	Numerical	2	The precipitation amounts at the departure airport
Condition	Categorical	32	The weather condition at the departure airport (clear, cloudy, rainy, etc.)

*Weather data has been formatted appropriately while creating the dataset.

Five-Number Summaries of the obtained data are shown in Table 4.2.

Table 4.2 Five-Number summaries of variables

Variable	Statistics	
Number of Daily Trips	Min	1
	Q1	1.0
	Median	1.0
	Q3	2.0
	Max	5
Number of Monlty Trips	Min	1
	Q1	10.0
	Median	15.0
	Q3	21.0
	Max	46
Capacity	Min	132
	Q1	160.0
	Median	186.0
	Q3	300.0
	Max	370

Table 4.2 (continued) Five-Number summaries of variables

Variable	Statistics	
Temperature	Min	0.0
	Q1	68.0
	Median	75.0
	Q3	82.0
	Max	113.0
Dew Point	Min	-2.0
	Q1	54.0
	Median	61.0
	Q3	66.0
	Max	86.0
Humidity	Min	0.0
	Q1	48.0
	Median	65.0
	Q3	79.0
	Max	100.0
Wind Speed	Min	0.0
	Q1	5.0
	Median	8.0
	Q3	13.0
	Max	32.0
Wind Gust	Min	0.0
	Q1	0.0
	Median	0.0
	Q3	0.0
	Max	44.0
Pressure	Min	0.0
	Q1	29.32
	Median	29.59
	Q3	29.74
	Max	32.2
Precipitation	Min	0.0
	Q1	0.0
	Median	0.0
	Q3	0.0
	Max	0.1

Parameter Determination for the Model:

The data have been obtained from public sources, and due to the unavailability of sensitive no-show data, appropriate parameters must be determined to test the model. The no-show rate calculation involves certain assumptions, which are explained in this section as part of the data preparation process. Each aircraft has a capacity defined by CP as shown in Eq. 4.1. Each feature that could potentially cause a no-show (independent variables in the dataset) is defined by FR_i as stated in Eq. 4.2, where $i = 1,2, \ldots n$, and n is the total number of features, which is 13 in this dataset. The number of potential no-shows for each flight is defined by NS as given in Eq. 4.3. The situation where all passengers show up is given in Eq. (4.4), and the potential number of no-shows (NS) is addressed in Eq. (4.5). The assumptions for determining appropriate parameters are as follows:

$$\text{Capacity} = \text{CP} \tag{4.1}$$

$$No\ show\ reason\ (futures) = FR_i\ , \quad i = 1,2, \ldots, n \tag{4.2}$$

$$No\ Show = NS \tag{4.3}$$

$$NS = \begin{Bmatrix} If\ all\ the\ passengers\ come & = & 0, \\ otherwise & = & 1 \end{Bmatrix} \tag{4.4}$$

$$NS = \begin{Bmatrix} 1 - \dfrac{CP * \prod_{i=1}^{n} FR_i}{(n-1)\sum_{i=1}^{n} FR_i} & n \geq 2, \\ CP * FR & n = 1 \end{Bmatrix} \tag{4.5}$$

In this proposed study, determining no-shows requires setting parameters (assigning coefficients) for each independent variable (feature). The features in the dataset are: FR_{day}= Day as stated in Equation 4.6,FR_{hour}= Houras shown in Equation 4.7 , $FR_{daily\ trip}$ = Number of Daily Trips as given in Equation 4.8, $FR_{monthly\ trip}$= Number of Monthly Trips as mentioned in Equation 4.9, $FR_{\text{special day(dep)}}$= Special Day for Departure Location as indicated in Equation 4.10, $FR_{\text{special day(des)}}$ = Special Day for Destination Location as described in Equation 4.11, $FR_{condition}$ = Condition as stated in Equation 4.12, FR_{line} = Line as given in Equation 4.13, $FR_{temperature}$ Temperature as indicated in Equation 4.15,

All assignments, except for Temperature, are based on the values in the dataset. For Temperature, the assignment is based on the deviation from the average temperature. The reason for this is that the dataset covers only one month, and the raw temperature values alone do not provide sufficient comparative data to measure their impact.

Day;

$$FR_{day} = \begin{bmatrix} Monday \\ Tuesday \\ Wednesday \\ Thursday \\ Friday \\ Saturday \\ Sunday \end{bmatrix} = \begin{bmatrix} 0.010 \\ 0.003 \\ 0.003 \\ 0.003 \\ 0.010 \\ 0.005 \\ 0.005 \end{bmatrix} \quad \textbf{(4.6)}$$

Hour;

$$FR_{hour} = \begin{Bmatrix} 03{:}00\ AM - 04{:}40\ AM & 0.002, \\ 05{:}00\ AM - 06{:}40\ AM & 0.006, \\ 07{:}00\ AM - 08{:}40\ AM & 0.001, \\ 09{:}00\ AM - 10{:}40\ AM & 0.004, \\ 11{:}00\ AM - 12{:}40\ PM & 0.002, \\ 01{:}00\ PM - 02{:}40\ PM & 0.004, \\ 03{:}00\ PM - 04{:}40\ PM & 0.006, \\ 05{:}00\ PM - 06{:}40\ PM & 0.014, \\ 07{:}00\ PM - 08{:}40\ PM & 0.008, \\ 09{:}00\ PM - 10{:}40\ PM & 0.002, \\ 11{:}00\ PM - 12{:}40\ AM & 0.000, \\ 01{:}00\ AM - 02{:}40\ AM & 0.002, \end{Bmatrix} \quad \textbf{(4.7)}$$

Number of Daily Trips;

$$FR_{daily\ trip} = \begin{Bmatrix} 1 & 0.0010, \\ 2 & 0.0020, \\ 3 & 0.0025, \\ 4 & 0.0030, \\ 5 & 0.0035, \\ otherwise & 0.0040 \end{Bmatrix} \quad \textbf{(4.8)}$$

Number of Monthly Trips;

$$FR_{monthly\ trip} = \begin{Bmatrix} 0-20 & 0.00, \\ 21-40 & 0.0003, \\ 41-80 & 0.0005, \\ 81-\infty & 0.0007 \end{Bmatrix} \quad \textbf{(4.9)}$$

Special Day (for Departure Location);

$$FR_{\text{special day(dep)}} = \begin{Bmatrix} If\ there\ is\ a\ special\ day & = & 0.001, \\ otherwise & = & 0.00 \end{Bmatrix} \quad \textbf{(4.10)}$$

Special Day (for Destination Location);

$$FR_{\text{special day(des)}} = \begin{Bmatrix} If\ there\ is\ a\ special\ day & = & 0.001, \\ otherwise & = & 0.00 \end{Bmatrix} \quad \textbf{(4.11)}$$

Condition;

$$FR_{condition} = \begin{Bmatrix} Fair & 0.00, \\ Fog & 0.0010, \\ Fair/Windy & 0.00, \\ Mostly\ Cloudy & 0.00, \\ Light\ Rain\ Shower & 0.0020, \\ Mist & 0.00, \\ Partly\ Cloudy & 0.00, \\ Mostly\ Cloudy\ /\ Windy & 0.00, \\ Cloudy & 0.00, \\ Light\ Snow\ Shower & 0.0100, \\ Cloudy\ /\ Windy & 0.00, \\ Thunder & 0.0500, \\ Light\ Drizzle & 0.0010, \\ Patches\ of\ Fog & 0.0005, \\ T-Storm & 0.0450, \\ Rain & 0.0100, \\ Light\ Rain\ Shower/Windy & 0.0010, \\ Light\ Rain & 0.0010, \\ Partly\ Cloudy/Windy & 0.0010, \\ Light\ Rain\ with\ Thunder & 0.0150, \\ Thunder\ in\ the\ Vicinity & 0.0250, \\ Heavy\ Rain/\ Windy & 0.0250, \\ Thunder/Windy & 0.0350, \\ Partial\ Fog & 0.0010, \\ Rain\ Shower & 0.0200, \\ Haze & 0.00, \\ Shower\ in\ the\ Vicinity & 0.0200, \\ Widespread\ Dust & 0.0010, \\ Drizzle & 0.0005, \\ Heavy\ Rain & 0.0100, \\ Thunder\ and\ Hail & 0.0100, \\ Sand/Windy & 0.001 \end{Bmatrix} \quad \textbf{(4.12)}$$

* The terms used for weather-related data come from the dataset that encompasses weather information.

Line;

$$FR_{\text{line}} = \begin{Bmatrix} If\ domestic & = & 0.0008, \\ if\ international & = & 0.0001 \end{Bmatrix} \quad \textbf{(4.13)}$$

Therefore, to reflect the significance of the temperature value for the specific location (airport), the assignment is based on the average temperature over the past 10 years and the deviation from this average. Consequently, TAC as stated in Equation 4.14a

represents the actual temperature at the airport, while the average temperature at the same airport is symbolized as TAV as shown in Equation 4.14b. The difference between the actual temperature and the average temperature is denoted by d as given in Equation 4.14c. σ_d represents the standard deviation of the temperature differences as mentioned in Equation 4.14d. FR_{tc} represents the distribution of the c-th data point within the standard deviation in the dataset as indicated in Equation 4.14e. Thus, the values it can take are used in the assignment of $FR_{temperature}$ as described in Equation 4.14f.

Temperature;

i. Determining Temperature Difference;

$$Actual\ Temperature = TAC \quad \textbf{(4.14a)}$$

$$Average\ Temperature = TAV \quad \textbf{(4.14b)}$$

$$d = |TAC - TAV| \quad \textbf{(4.14c)}$$

$$\sigma_d = standard\ deviation(d) \quad \textbf{(4.14d)}$$

$$FR_{tc} = \frac{d}{\sigma_d} \quad \textbf{(4.14e)}$$

$$FR_{temperature} = \begin{Bmatrix} 1 & if\ FR_{tc} \leq 1, \\ 2 & if\ 1 < FR_{tc} \leq 2, \\ 3 & if\ 2 < FR_{tc} \leq 3 \end{Bmatrix} \quad \textbf{(4.14f)}$$

ii. Assing Coefficients:

$$FR_{temperature} = \begin{Bmatrix} 1 & 0.0001, \\ 2 & 0.0005, \\ 3 & 0.007 \end{Bmatrix} \quad \textbf{(4.15)}$$

$FR_{capacity}$ = Capacity (Eq. 4.16), $FR_{wind\ speed}$ = Wind Speed (Eq. 4.17), $FR_{departure}$ = Departure (Eq. 4.18), and $FR_{destination}$ = Destination (Eq. 4.19).

Capacity;

$$FR_{capacity} = \begin{Bmatrix} 0-100 & 0.0001, \\ 101-200 & 0.0002, \\ 201-300 & 0.0004 \\ 301-400 & 0.0008, \\ 401-500 & 0.0016, \\ 501-\ 600 & 0.0032, \\ 601-\ \infty & 0.0064 \end{Bmatrix} \quad \textbf{(4.16)}$$

Wind Speed;

$$FR_{wind\ speed} = \begin{Bmatrix} 0-5 & 0.001, \\ 6-10 & 0.002, \\ 11-15 & 0.004, \\ 16-\ 20 & 0.006, \\ 21-25 & 0.007, \\ 26-30 & 0.009, \\ 31-\ \infty & 0.01 \end{Bmatrix} \quad \textbf{(4.17)}$$

Departure;

If each unique value in the $FR_{departure}$ dataset is expressed as d_i, then;

$$\forall d_i \in FR_{departure},\ r_{d_i} \sim U\ (0.001, 0.01) \quad \textbf{(4.18)}$$

Destination;

If each unique value in the $FR_{destination}$ dataset is expressed as t_i, then;

$$\forall t_i \in FR_{destination},\ r_{t_i} \sim U\ (0.001, 0.009) \quad \textbf{(4.19)}$$

4.3.3 Data Preprocessing

Raw data often contains many missing or inappropriate values. Therefore, it is necessary to transform the dataset into a suitable format for the data analysis methods to be applied in this study. By enhancing data quality, the aim is to conduct data analysis more effectively. Through data cleaning, missing values in the raw data are addressed, unnecessary variables are removed, and invalid values are transformed into appropriate forms via outlier analysis. Additionally, categorical data are converted into numerical values at this stage.

As shown in Table 4.1, the dataset includes both numerical and categorical features. A description of the data is provided in the table. Temperature and Dew Point are measured in “°F”, Humidity in “%”, and Wind Speed and Wind Gust in “°mph”.

4.3.3.1 Missing Values

When applying machine learning methods, it is crucial to address missing values. These values should either be removed from the dataset or filled using statistical methods such as mean, median, mode, Multiple Imputation, Maximum Likelihood, etc. Although filling in missing values does not perfectly replace real data, and given that there are relatively few missing values in a large dataset, this study opted to remove rows with missing values. The missing data primarily stemmed from unavailable weather data for the relevant flights. In this research, therfore, a total of 9 rows were removed from the dataset.

4.3.3.2 Data Reduction & Data Balancing

In the dataset, some flights and airports have more detailed (more observations) weather data, while others have significantly fewer observations. To address this imbalance, weather data from 2 hours before the flight (since passengers typically check-in about 1 hour before the flight) was used. If only one observation was available, it was directly included in the dataset; if multiple observations (every 10 or 20 minutes) were available, their average was taken.

The flight number, being a unique code assigned to each flight, was removed from the dataset as it had no impact on the no-show rate. The city attribute was also removed to avoid redundancy, as departure and arrival airports already include this information. The direction variable was excluded since it only contained departure and arrival information, which was already covered by the departure and arrival airports. The flight date was not included because the dataset only covers one month, making it impossible to observe its effect. The week attribute, containing only weekday and weekend information, was also removed as this information is included in the 'Day' attribute. The aircraft information was excluded because it did not have a significant impact on passengers and no-shows. Precipitation was removed from the dataset since

it was already represented within the 'Condition' variable for weather data. Finally, Dew Point, Humidity, Wind, Wind Gust, and Pressure were removed from this research data set because they were either already represented by Temperature and Condition variables or did not have a direct impact on weather conditions. In the final dataset, 13 features and 6294 rows of data remain.

A table showing the relationship between continuous variables and the target variable is provided in Table 4.3

Table 4.3 Relations between continuous variables and target variable

Variable (No-Show)	Pearson Correlation Coefficient	Spearman's Rank Correlation Coefficient	Kendall's Tau Rank Correlation Coefficient
Number of Daily Trips	0.0471	0.0657	0.0553
Number of Montly Trips	-0.0114	-0.0056	-0.0041
Capacity	0.6816	0.6691	0.5339
Tempreture	-0.0280	-0.0110	-0.0082
Wind Speed	0.1605	0.1774	0.1311

Correlation coefficients measure the linear relationships between numerical variables and the No-Show variable. Analyses using different correlation coefficients (Pearson, Spearman, and Kendall) yield the following results:

There are weak relationships observed in the analysis. For the Number of Daily Trips, there is a weak positive relationship with No-Show, indicated by low correlation coefficients. An increase in the number of daily trips is associated with a slight increase in the likelihood of No-Show. Similarly, the Number of Monthly Trips shows a very weak positive relationship with No-Show, suggesting minimal impact. The Line also exhibits a very weak positive relationship with No-Show, with no significant relationship observed. Additionally, the Special Day (for Departure) demonstrates a very weak relationship, indicating that the presence of holidays at the departure location does not significantly affect the likelihood of No-Show. The Special Day (for

Destination) shows a weak positive relationship, suggesting that holidays at the destination have a minimal effect on the likelihood of No-Show.

In terms of moderate relationships, the Temperature variable shows a moderate positive relationship with No-Show. An increase in temperature is moderately associated with an increased likelihood of No-Show, although the relationship is not particularly strong. Wind Speed exhibits a weak positive relationship with No-Show, indicating that as wind speed increases, the likelihood of No-Show increases slightly. The Day shows a moderate positive relationship, with an increase in the Day associated with an increase in the likelihood of No-Show. The Hour has a moderate to strong positive relationship with No-Show, suggesting that as the Hour increases, the likelihood of No-Show increases significantly. The Condition also has a moderate positive relationship with No-Show, indicating that flight conditions have a noticeable effect on the likelihood of No-Show.

A strong relationship is observed with the Capacity, which shows a strong positive relationship with No-Show. An increase in capacity is strongly associated with an increased likelihood of No-Show, highlighting that flights with higher capacities have significantly higher No-Show rates. This pronounced relationship underscores the importance of considering capacity in predicting No-Show rates.

Table 4.4 presents the results of the one-way ANOVA tests for each categorical variable:

Table 4.4 Relations between categorical variables and target variable

Variable (No-Show)	F Value	p-Value
Departure	7.496	2.80E-65
Destination	1.431	8.624
Day	62.749	6.35E-73
Hour	13.191	2.89E-136
Line	3.501	6.138
Special Day (for Departure)	1.355	17.075
Special Day (for Destination)	982	496.039
Condition	46.355	9.38E-250

ANOVA tests were conducted to examine the relationship between independent categorical variables and the dependent variable (No-Show). The hypotheses for the ANOVA tests were formulated as follows:

- H_0: There is no significant relationship between the categorical variable and No-Show.

- H_1: There is a significant relationship between the categorical variable and No-Show.

ANOVA tests were conducted to examine the relationship between each independent categorical variable and the dependent variable, No-Show, using a One-Way ANOVA approach. For the Departure variable, the F value is 7.496 with a p-value of 2.80E-65. indicating a significant effect of the departure location on the likelihood of No-Show. The Day variable shows an F value of 62.749 and a p-value of 6.35E-73, suggesting that the day of the flight significantly influences the No-Show rate. The Hour variable, with an F value of 13.191 and a p-value of 2.89E-136, indicates that flight time is a significant determinant of No-Show likelihood. The Condition variable, which includes factors such as weather, presents an F value of 46.355 with a p-value of 9.38E-250, indicating a strong significant effect on the likelihood of No-Show.

In contrast, the Destination variable has an F value of 1.431 and a p-value of 8.624, suggesting that the destination does not significantly impact the No-Show rate. The Line variable, with an F value of 3.501 and a p-value of 6.138, does not provide sufficient evidence to reject the null hypothesis, indicating that the flight route does not significantly affect the No-Show rate. Additionally, the variables Special Day (for Departure) and Special Day (for Destination) have F values of 1.355 and 982, respectively, with p-values of 17.075 and 496.039, suggesting that whether the departure or arrival day is a holiday does not significantly affect the likelihood of No-Show.

These results indicate that certain categorical variables, such as departure location, flight day, flight hour, and weather conditions, have a significant impact on the No-Show rates, while other factors like destination and flight line do not have a statistically

significant effect. Each categorical variable was analyzed separately using One-Way ANOVA to determine its unique contribution to the variability in No-Show rates.

After converting categorical variables into numerical (continuous) variables, the same tests were repeated. The categorical variables Day, Hour, Line, Special Day (for Departure), Special Day (for Destination), and Condition were transformed into continuous variables, while the Departure and Destination variables remained categorical. Additionally, some numerical variables were also converted using specific criteria. For example, Day was encoded using unique identifiers for each day of the week, and Hour was categorized into specific time intervals. The details of these transformations, including the methodology for encoding numerical values like Number of Daily Trips, Number of Monthly Trips, Capacity, Temperature, and Wind Speed, are provided in the equations (4.6) through (4.19). Specifically, the transformation was based on factors such as the distribution of the data, historical averages (for variables like temperature), and specific ranges for variables like capacity. This process was applied to ensure a consistent scale and to categorize the data into meaningful intervals. The results for the numerical variables are presented in Table 4.5.

Table 4.5 Relations between continuous variables and target variable after convert

Variable (No-Show)	Pearson Correlation Coefficient	Spearman's Rank Correlation Coefficient	Kendall's Tau Rank Correlation Coefficient
Number of Daily Trips	0.0471	0.0657	0.0553
Number of Montly Trips	0.0015	0.0053	0.0046
Capacity	0.6629	0.6780	0.5839
Tempreture	0.2026	0.1911	0.1615
Wind Speed	0.1600	0.1722	0.1384
Day	0.2309	0.2654	0.2207
Hour	0.3474	0.3714	0.2937

Table 4.5 (continued) Relations between continuous variables and target variable after convert

Variable (No-Show)	Pearson Correlation Coefficient	Spearman's Rank Correlation Coefficient	Kendall's Tau Rank Correlation Coefficient
Line	0.0236	0.0182	0.0159
Special Day (for Departure)	-0.0021	0.0027	0.0024
Special Day (for Destination)	0.0371	0.0393	0.0343
Condition	0.4134	0.1820	0.1583

According to the obtained results, the correlation values for the continuous variables indicate the following:

There are weak relationships observed in the analysis. For the Number of Daily Trips, there is a weak positive relationship with No-Show. Although the correlation coefficients are low, an increase in the number of daily trips is associated with an increased likelihood of No-Show. The Number of Monthly Trips shows a very weak positive relationship with No-Show, indicating that this variable has a minimal effect on the likelihood of No-Show. Similarly, Line exhibits a very weak positive relationship with No-Show, with no significant relationship observed. Special Day (for Departure) also demonstrates a very weak relationship, suggesting that the presence of holidays at the departure location does not significantly affect the likelihood of No-Show. Lastly, Special Day (for Destination) shows a weak positive relationship, implying that holidays at the destination have a minimal effect on the likelihood of No-Show.

In terms of moderate relationships, Temperature shows a moderate positive relationship with No-Show. An increase in temperature is associated with an increased likelihood of No-Show, though this relationship is not strong. Wind Speed exhibits a weak positive relationship with No-Show, indicating that as wind speed increases, the likelihood of No-Show could also increase slightly. Day has a moderate positive relationship with No-Show, with an increase in Day being associated with an increase in the likelihood of No-Show. Hour demonstrates a moderate to strong positive

relationship with No-Show, suggesting that as Hour increases, the likelihood of No-Show could increase significantly. Finally, Condition shows a moderate positive relationship with No-Show, indicating that flight conditions have a noticeable effect on the likelihood of No-Show.

A strong relationship is observed with Capacity. There is a strong positive relationship between Capacity and No-Show, where an increase in capacity is associated with an increased likelihood of No-Show. This relationship is quite pronounced, underscoring the importance of considering capacity when predicting No-Show rates.

The results obtained from converting categorical variables to continuous variables can be comparatively evaluated against the results obtained before the conversion. Accordingly:

i. Numerical Variables

In both analyses, a positive and weak relationship was observed for Daily Trip Count. The correlation coefficients remained similar, indicating that the effect of this variable is stable.

The relationship for Monthly Trip Count was initially weakly negative but became positive and very weak after the conversion, indicating a minimal impact on the likelihood of No-Show.

A strong positive relationship was observed for Capacity in the first analysis, and this strong positive relationship was maintained in the second analysis. Although there were slight changes in the correlation coefficients, the overall trend remained the same.

The relationship for Temperature shifted from a weakly negative to a positive correlation in the second analysis. This change suggests that the effect of temperature on the likelihood of No-Show can vary depending on the quantification process used.

A positive and weak relationship was observed for Wind Speed in both analyses. The correlation coefficients remained similar, indicating that the effect of this variable is stable.

ii. Categorical Variables

The conversion of categorical variables into numerical forms did not affect the significance of the relationship between certain variables and No-Show. For example, variables like Line and Special Day (for both Departure and Destination) did not exhibit significant correlations in either analysis, indicating that these factors do not have a substantial impact on the No-Show rates. The consistent results highlight the stability of these relationships, irrespective of the data transformation approach used.

The effect of these variables remained constant. The comparative evaluation reveals that the quantification process can affect the observed relationships between variables and the likelihood of No-Show. Specifically, noticeable changes were observed for temperature and monthly trip count. However, the effects of capacity, daily trip count, wind speed, and departure/destination points remained largely stable. These findings underscore the importance of having a robust dataset and careful consideration of variable transformation in the analysis process.

The distributions of continuous independent variables are shown as follows: Daily Trip Count in Figure 4.1, Monthly Trip Count in Figure 4.2, Capacity in Figure 4.3, Temperature in Figure 4.4, and Wind Speed in Figure 4.5.

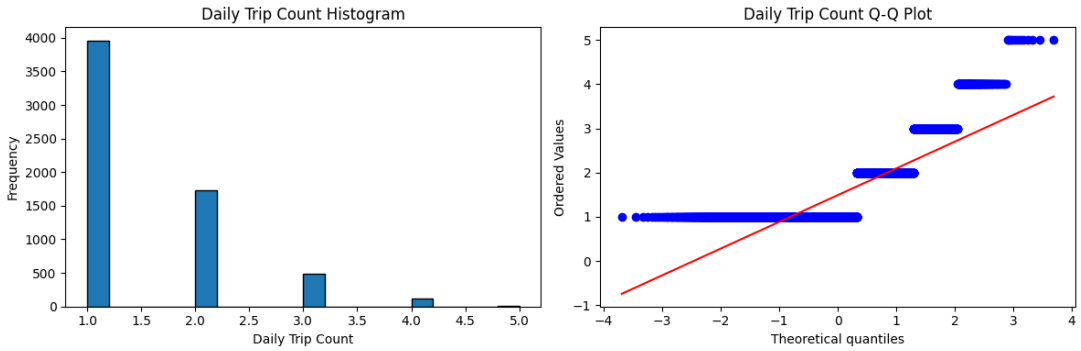

Figure 4.1 Daily Trip Count histogram

Figure 4.1 shows the distribution of the daily trip count. As can be seen from the graph, the majority of the trips are concentrated between 1 and 2. The number of observations with a daily trip count of 1 is approximately 4000, while the number of observations

with a daily trip count of 2 is approximately 1500. This indicates that most days in the dataset result in only one or two trips. The histogram exhibits a long tail to the right, indicating that the number of observations decreases rapidly as the trip count increases.

The Q-Q Plot demonstrates how well the daily trip count follows a theoretical normal distribution. Most of the points do not align along a straight line, indicating deviations from a normal distribution. It can be particularly noted that the data has a long tail to the right, suggesting positive skewness.

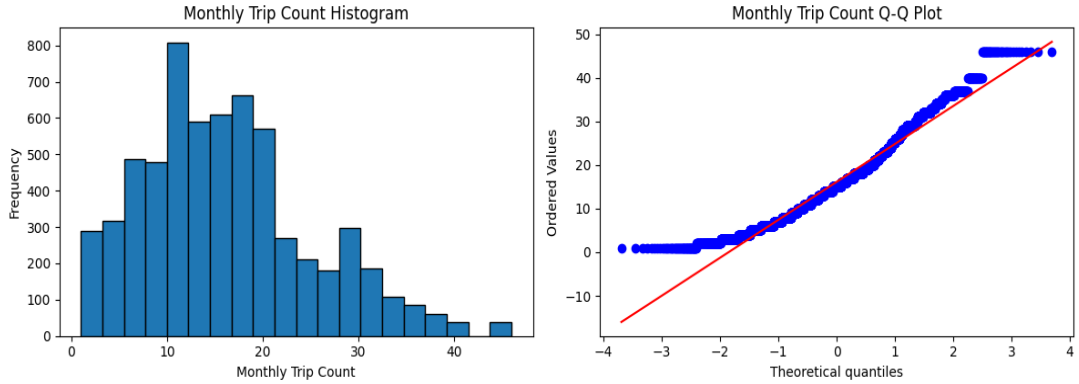

Figure 4.2 Montly Trip Count histogram

Figure 4.2 shows the distribution of the monthly trip count. It is observed that the monthly trip count is more varied and exhibits a more balanced distribution. The highest frequency is around 10 trips, and the distribution is symmetrical around this number. However, at higher trip counts (e.g., 30 and above), the frequency decreases, and the right tail of the graph is long.

The Q-Q Plot demonstrates how well the monthly trip count follows a normal distribution. Most of the points are aligned along a straight line, indicating that the data generally approximates a normal distribution. However, deviations from the normal distribution are observed at the extreme values, particularly at higher trip counts.

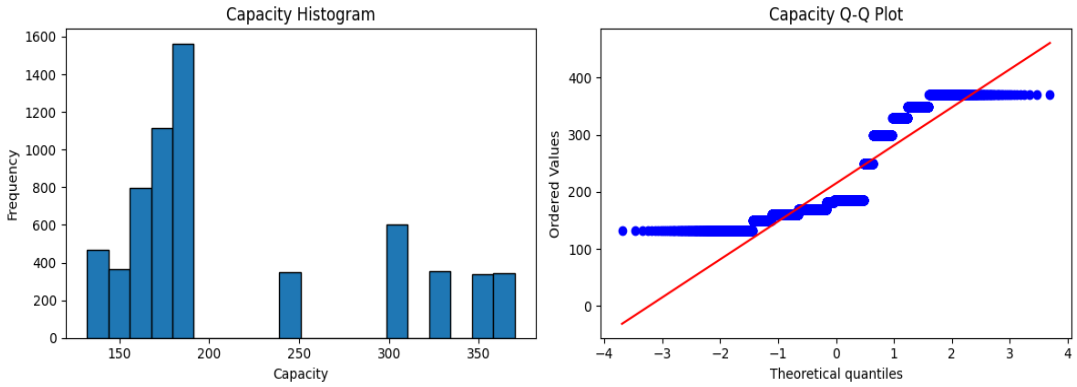

Figure 4.3 Capacity histogram

Figure 4.3 shows the distribution of bus capacity. As can be seen from the graph, capacities are concentrated in the range of 150-200. The highest frequency is observed at a capacity of 200. Buses with capacities over 300 have a notably low frequency. The distribution of capacities is asymmetrical and exhibits a long tail to the right.

The Q-Q Plot demonstrates how well the bus capacity follows a normal distribution. Most of the points do not align along a straight line, indicating deviations from a normal distribution. Notably, high-capacity values in the right tail significantly deviate from the normal distribution.

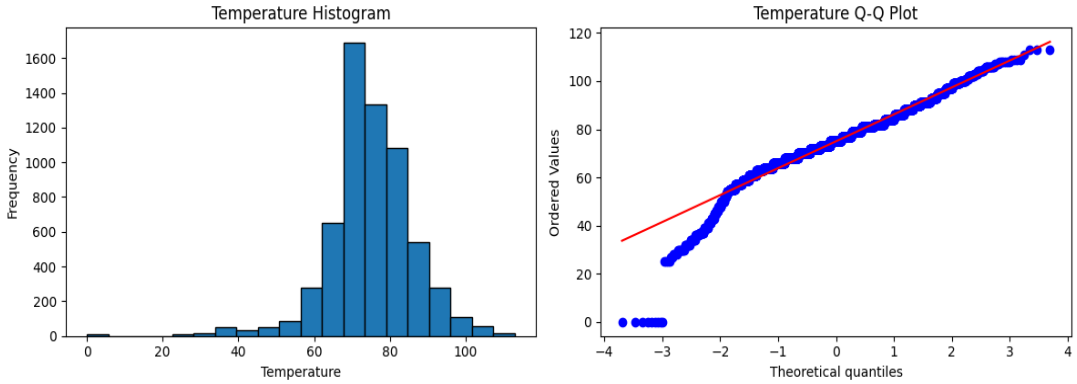

Figure 4.4 Temperature histogram

Figure 4.4 shows the distribution of temperature values. As can be seen from the graph, temperatures are concentrated in the range of 60 °F - 80 °F. The highest frequency is observed in the 70 °F - 80 °F range. The temperature distribution is generally symmetrical and concentrated around the average.

The Q-Q Plot demonstrates how well the temperature values follow a normal distribution. Most of the points are aligned along a straight line, indicating that the data generally conforms to a normal distribution. However, deviations from the normal distribution are observed at the extreme values, particularly at lower temperatures.

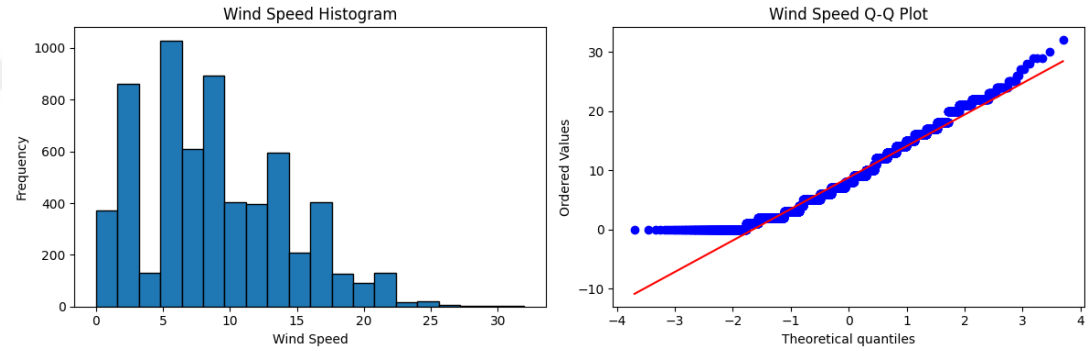

Figure 4.5 Wind Speed histogram

Figure 4.5 shows the distribution of wind speed. As can be seen from the graph, wind speeds are concentrated in the range of 0 ° mph - 10 ° mph, with the highest frequency around 5 ° mph. As wind speed increases, the frequency decreases, and the graph exhibits a long tail to the right.

The Q-Q Plot demonstrates how well the wind speed values follow a normal distribution. Most of the points are aligned along a straight line, indicating that the data generally approximates a normal distribution. However, deviations from the normal distribution are observed at the extreme values, particularly at higher wind speeds.

The distributions of categorical variables are shown as follows: Hour density in Figure 4.6, Day density in Figure 4.7, and Condition density in Figure 4.8.

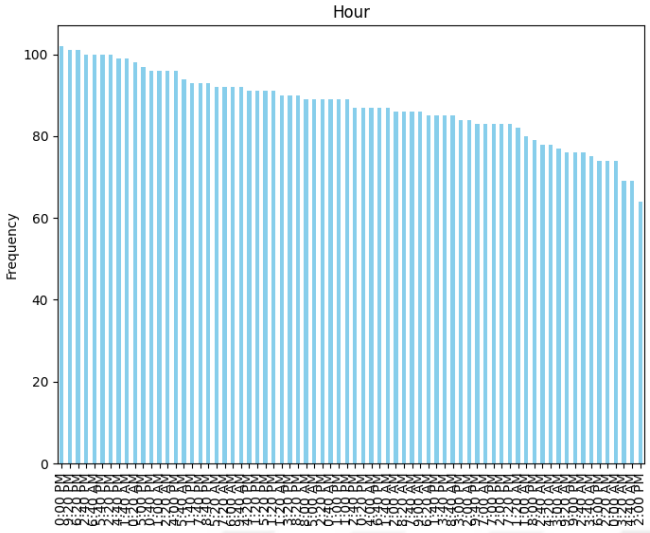

Figure 4.6 Hour frequency

Figure 4.6 shows the frequency of trips by hour. The distribution of trip counts by hour exhibits a consistent decreasing trend from hour 1 to hour 24. The busiest trip hours are observed to be in the early hours, with the highest number of trips occurring in the first hours after midnight, and the number of trips decreasing as the day progresses. This indicates that demand is higher during the early hours, leading to a concentration of trips at these times.

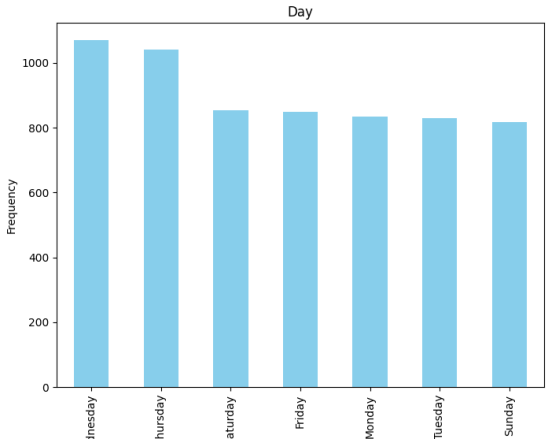

Figure 4.7 Day frequency

Figure 4.7 shows the frequencies of trips by day. Wednesday and Thursday have the highest number of trips, while Sunday has the lowest. This indicates that trips are more frequent in the middle of the week and decrease towards the weekend. Monday, Tuesday, and Friday have a balanced distribution of trips. This pattern suggests an increase in trip frequency in the middle of the week, with a decline towards the weekend.

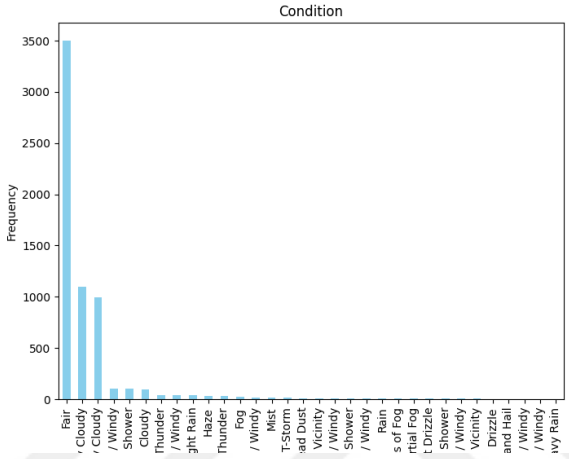

Figure 4.8 Condition frequency

Figure 4.8 shows the frequencies of trips by weather conditions. The highest frequency is observed under "Fair" conditions, with approximately 3500 trips. This is followed by "Cloudy" and "Partly Cloudy" conditions. Other weather conditions (e.g., rain, snow, fog) have much lower frequencies. This distribution indicates that most trips occur in fair and cloudy weather, while the number of trips significantly decreases under adverse weather conditions, or it suggests that adverse weather conditions were not prevalent during the observation period.

4.3.3.3 Converting Categorical Covariates

Categorical data must be converted into numerical values for use in machine learning. This conversion can be done in two main ways: by assigning specific numerical values to each categorical variable or by creating matrices to define a scalar magnitude.

For Departure and Destination, the OneHotEncoder method is used [224]. According to this method, a matrix is created where the value for the relevant airport is 1, and the values for other airports are 0. Since there are a total of 72 airports in the dataset, 72 new columns are created, with the relevant airport being assigned a value of 1 and the rest assigned a value of 0 (Eq. 4.20). To illustrate, let's assume the starting point is "AAA":

$$Departure = Dep_{AAA}\,\mathbf{1}_{\{Dep=AAA\}} + Dep_{BBB}\,\mathbf{1}_{\{Dep=BBB\}} + \cdots \quad \textbf{(4.20)}$$

By this way, the departure from "AAA" is represented by a “1” in the column corresponding to "AAA" and “0” in all other columns. The corresponding matrix representation is presented in Table 4.6.

Table 4.6 Example matrix structure for an encoded variable

Departure AAA	**Departure BBB**	**Departure CCC**	**…**	**Departure ZZZ**
1	0	0	…	0

Similarly, this process is applied to the Destination variable, resulting in a comprehensive matrix for both departure and destination airports, allowing the model to process the categorical data effectively.

4.3.4 Model Selection

The hypothesis and the obtained dataset aim to explain the relationship between many independent variables and the target variable (no-show). Therefore, the preferred models should be capable of learning this relationship (pattern). For this reason, 'learning' models were chosen. Another step involves selecting a suitable model for the target variable. Since the target variable consists of continuous (numerical) data, regression models were preferred. Six different regression models (Random Forest Regression, Gradient Boosting Regression, Linear Regression, Lasso Regression, Support Vector Regression) were tested.

4.3.5 Technical Specifications of the Computational Environment

All coding was conducted using the Python programming language. The computer utilized for this research is equipped with a 13th Gen Intel(R) Core (TM) i5-13420H processor operating at 2.10 GHz, 8.00 GB of RAM (7.73 GB usable), and runs on Windows 11 Pro. The system architecture is a 64-bit operating system based on an x64 processor. The computer brand is Lenovo.

4.4 Results and Sensitivy Analysis

4.4.1 Results

The results obtained from applying machine learning methods to the models are evaluated separately for each algorithm. To summarize the performance metrics of the models, tables containing MSE, RMSE, MAE, and R^2 values have been prepared. The performance values of the Random Forest Regression algorithm are shown in Table 4.7, the results of the Gradient Boosting Regression algorithm are shown in Table 4.8, the results of the Linear Regression algorithm are shown in Table 4.9, the results of the Ridge Regression algorithm are shown in Table 4.10, the results of the Lasso Regression algorithm are shown in Table 4.11, and the results of the Support Vector Regression algorithm are shown in Table 4.12.

4.4.1.1 Random Forest Regression

Table 4.7 Random Forest Regression performance metrics

Random Forest Regression	
Mean Squared Error	0.7320
Root Mean Squared Error	0.8556
Mean Absolute Error	0.5982
R-squared	0.8920

The Random Forest model exhibits quite low error rates in terms of MSE and RMSE values, indicating that the model's predictions do not deviate significantly from the actual values. The low MAE value further suggests that the majority of the errors are

small. An R^2 value of 0.8919 indicates that the model explains 89.2% of the variance in the dataset, implying that the model performs very well.

4.4.1.2 Gradient Boosting

Table 4.8 Gradient Boosting Regression performance metrics

Gradient Boosting Regression	
Mean Squared Error	0.5739
Root Mean Squared Error	0.7576
Mean Absolute Error	0.5622
R-squared	0.9153

The Gradient Boosting model has lower MSE and RMSE values compared to the Random Forest model, indicating that its predictions are more aligned with the actual values. The low MAE value further suggests that this model's error rates are small. An R^2 value of 0.9153 indicates that the model explains 91.5% of the variance in the dataset. Consequently, the Gradient Boosting model is observed to perform better than the Random Forest model.

4.4.1.3 Linear Regression

Table 4.9 Linear Regression performance metrics

Linear Regression	
Mean Squared Error	0.8078
Root Mean Squared Error	0.8988
Mean Absolute Error	0.6382
R-squared	0.8808

The Linear Regression model has higher MSE and RMSE values compared to the Random Forest and Gradient Boosting models, indicating that its predictions deviate more from the actual values. The MAE value of 0.6382 is also higher than that of the other models. An R^2 value of 0.8807 indicates that the model explains 88.1% of the variance in the dataset. Therefore, this model exhibits lower performance compared to the other two models.

4.4.1.4 Ridge Regression

Table 4.10 Ridge Regression performance metrics

Ridge Regression	
Mean Squared Error	1.6288
Root Mean Squared Error	1.2762
Mean Absolute Error	0.7585
R-squared	0.7597

The Ridge Regression model has higher MSE, RMSE, and MAE values compared to the Linear Regression model. This indicates that the predictions of the Ridge Regression model deviate more from the actual values and that the error rates are larger. An R^2 value of 0.7596 indicates that the model explains 75.9% of the variance in the dataset. This demonstrates that the Ridge Regression model has lower performance compared to the other models.

4.4.1.5 Lasso Regression

Table 4.11 Lasso Regression performance metrics

Lasso Regression	
Mean Squared Error	5.1711
Root Mean Squared Error	2.2740
Mean Absolute Error	1.6183
R-squared	0.2369

The Lasso Regression model has the highest MSE, RMSE, and MAE values among all models. This indicates that the model's predictions significantly deviate from the actual values and that the error rates are quite large. An R^2 value of 0.2369 indicates that the model explains only 23.7% of the variance in the dataset. This suggests that the model has the lowest performance and is not suitable for the dataset.

4.4.1.6 Support Vector Regression (SVR)

Table 4.12 Support Vector Regression performance metrics

Support Vector Regression Regression	
Mean Squared Error	1.6570
Root Mean Squared Error	1.2873
Mean Absolute Error	0.6381
R-squared	0.7555

The Support Vector Regression model has MSE and RMSE values similar to those of the Ridge Regression model, indicating significant deviations of predictions from the actual values. The MAE value is relatively low, suggesting that most error rates are small. An R^2 value of 0.7555 indicates that the model explains 75.5% of the variance in the dataset. This model performs similarly to the Ridge Regression model but has lower performance compared to other models.

In conclusion, the Gradient Boosting model has the lowest MSE, RMSE, and MAE values and the highest R^2 value, indicating the best performance among all models. The Random Forest model also performs well, ranking second after the Gradient Boosting model. Linear Regression and Ridge Regression models exhibit moderate performance, while the Lasso Regression model shows the lowest performance. The Support Vector Regression model performs similarly to the Ridge Regression model.

The research primarily focuses on predicting no-show rates by analyzing the impact of environmental and external factors on these rates using machine learning algorithms. Therefore, the aim is to develop the best regression model by identifying the features (independent variables) that best predict no-show rates. Regression models were constructed to investigate the minimum error rates and the extent to which independent variables explain the variance of the target variable.

Gradient Boosting and Random Forest models have demonstrated better performance compared to other models. They have lower error rates and explain a higher proportion of the target variable's variance through independent variables. This indicates that a large portion of the observed variability can be explained by the model's input

variables. Consequently, Gradient Boosting and Random Forest models are observed to have better generalization capabilities compared to other models.

Due to their superior performance in regression analyses conducted for the established model, it is recommended to use Gradient Boosting and Random Forest models for predicting no-show rates.

4.4.2 Model Validation

Model validation is essential for ensuring the reliability and accuracy of machine learning models. It helps identify and prevent overfitting, where the model performs well on training data but poorly on new, unseen data, and underfitting, where the model fails to capture the underlying patterns of the data. Proper validation ensures that the model generalizes well to real-world scenarios, providing dependable predictions and enhancing the model's overall performance.

In this study, K-Fold Cross-Validation was selected due to its robustness and efficiency. K-Fold Cross-Validation provided a comprehensive performance estimate by dividing the data into multiple subsets, ensuring that each data point was used for both training and testing. This method struck a balance between computational efficiency and thoroughness, making it ideal for accurately validating the models. The model validation results are summarized in Table 4.13.

Table 4.13 Model validation test results

Model	K-Fold Mean R^2	K-Fold Std Dev R^2	Test R^2
Random Forest	0.9058	0.0090	0.8920
Gradient Boosting	0.9222	0.0092	0.9153
Linear Regression	0.8782	0.0109	0.8808
Ridge Regression	0.7503	0.0469	0.7596
Lasso Regression	0.2551	0.0291	0.2369
SVM Regression	0.7397	0.0672	0.7555

The K-Fold Cross-Validation results indicate that Gradient Boosting has the highest mean R^2 value (0.9222) with a low standard deviation (0.0092), suggesting that it consistently performs well across different folds. Random Forest follows closely with a mean R^2 of 0.9058 and a standard deviation of 0.0090, indicating reliable

performance. Linear Regression also performs reasonably well with a mean R^2 of 0.8782 but shows a slightly higher standard deviation of 0.0109, suggesting more variability in performance. In contrast, Ridge Regression, Lasso Regression, and Support Vector Regression exhibit significantly lower mean R^2 values and higher standard deviations, indicating poorer and more inconsistent performance.

When evaluating the models on the test set, Gradient Boosting again demonstrates superior performance with the lowest Mean Squared Error (MSE) of 0.5739 and the highest R^2 value of 0.9153. Random Forest also performs well on the test set, with an MSE of 0.7320 and an R^2 value of 0.8920. Linear Regression, while showing reasonable performance with an R^2 value of 0.8808, has higher error metrics compared to Gradient Boosting and Random Forest. Ridge Regression, Lasso Regression, and Support Vector Regression continue to show lower performance on the test set, with Ridge and Lasso Regression particularly underperforming.

Comparing the results from K-Fold Cross-Validation and the test set, it is evident that Gradient Boosting is the best-performing model for predicting no-show rates in the airline industry. Its high and consistent R^2 values, along with low error metrics, confirm its robustness and reliability. Random Forest also emerges as a strong contender, offering reliable performance both in cross-validation and on the test set.

The consistent performance of Gradient Boosting across different validation techniques underscores its suitability for this prediction problem. The slight variability observed in Linear Regression's performance suggests it is less robust compared to ensemble methods like Gradient Boosting and Random Forest. The poor performance of Ridge and Lasso Regression indicates that these models are not well-suited for this particular dataset and problem.

4.4.3 Sensitivity Analysis

The sensitivity of the study to the independent variables (features) is crucial and must be analyzed. It is important to examine the contribution of the selected features to the model's outcome, as this can guide future studies. Therefore, in this section, the contribution of the features to the model will be examined.

The impact of each removed feature on model performance has been analyzed based on the algorithm, and the results, including the R^2 values for each algorithm, are presented in Table 4.14. The abbreviations used in the analysis refer to specific features removed from the dataset to assess their impact on model performance. The Base Model represents the model using all features. “D' Dep.” indicates the removal of the departure feature, while “D' Dest.” signifies the removal of the destination feature. When both the departure and destination features are removed, it is denoted as “D' Dep. & Dest.”. “D' Day” refers to the removal of the day feature, and “D’ Hour” indicates the removal of the hour feature. “D' Line” represents the removal of the flight line feature, which distinguishes between domestic and international flights. “D' DAoT” denotes the removal of the daily trip count feature, and “D' MAoT” signifies the removal of the monthly trip count feature. “D' Cap.” refers to the removal of the capacity feature, a critical determinant in the analysis. Special days are also considered, with “D' SDfD1” indicating the removal of the special day for departure destination, “D' SDfD2” for the special day for destination, and “D' SD” for special days affecting both departure and destination. Weather-related features include D' Temp.”, which denotes the removal of the temperature feature, “D' Cond.” for the removal of the condition (weather) feature, and “D' WS”, representing the removal of the wind speed feature.

Table 4.14 R^2 values according to models after feature extraction

	RF	**GB**	**LG**	**RR**	**LG**	**SVM**
Base Model	0.894751	0.916284	0.884387	0.763194	0.232895	0.75550
D' Dep.	0.879760	0.896575	0.857214	0.685784	0.065789	0.68670
D' Dest.	0.881424	0.897628	0.870290	0.712937	0.168207	0.74693
D' Dep. & Dest	0.918557	0.917498	0.850612	0.671521	0.148954	0.71084
D' Day	0.874663	0.891568	0.861291	0.690322	0.090366	0.66256
D’ Hour	0.877404	0.896355	0.864757	0.695401	0.201795	0.65342
D' Line	0.876591	0.899321	0.872632	0.700252	0.170837	0.74573
D' DAoT	0.878399	0.896876	0.868586	0.696481	0.220368	0.73879
D' MAoT	0.879358	0.904324	0.869849	0.752554	0.230247	0.74853
D' Cap.	0.299400	0.350299	0.226601	0.237100	0.099709	0.15779
D' SDfD1	0.878399	0.895812	0.873516	0.713826	0.249406	0.748674
D' SDfD2	0.879358	0.895643	0.874605	0.718485	0.251213	0.74852

Table 4.14 (continued) R^2 values according to models after feature extraction

	RF	**GB**	**LG**	**RR**	**LG**	**SVM**
D' SD	0.878526	0.895481	0.870299	0.705216	0.234807	0.74848
D' Temp.	0.884362	0.899293	0.873117	0.718374	0.243892	0.72911
D' Cond.	0.880873	0.899384	0.876293	0.719029	0.251815	0.89394
D' WS	0.677797	0.861587	0.835514	0.672973	0.172851	0.74884

The analysis of the R-squared (R^2) values across different models and feature extractions provides comprehensive insights into the importance of various features in predicting the dependent variable (no-show). Each model's performance varies significantly based on which features are included or excluded. This variability highlights the necessity of feature selection and its impact on model accuracy.

The removal of the capacity feature (D' Cap.) results in a dramatic decrease in the R^2 values across all models, indicating its crucial role in predictive performance. For instance, in the Random Forest Model, the R^2 value drops from 0.894751 to 0.299400, and in the Ridge Regression Model, it falls from 0.763194 to 0.237100. This consistent pattern underscores the critical role of capacity in accurately predicting no-show rates.

Similarly, the exclusion of departure (D' Dep.) and destination (D' Dest.) features affects the models' performance but to a lesser extent than capacity. For example, in the Gradient Boosting Model, the R^2 value changes from 0.916284 to 0.896575 with the removal of the departure feature and to 0.897628 with the removal of the destination feature. This suggests that while these features are important, they may have some degree of redundancy. Interestingly, the simultaneous exclusion of both features (D' Dep. & Dest.) leads to a slight increase in the R^2 value for some models, such as Random Forest, from 0.894751 to 0.918557.

Temporal features, such as the day of the trip (D' Day), the hour of the trip (D’ Hour), the daily trip count (D' DAoT), and the monthly trip count (D' MAoT), show varying degrees of importance. The exclusion of these features results in moderate decreases in R^2 values across models. For instance, in the Linear Regression Model, the R^2 value decreases from 0.884387 to 0.861291 with the removal of the day feature, to 0.864757 with the removal of the hour feature, to 0.868586 with the removal of the daily trip count, and to 0.869849 with the removal of the monthly trip count. These temporal

features are essential but do not impact the model's performance as significantly as capacity.

The impact of special day features varies across models. In the Ridge Regression Model, the removal of special day features results in slight changes in the R^2 values, indicating that these features provide some additional predictive power but are not as critical as capacity or departure/destination features. For example, the R^2 value changes from 0.763194 to 0.713826 with the removal of the first special day feature for departure destination (D' SDfD1).

Weather-related features, including temperature (D' Temp.), condition (D' Cond.), and wind speed (D' WS), show varying importance across models. In the Support Vector Regression Model, the removal of wind speed (D' WS) leads to a notable decrease in the R^2 value from 0.75550 to 0.74884, indicating its significant role. Conversely, the removal of temperature and condition features shows less impact on the models' performance.

The flight line feature (whether the flight is domestic or international) also plays a role in the predictive performance of the models. Its exclusion leads to moderate decreases in R^2 values across different models. For instance, in the Random Forest Model, the R^2 value decreases slightly from 0.894751 to 0.876591 when the flight line feature is removed.

The comprehensive analysis reveals that certain features, particularly capacity, departure, destination, and specific weather conditions, are vital for accurate prediction model. The consistent drop in model performance with the exclusion of capacity emphasizes its pivotal role.

The Random Forest and Gradient Boosting models exhibit higher resilience and better performance across various feature exclusions, indicating their robustness in handling complex datasets with potentially redundant or correlated features. Lasso Regression, with its feature selection capabilities, shows high sensitivity, suggesting it requires careful tuning and appropriate feature selection to avoid overfitting.

For airlines, understanding the importance of these features can guide data collection, preprocessing, and model development efforts. Focusing on critical features like capacity and weather conditions can enhance predictive accuracy and operational efficiency. This analysis can also inform dynamic pricing strategies, overbooking policies, and resource allocation, ultimately improving customer satisfaction and profitability.

In conclusion, the detailed examination of feature importance and model performance underscores the necessity of targeted feature selection and robust modeling techniques in predictive analytics.

Random Forest Algortihm

The R^2 values for the Random Forest Model demonstrate a high level of variance depending on the feature removed. The removal of the capacity feature (D' Cap.) leads to a significant drop in the R^2 value, indicating its critical importance. Conversely, the simultaneous removal of departure and destination features (D’ Dep. & Dest.) results in an improvement in the model's R^2 value, suggesting possible redundancy or noise in these features. The changes in R^2 values for the Random Forest Model are illustrated in Figure 4.9.

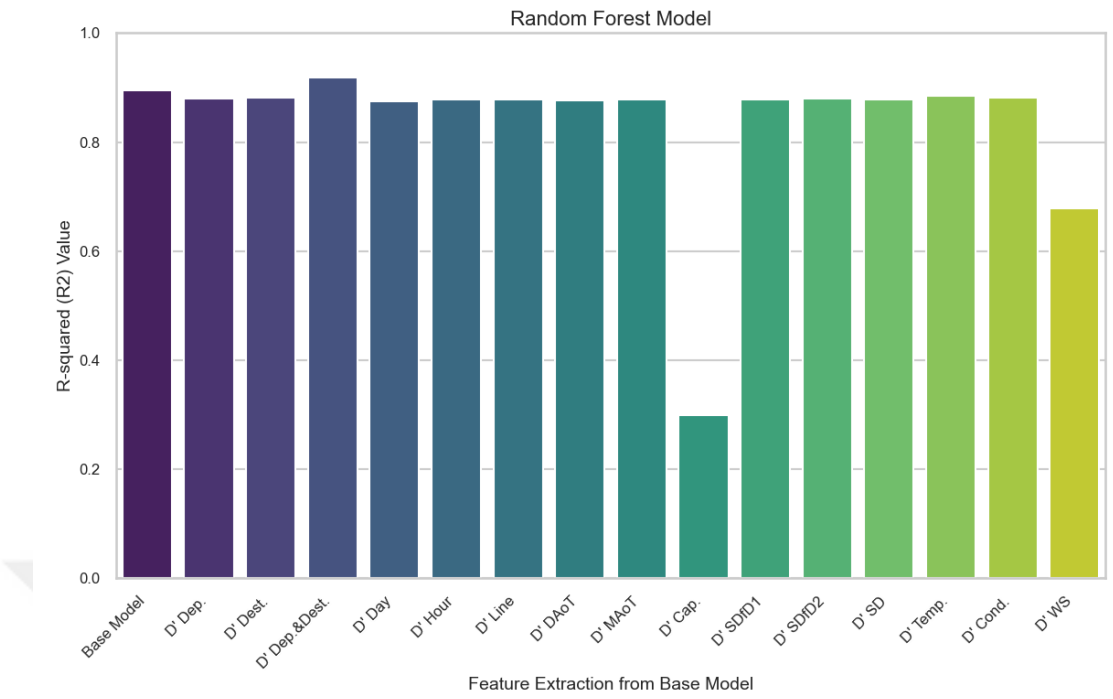

Figure 4.9 Random Forest R^2 value changes

Gradient Boosting Algortihm

Gradient Boosting maintains high R^2 values across most feature exclusions, except for the capacity feature (D' Cap.), where a significant drop is observed. The changes in R^2 values for the Gradient Boosting Model are illustrated in Figure 4.10.

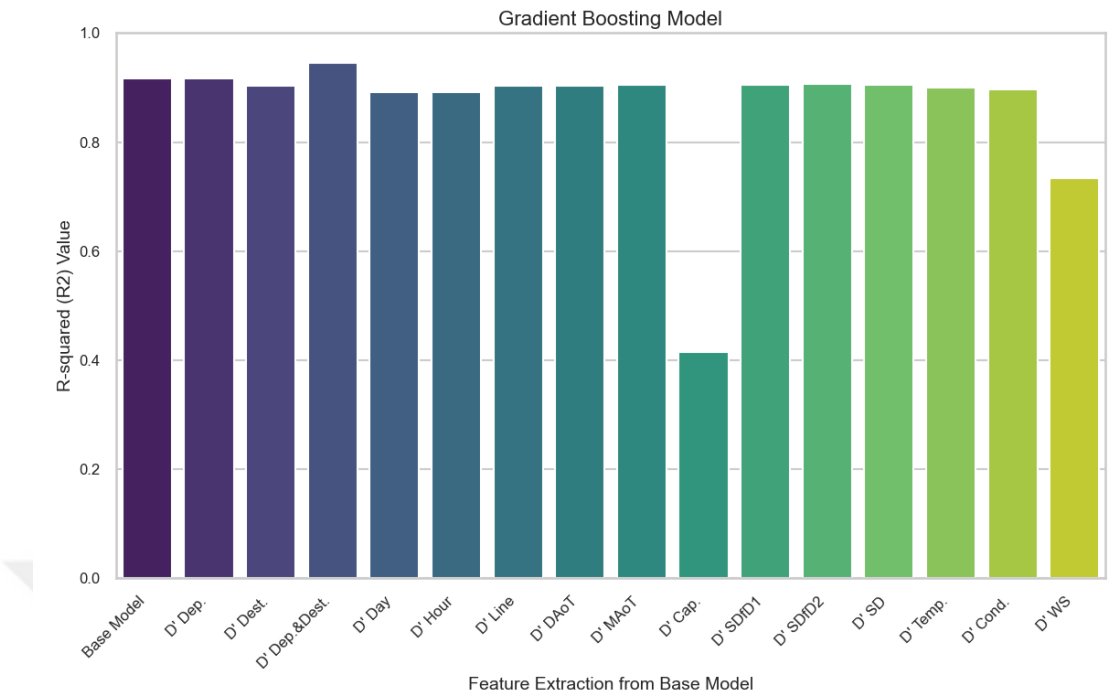

Figure 4.10 Gradient Boosting R^2 value changes

Linear Regression Algortihm

This model also underscores the critical role of the capacity feature (D' Cap.) with a substantial decrease in R^2 when it is removed. The changes in R^2 values for the Linear Regression Model are illustrated in Figure 4.11.

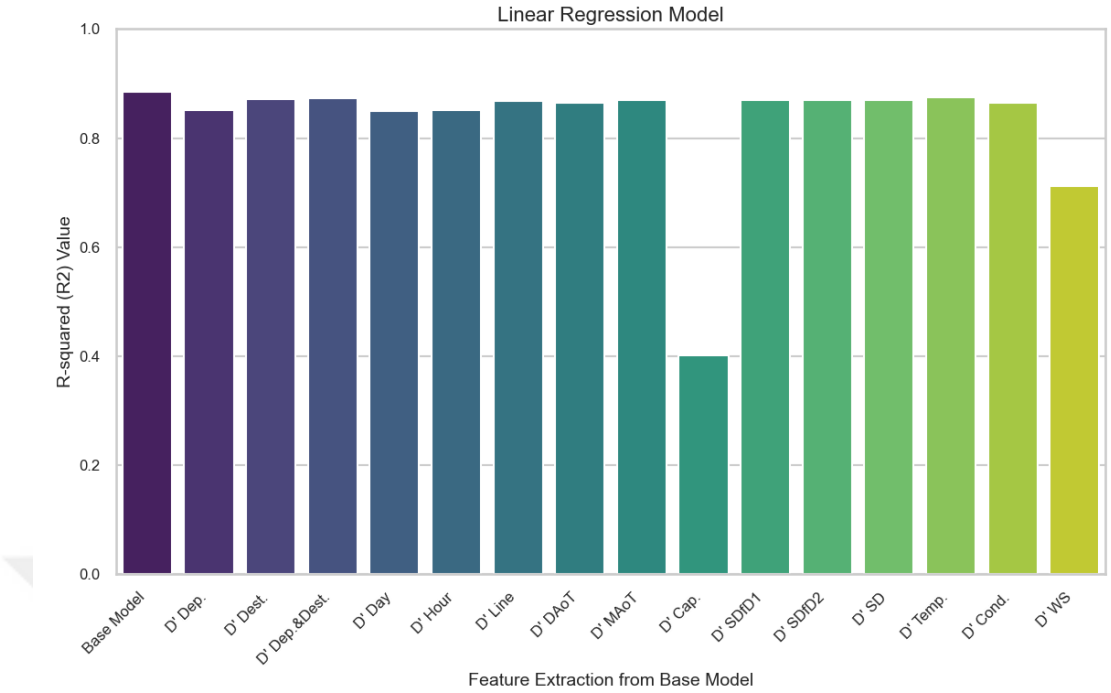

Figure 4.11 Linear Regression R^2 value changes

Ridge Regression Algortihm

Similar to the Random Forest Model, the Ridge Regression Model's R2 value drops significantly when the capacity feature (D' Cap.) is excluded, highlighting its importance. The changes in R^2 values for the Ridge Regression Model are illustrated in Figure 4.12.

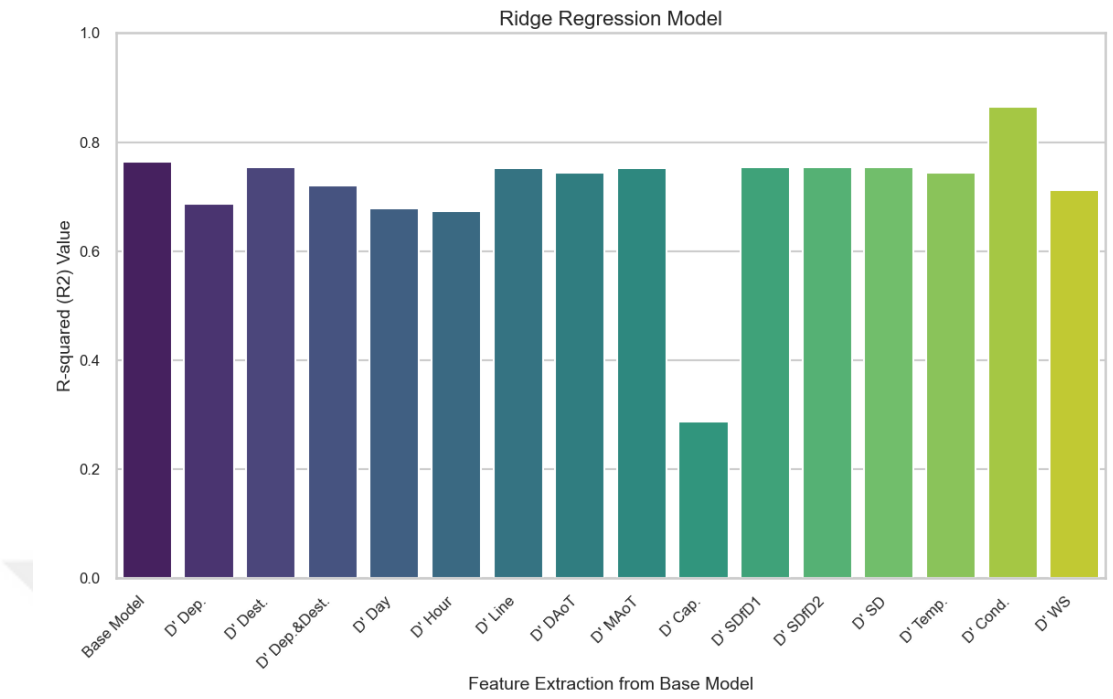

Figure 4.12 Ridge Regression R^2 value changes

Lasso Regression Algortihm

The Lasso Regression Model is highly sensitive to the removal of features, indicating that it may struggle with overfitting or that the data set requires careful feature selection for optimal performance. The changes in R^2 values for the Lasso Regression Model are illustrated in Figure 4.13.

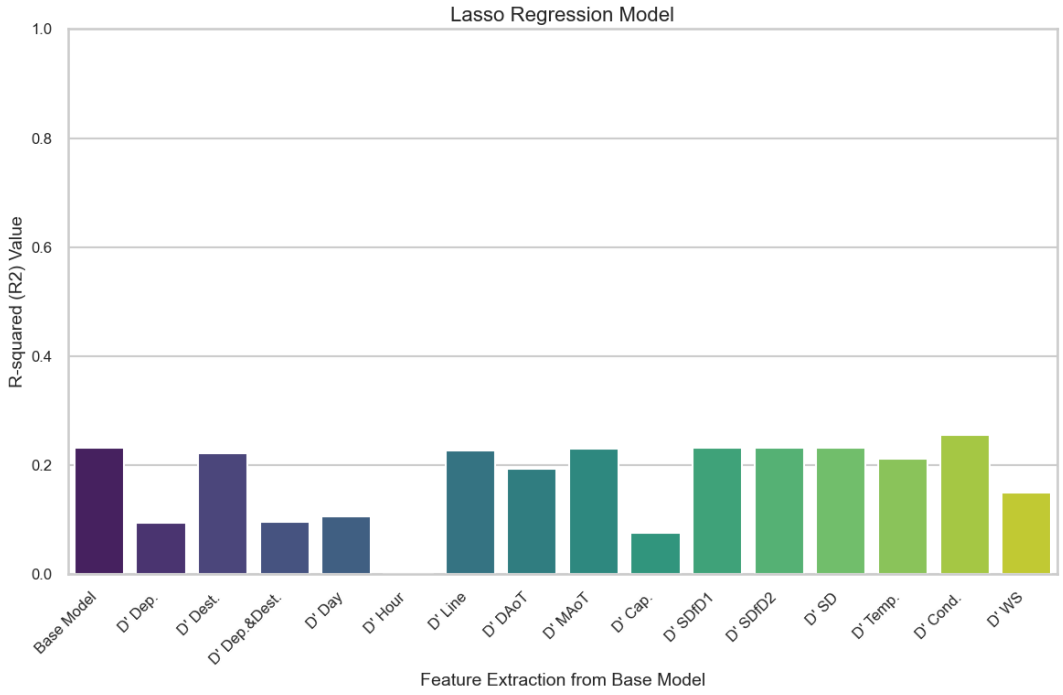

Figure 4.13 Lasso Regression R^2 value changes

Support Vector Regression Algortihm

Support Vector Regression also shows significant sensitivity to the removal of the capacity feature, indicating its importance in the model's predictive capability. The changes in R^2 values for the Support Vector Regression Model are illustrated in Figure 4.14.

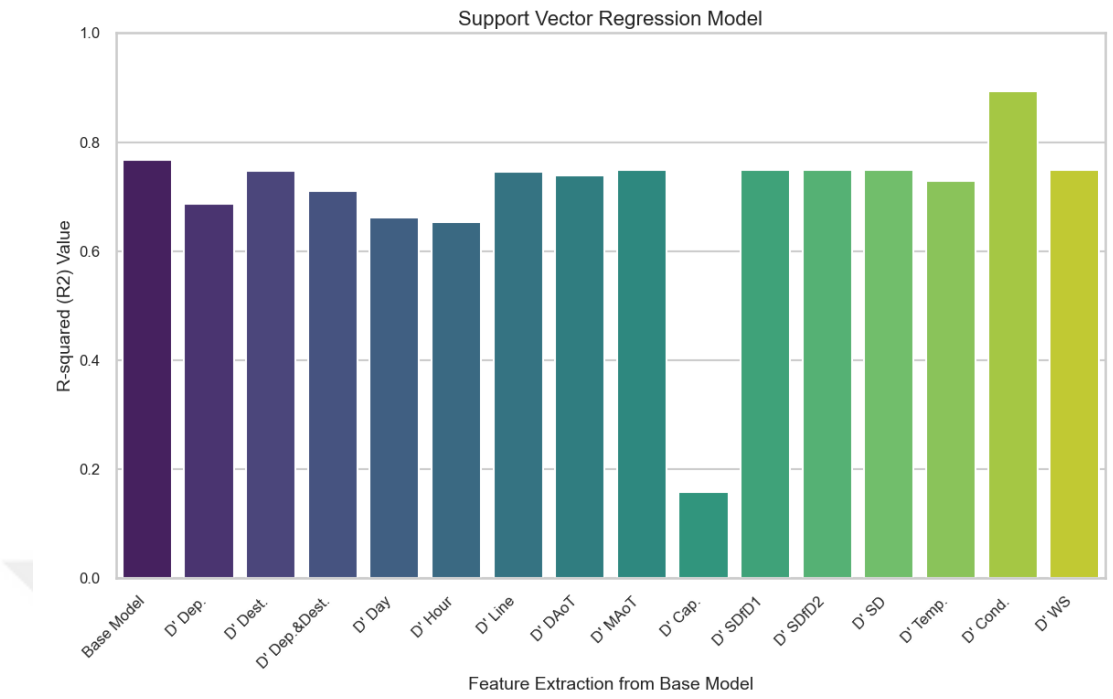

Figure 4.14 SVM Regression R^2 value changes

The Figure 4.15 show that, the negative performance change indicates a decrease in model accuracy (R^2) when the respective feature is excluded. Notably, the removal of the capacity feature (D' Cap.) leads to the most significant drop in performance, emphasizing its crucial role in the model. Similarly, the wind speed feature (D' WS) also shows a substantial negative impact on the model's accuracy when removed. Conversely, the removal of other features, such as the combined departure and destination feature (D' Dep. & Dest.) has a relatively minor effect, indicating potential redundancy or less critical importance in the model.

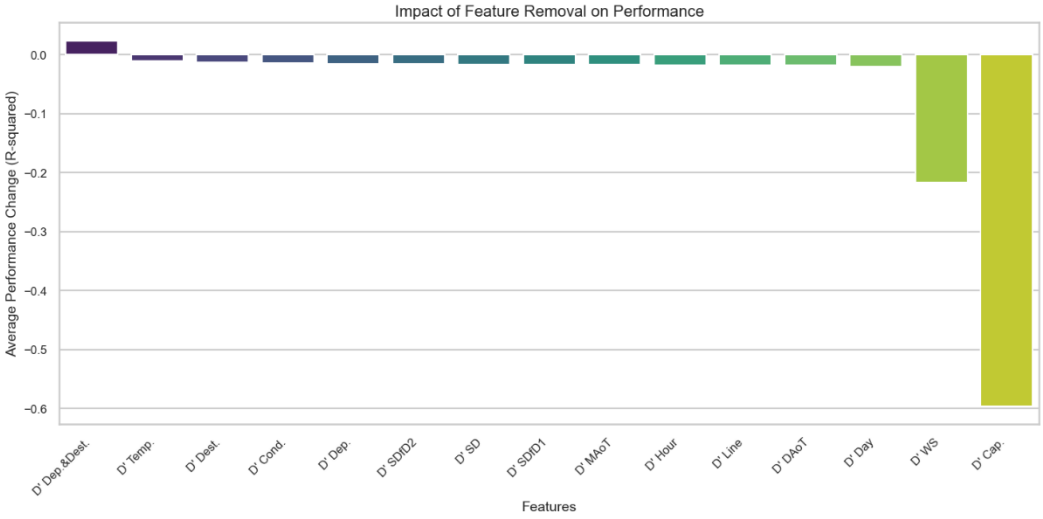

Figure 4.15 Impact of feature removal on performance

The Figure 4.16 shows that Gradient Boosting has the highest mean R-squared (R^2) value, indicating it performs the best among the algorithms tested, while Lasso Regression has the lowest mean R^2, suggesting it is the least effective model in this analysis.

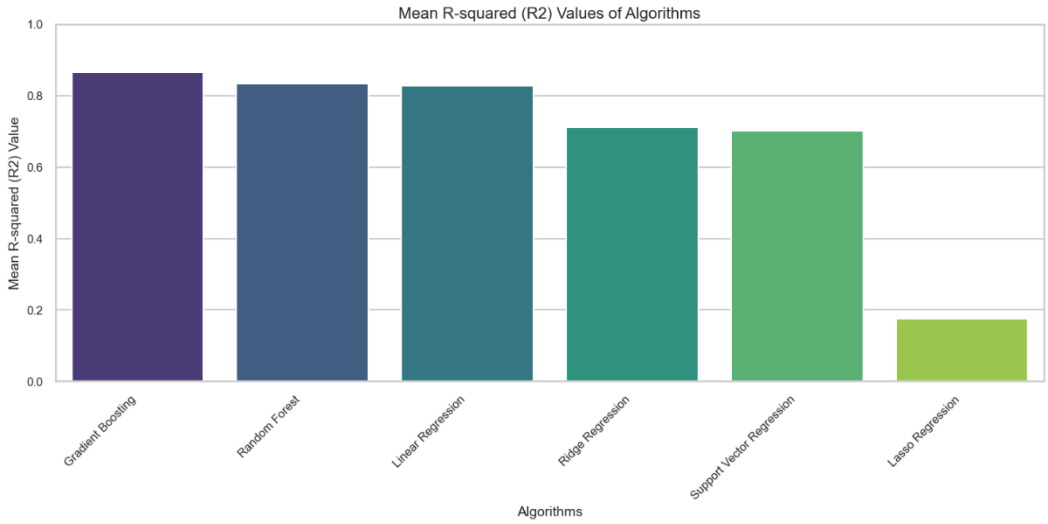

Figure 4.16 Algorithms average R^2 results

CHAPTER 5

CONCLUSION

This research primarily aims to examine the relationship between no-show rates and environmental-external factors. Subsequently, these examined factors are treated as independent variables to predict no-show rates. The goal is for the predicted no-show rates, obtained using machine learning methods, to contribute significantly to the aviation industry and current research in this field.

In alignment with the research topic, the aviation industry was initially examined. The historical development of the aviation industry and the impact of international developments on the sector were thoroughly analyzed. The challenges faced by companies in the aviation industry due to increasing competition, rapidly changing customer demands, and the capital power of multinational global companies were also investigated. Emphasis was placed on the importance of gaining a competitive advantage due to the industry's low profit margins and the pressures created by the global economy.

The significance of data science and machine learning, which are foundational to the major advancements in artificial intelligence, for stakeholders looking to increase market share was highlighted. Generally, machine learning and big data analytics have been successful in solving complex problems and have the potential to play a significant role in the aviation field.

All companies operating in the aviation sector (including those operating globally and locally across all domains) are making intense efforts to increase efficiency. All companies actively work to improve operational activities and reduce costs. Moreover, generating additional revenue in aviation is critical for creating a competitive advantage under variable conditions. Therefore, the concept of overbooking in the literature was addressed.

The importance of the overbooking concept for airlines was highlighted, and the detailed benefits for companies that can optimize their operational planning processes by overbooking were discussed. Currently, many airlines have overbooking policies, but optimizing these policies is quite challenging. Accurate calculation of the no-show rate is a critical factor that directly impacts the operational and financial performance of companies with overbooking policies. Although overbooking provides additional revenue, it requires allocating resources to develop the correct policy, which should also have the potential to generate revenue for the company.

There are disadvantages for small airlines in this regard. Companies without an extensive flight network may find it challenging to have sufficient data sets, resources, and manage the penalties (denied-boarding) that can occur if all passengers show up. The reputational loss, financial loss, and resources required for setting an overbooking policy make it difficult for small airlines to compete with larger ones. Small airlines, already facing many disadvantages compared to larger, more established airlines, may lose the competitive advantage of adapting quickly to dynamic conditions. Small airlines face higher operational and marketing costs, greater needs for working capital due to financial resource shortages, and limited resources for technology and innovation, coupled with a lack of trained personnel.

If they can optimize their overbooking policies, they can gain a significant advantage over their competitors. Conversely, failure to optimize these policies can lead to more significant problems. For large airlines, competing for market share with other large competitors, the potential to improve low profit margins and increase market share through improved overbooking policies is significant.

This study aims to provide a different perspective on the overbooking concept discussed in the literature. Most studies in the literature aim to determine an optimal upper limit for overbooking using historical data, including balancing the revenue from ticket sales and the penalties paid. No-show rates are generally predicted through statistical analysis of historical data. While optimizing these methods for airlines contributes to company revenues, there are also disadvantages.

Primarily, these studies aim to predict the total number of extra tickets that can be sold using historical no-show rates rather than predicting no-show rates for specific flights. This approach can produce efficient results for flights under similar conditions. However, the 'similar conditions' of flights are not detailed. For instance, two flights departing and arriving at the same locations with the same capacity on the same day, month, and hour but a year apart are considered similar based on historical data. The calculation of the optimal overbooking amount assumes similarity in no-show rates based on previous years' flights. Although the revenue from selling extra tickets and the penalty if all passengers show up are included in the calculations, the potential revenue can significantly decrease if the flights are not actually under 'similar conditions.' This highlights a significant potential disadvantage when flight information is explained only by apparent similarities.

In the literature, another method for predicting no-show rates involves using machine learning techniques. However, few studies use this method. Predicting no-show rates directly for flights (if the prediction is accurate) can potentially generate much higher revenue than estimating the maximum number of extra tickets that can be sold. The primary reason is that it allows for a dynamic overbooking model that adapts better to different conditions without idealizing flight conditions. In studies directly predicting no-show rates using PNR (Passenger Name Record) and flight information, machine learning methods have been applied. These methods align methodologically with the complexity of predicting no-show rates, which involve many independent variables.

However, models based on PNR and flight information have significant disadvantages. They predict no-show rates based on passenger information (age, gender, marital status) and flight information (ticket class, flight direction). Essentially, passengers are grouped into 'classes' based on their similarities. While the studies do not explicitly aim to classify passengers, the working principle of prediction models examines the relationship (pattern) between independent variables (inputs) and the dependent variable (no-show), conducting regression analysis accordingly. Thus, similar passenger characteristics are assumed to correspond to similar no-show potential. However, passenger information provided during ticket purchase only includes necessary procedural details, ignoring many unpredictable and variable factors that

can influence a passenger's ability to travel, such as sleep status, psychological state, travel reason, financial status, and job situation.

Additionally, the limited information provided by PNR numbers is not sufficient for making 'distinctive' classifications for no-show predictions. In simpler terms, the no-show behavior of a passenger in a particular age group and gender, even when considering additional information like marital status, cannot be distinctly explained when compared to thousands of other passengers with similar profiles. This makes it difficult to statistically approach no-show propensity. Furthermore, each person's unique characteristics, combined with their unique circumstances, make it challenging to develop a suitable method for predicting specific no-shows based on the limited information provided by a PNR number.

This study aims to analyze the environmental-external conditions that could cause no-shows on flights and make no-show predictions accordingly. Since environmental conditions (weather, traffic congestion) and external factors (flight time, holidays) occur under 'equal' conditions for all passengers, the aim is to overcome the disadvantages caused by internal passenger information in no-show predictions. Additionally, the similarity of flights under 'similar' conditions is ensured by considering environmental and external factors. In other words, two flights departing and arriving at the same locations with the same capacity on the same day, month, and hour but different years will be affected differently due to varying weather conditions. Similarly, the impact of holidays on passenger density (traffic, job-related effects) will cause differences. This study aims to analyze these similarities and differences under environmental-external conditions and make specific no-show predictions for each flight.

Moreover, the rapidly changing climate conditions in recent years, leading to deviations in historical weather data, make the analysis of 'weather conditions' and their impact on no-show rates even more critical. The frequent occurrence of historically anomalous weather events also allows for the 'biased' evaluation of their impact on no-show rates.

Machine learning methods facilitate the analysis of complex models with numerous independent variables. In this study, the model includes 13 different independent variables to analyze their impact on no-show rates. The main limitation of the established model is the inability to obtain no-show information, considered sensitive commercial data. Other information is public. Weather data, for example, contains actual data for the relevant dates.

As a result, parameters for no-show predictions were determined using publicly available data. These parameters were used to create a dataset containing no-show information. The target variable (no-show rates) in this dataset naturally contains continuous (numerical) data. Therefore, regression algorithms compatible with the dataset were chosen for the model's solution. The model was analyzed using six different Supervised Machine Learning Algorithms (Gradient Boosting Regression, Random Forest Regression, Linear Regression, Ridge Regression, Lasso Regression, and Support Vector Regression). Gradient Boosting achieved the lowest MSE (0.57), RMSE (0.76), and MAE (0.56) values, with the highest R^2 value of 0.91. The model produced the best results across all performance metrics, indicating that the Gradient Boosting Algorithm has the highest generalization capability for this model. The Random Forest model also performed well, producing results close to those of the Gradient Boosting model. It is important to note that the results may vary depending on the dataset used, but the analytical method employed is effective in examining different datasets. Ultimately, as the dataset changes, the results will also change.

In conclusion, Gradient Boosting and Random Forest models have shown better generalization capabilities compared to other models. This study focuses on the unique and critical need to improve operational planning processes in the aviation industry. It also aims to contribute to understanding the challenges in the aviation industry and addressing the limitations in calculating no-show rates. The importance and impact of accurately predicting no-show rates were thoroughly examined, and a model was presented using machine learning algorithms to address the problem. Due to their superior performance in regression analyses for the established model, it is recommended to use Gradient Boosting and Random Forest models for predicting no-show rates.

To ensure the reliability and accuracy of the machine learning models, k-Fold Cross-Validation was used, confirming their robustness and generalization capability. Gradient Boosting achieved a mean R^2 value of 0.9222 with a low standard deviation of 0.0092, and Random Forest followed closely with a mean R^2 value of 0.9058 and a standard deviation of 0.0090.

Considering the limitations of other methods in the literature, this study highlights the benefits of addressing environmental-external factors for the aviation sector. Accurate prediction of no-show rates using machine learning algorithms allows airlines to optimize their revenue management strategies and improve operational planning. An important finding of this study is the effectiveness of machine learning algorithms like Gradient Boosting and Random Forest in predicting no-show rates. These methods are recommended for addressing complex problems, such as no-show predictions, and providing methodological benefits to airlines. This enables companies to gain a competitive advantage and achieve more sustainable performance compared to the industry average.

In contrast to existing literature, this research explores how external factors, such as weather conditions, traffic density, flight time, and flight day, influence no-show rates. This study is pioneering in its focus on these external elements, considering the entire flight context and anomalies like public holidays and flight routes. Traditional prediction methods relying on historical data fall short in integrating environmental factors, often assuming a normal distribution of no-show rates for similar flights. To address this, Machine Learning methods are utilized, enabling the analysis of categorical data and identification of hidden patterns. This strategy aims to improve no-show rate predictions by considering variations in similar flights and providing more precise insights. Additionally, the limited information provided by PNR numbers is not sufficient for making distinctive classifications for no-show predictions. The no-show behavior of a passenger in a particular age group and gender, even when considering additional information like marital status, cannot be distinctly explained compared to thousands of other passengers with similar profiles. This makes it difficult to statistically approach no-show propensity, as each person's unique characteristics

and circumstances pose challenges for predicting specific no-shows based on limited PNR information.

Expanding the dataset to include more passenger behavior analysis could enhance the effectiveness of no-show predictions. Experts highlight that specific flight details (domestic-international, business-economy) significantly affect passenger no-show behavior, suggesting the study's scope could be broadened with such information. However, due to the personal data involved in flight information, airlines are reluctant to share this type of data, hindering contributions to the literature through published studies. Additionally, the study's scope could be expanded by comparing the performance of reinforcement learning methodologies, a type of agent training approach, with the supervised learning methods used in this study.

REFERENCES

[1] Schwab, K., "The fourth industrial revolution", Crown Currency, 2017.

[2] Schifrin, A. M. M., "Forbes 2024 Global 2000 List - The world's largest companies ranked", Retrieved from https://www.forbes.com/lists/global2000/, 2024, June 06.

[3] Duarte, F. "Amount of data created daily (2024)", Exploding Topics, https://explodingtopics.com/blog/data-generated-per-day, 2024, June 13.

[4] Talmage-Rostron, M. "How will artificial intelligence affect jobs 2024-2030", Nexford University, https://www.nexford.edu/insights/how-will-ai-affect-jobs, 2024, June 27.

[5] Flightradar. "Live flight tracker- Real-Time flight tracker Map", Flightradar24, https://www.flightradar24.com/data/statistics, n.d.

[6] US Department of Transportation, "Bumping & oversales", Retrieved from https://www.transportation.gov/individuals/aviation-consumer-protection/bumping-oversales, n.d.

[7] Pang, J., "Big Data Analysis Method based on Statistical Machine Learning: A Case Study of Financial Data Modeling", Proceedings of the 2nd International Conference on Information, Control and Automation, ICICA 2022, Vol. 2, pp. 2-4, Chongqing, China, March 2023.

[8] Sundarakani, B., Ajaykumar, A., & Gunasekaran, A. "Big data driven supply chain design and applications for blockchain: An action research using case study approach", Omega, 102, 102452, 2021.

[9] Nassar, A., & Kamal, M. "Machine Learning and Big Data analytics for Cybersecurity Threat Detection: A Holistic review of techniques and case studies", Journal of Artificial Intelligence and Machine Learning in Management, 5(1), 51-63, 2021.

[10] López-Martínez, F., Núñez-Valdez, E. R., García-Díaz, V., & Bursac, Z. "A case study for a big data and machine learning platform to improve medical decision support in population health management", Algorithms, 13(4), 102, 2020.

[11] Sohangir, S., Wang, D., Pomeranets, A., & Khoshgoftaar, T. M. "Big Data: Deep Learning for financial sentiment analysis", Journal of Big Data, 5(1), 1-25, 2018.

[12] Talón-Ballestero, P., González-Serrano, L., Soguero-Ruiz, C., Muñoz-Romero, S., & Rojo-Álvarez, J. L. "Using big data from customer relationship

management information systems to determine the client profile in the hotel sector", Tourism Management, 68, 187-197, 2018.

[13] Shi, Q., & Abdel-Aty, M. "Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways", Transportation Research Part C: Emerging Technologies, 58, 380-394, 2015.

[14] Luckow, A., Kennedy, K., Manhardt, F., Djerekarov, E., Vorster, B., Apon, A., "Automotive big data: Applications, workloads and infrastructures", 2015 IEEE International Conference on Big Data (Big Data), Vol. 1, pp. 1201-1210, IEEE, October 2015.

[15] Helgo, M. "Deep Learning and Machine Learning Algorithms for Enhanced Aircraft Maintenance and Flight Data Analysis", J. Robotics Spectrum, 1, 90-99, 2023.

[16] Adamopoulou, E., & Daskalakis, E. "Applications and Technologies of Big Data in the Aerospace Domain", Electronics, 12(10), 2225, 2023.

[17] Xu, X., Zhang, Y., McGrory, C. A., Wu, J., & Wang, Y. G. "Forecasting stock closing prices with an application to airline company data", Data Science and Management, 6(4), 239-246, 2023.

[18] Huang, C., & Cheng, X. "Estimation of aircraft fuel consumption by modeling flight data from avionics systems", Journal of Air Transport Management, 99, 102181, 2022.

[19] Gao, Z., & Mavris, D. N. Statistics and machine learning in aviation environmental impact analysis: A survey of recent progress, Aerospace, 9(12), 750, 2022.

[20] Park, S. H., Kim, M. Y., Kim, Y. J., & Park, Y. H. "A deep learning approach to analyze airline customer propensities: the case of South Korea", Applied Sciences, 12(4), 1916, 2022.

[21] Brunton, S. L., Nathan Kutz, J., Manohar, K., Aravkin, A. Y., Morgansen, K., Klemisch, J., et al. Data-driven aerospace engineering: reframing the industry with machine learning, AIAA Journal, *59(8)*, 2820-2847, 2021.

[22] Ning, S., Sun, J., Liu, C., Yi, Y., "Applications of deep learning in big data analytics for aircraft complex system anomaly detection", Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, Vol. 235, No. 5, pp. 923-940, 2021.

[23] Jiang, Y., Liu, Y., Liu, D., Song, H., "Applying machine learning to aviation big data for flight delay prediction", 2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), Vol. 1, pp. 665-672, IEEE, August 2020.

[24] Huo, J., Keung, K. L., Lee, C. K. M., Ng, K. K., Li, K. C., "The prediction of flight delay: Big data-driven machine learning approach", 2020 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Vol. 1, pp. 190-194, IEEE, December 2020.

[25] S Hwang, S., Kim, J., Park, E., & Kwon, S. J. "Who will be your next customer: A machine learning approach to customer return visits in airline services", Journal of Business Research, 121, 121-126, 2020.

[26] Gui, G., Liu, F., Sun, J., Yang, J., Zhou, Z., & Zhao, D. "Flight delay prediction based on aviation big data and machine learning", IEEE Transactions on Vehicular Technology, 69(1), 140-150, 2019.

[27] Oehling, J., & Barry, D. J. "Using machine learning methods in airline flight data monitoring to generate new operational safety knowledge from existing data", Safety science, 114, 89-104, 2019.

[28] Korfiatis, N., Stamolampros, P., Kourouthanassis, P., & Sagiadinos, V. "Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews", Expert Systems with Applications, 116, 472-486, 2019.

[29] Kumar, S., & Zymbler, M. "A machine learning approach to analyze customer satisfaction from airline tweets", Journal of Big Data, 6(1), 1-16, 2019.

[30] Ma, J., Tse, Y. K., Wang, X., & Zhang, M. "Examining customer perception and behaviour through social media research–An empirical study of the United Airlines overbooking crisis", Transportation Research Part E: Logistics and Transportation Review, 127, 192-205, 2019.

[31] Sternberg, F., Pedersen, K. H., Ryelund, N. K., Mukkamala, R. R., Vatrapu, R., "Analysing customer engagement of Turkish airlines using big social data", 2018 IEEE international congress on big data (BigData Congress), Vol. 1, pp. 74-81, IEEE, July 2018.

[32] Burmester, G., Ma, H., Steinmetz, D., & Hartmannn, S. "Big data and data analytics in aviation", Advances in Aeronautical Informatics: Technologies Towards Flight 4.0, 55-65, 2018.

[33] Tong, C., Yin, X., Wang, S., & Zheng, Z. "A novel deep learning method for aircraft landing speed prediction based on cloud-based sensor data". Future Generation Computer Systems, 88, 552-558, 2018.

[34] Li, S., Yang, Y., Yang, L., Su, H., Zhang, G., Wang, J., "Civil aircraft big data platform", 2017 IEEE 11th International Conference on Semantic Computing (ICSC), Vol. 1, pp. 328-333, IEEE, January 2017.

[35] Zheng, Y. J., Sheng, W. G., Sun, X. M., & Chen, S. Y. "Airline passenger profiling based on fuzzy deep machine learning", IEEE transactions on neural networks and learning systems, 28(12), 2911-2923, 2016.

[36] Garrow, L. A., & Koppelman, F. S. "Multinomial and nested logit models of airline passengers' no-show and standby behaviour", Journal of Revenue and Pricing Management, 3, 237-253, 2004.

[37] Petrović, B., & Milošević, P. "Decision support system for predicting the number of no-show passengers in airline industry", Tehnički vjesnik, 28(1), 123-134, 2021.

[38] Zenkert, D. "No-show forecast using passenger booking data", 2017.

[39] Neuling, R., Riedel, S., & Kalka, K. U. "New approaches to origin and destination and no-show forecasting: Excavating the passenger name records treasure", Journal of Revenue and Pricing Management, 3, 62-72, 2004.

[40] Taleb, I., Serhani, M. A., Dssouli, R., "Big data quality: A survey", 2018 IEEE International Congress on Big Data (BigData Congress), Vol. 1, pp. 166-173, IEEE, July 2018.

[41] Chen, D., & Zhao, H. "Data security and privacy protection issues in cloud computing", 2012 international conference on computer science and electronics engineering (Vol. 1, pp. 647-651). IEEE, 2012, March.

[42] Sun, Y., Zhang, J., Xiong, Y., & Zhu, G. "Data security and privacy in cloud computing", International Journal of Distributed Sensor Networks, 10(7), 190903, 2014.

[43] Bertino, E., & Ferrari, E. "Big data security and privacy", A comprehensive guide through the Italian database research over the last 25 years (pp. 425-439). Cham: Springer International Publishing, 2017.

[44] Marr, B. "Data strategy: how to profit from a world of big data, analytics and the internet of things", London: Kogan Page, 2017.

[45] Aktan, E. "Büyük veri: Uygulama alanları, analitiği ve güvenlik boyutu", Bilgi Yönetimi, 1(1), 1-22, 2018.

[46] Ali, A., Qadir, J., Rasool, R. U., Sathiaseelan, A., Zwitter, A., & Crowcroft, J. "Big data for development: applications and techniques", Big Data Analytics, 1, 1-24, 2016.

[47] Sagiroglu, S., & Sinanc, D. "Big data: A review", 2013 international conference on collaboration technologies and systems (CTS) (pp. 42-47). IEEE, 2013, May.

[48] Elgendy, N., & Elragal, A. "Big data analytics: a literature review paper", Advances in Data Mining. Applications and Theoretical Aspects: 14th Industrial Conference, ICDM 2014, St. Petersburg, Russia, July 16-20, 2014. Proceedings 14 (pp. 214-227). Springer International Publishing, 2014.

[49] Huang, Y., Zhu, F., Yuan, M., Deng, K., Li, Y., Ni, B., ... & Zeng, J. "Telco churn prediction with big data", Proceedings of the 2015 ACM SIGMOD international conference on management of data (pp. 607-618), 2015, May.

[50] Kitchin, R., & McArdle, G. "What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets", Big Data & Society, 3(1), 2053951716631130, 2016.

[51] Lu, R., Zhu, H., Liu, X., Liu, J. K., & Shao, J. "Toward efficient and privacy-preserving computing in big data era", IEEE Network, 28(4), 46-50, 2014.

[52] Hu, H., Wen, Y., Chua, T. S., & Li, X. "Toward scalable systems for big data analytics: A technology tutorial", IEEE access, 2, 652-687, 2014.

[53] McNeely, C. L., & Hahm, J. O. "The big (data) bang: Policy, prospects, and challenges", Review of Policy Research, 31(4), 304-310, 2014.

[54] Cappa, F., Oriani, R., Peruffo, E., & McCarthy, I. "Big data for creating and capturing value in the digitalized environment: unpacking the effects of volume, variety, and veracity on firm performance", Journal of Product Innovation Management, 38(1), 49-67, 2021.

[55] Wang, G., Gunasekaran, A., Ngai, E. W., & Papadopoulos, T. "Big data analytics in logistics and supply chain management: Certain investigations for research and applications", International journal of production economics, 176, 98-110, 2016.

[56] Ghasemaghaei, M., & Calic, G. "Assessing the impact of big data on firm innovation performance: Big data is not always better data", Journal of Business Research, 108, 147-162, 2020.

[57] Jin, X., Wah, B. W., Cheng, X., & Wang, Y. "Significance and challenges of big data research", Big data research, 2(2), 59-64, 2015.

[58] Johnson, J. S., Friend, S. B., & Lee, H. S. "Big data facilitation, utilization, and monetization: Exploring the 3Vs in a new product development process", Journal of Product Innovation Management, 34(5), 640-658, 2017.

[59] Hofmann, E. "Big data and supply chain decisions: the impact of volume, variety and velocity properties on the bullwhip effect", International Journal of Production Research, 55(17), 5108-5126, 2017.

[60] Liu, Y. "Big data and predictive business analytics", The Journal of Business Forecasting, 33(4), 40, 2014.

[61] Chen, M., Mao, S., Zhang, Y., & Leung, V. C. "Big data: related technologies, challenges and future prospects", Heidelberg: Springer, 2014.

[62] Dong, X. L., Srivastava, D., "Big data integration", 2013 IEEE 29th international conference on data engineering (ICDE), Vol. 1, pp. 1245-1248, IEEE, April 2013.

[63] Bendler, J., Wagner, S., Brandt, T., & Neumann, D. "Taming uncertainty in big data: Evidence from social media in urban areas", Business & Information Systems Engineering, 6, 279-288, 2014.

[64] Cai, L., & Zhu, Y. "The challenges of data quality and data quality assessment in the big data era", Data science journal, 14, 2-2, 2015.

[65] Rajaraman, V. "Big data analytics", Resonance, 21, 695-716, 2016.

[66] Cyganek, B., Graña, M., Krawczyk, B., Kasprzak, A., Porwik, P., Walkowiak, K., & Woźniak, M. "A survey of big data issues in electronic health record analysis", Applied Artificial Intelligence, 30(6), 497-520, 2016.

[67] Lee, I. "Big data: Dimensions, evolution, impacts, and challenges", Business horizons, 60(3), 293-303, 2017.

[68] Gholizadeh, H., Fazlollahtabar, H., & Khalilzadeh, M. "A robust fuzzy stochastic programming for sustainable procurement and logistics under hybrid uncertainty using big data", Journal of cleaner production, 258, 120640, 2020.

[69] Yang, T., & Jia, S. "Research on network security visualization under big data environment", 2016 International Computer Symposium (ICS) (pp. 660-662). IEEE, 2016, December.

[70] Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. "How 'big data'can make big impact: Findings from a systematic review and a longitudinal case study", International journal of production economics, 165, 234-246, 2015.

[71] Armor, F., Espinosa, J. A., Kaisler, S., & Money, W. "Big Data: Issues and Challenges Moving Forward", 46th Hawaii International Conference on System Sciences (HICSS), January, 7–10, 2013.

[72] Zhang, Y., Ren, J., Liu, J., Xu, C., Guo, H., & Liu, Y. "A survey on emerging computing paradigms for big data", Chinese Journal of Electronics, 26(1), 1-12, 2017.

[73] Zhou, L., Pan, S., Wang, J., & Vasilakos, A. V. "Machine learning on big data: Opportunities and challenges", Neurocomputing, 237, 350-361, 2017.

[74] Demchenko, Y., Grosso, P., De Laat, C., Membrey, P., "Addressing big data issues in scientific data infrastructure", 2013 International conference on collaboration technologies and systems (CTS), Vol. 1, pp. 48-55, IEEE, May 2013.

[75] Chen, M. S., Han, J., & Yu, P. S. "Data mining: an overview from a database perspective", IEEE Transactions on Knowledge and data Engineering, 8(6), 866-883, 1996.

[76] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. "CRISP-DM 1.0 step-by-step data mining guide", Technical Report, CRISP-DM, 2000.

[77] Saltz, J. "CRISP-DM is Still the Most Popular Framework for Executing Data Science Projects", Data Science Process Alliance. Retrieved from https://www.datascience-pm.com/crisp-dm-still-most-popular/, 2024, April 10.

[78] Hernández, M. A., & Stolfo, S. J. "Real-world data is dirty: Data cleansing and the merge/purge problem", Data mining and knowledge discovery, 2, 9-37, 1998.

[79] Rahm, E., & Do, H. H. "Data cleaning: Problems and current approaches", IEEE Data Eng. Bull., 23(4), 3-13, 2000.

[80] García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. "Big data preprocessing: methods and prospects", Big Data Analytics, 1(1), 1-22, 2016.

[81] Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. "Big data: Issues and challenges moving forward", 2013 46th Hawaii international conference on system sciences (pp. 995-1004). IEEE, 2013, January.

[82] Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M., & Widom, J. "Challenges and opportunities with Big Data 2011-1", 2011.

[83] Dean, J., & Ghemawat, S. "MapReduce: simplified data processing on large clusters", Communications of the ACM, 51(1), 107-113, 2008.

[84] Luengo, J., García-Gil, D., Ramírez-Gallego, S., García, S., & Herrera, F. "Big data preprocessing", Cham: Springer, 2020.

[85] García, S., Luengo, J., & Herrera, F. "Data preprocessing in data mining" (Vol. 72, pp. 59-139). Cham, Switzerland: Springer International Publishing, 2015.

[86] Gandomi, A., & Haider, M. "Beyond the hype: Big data concepts, methods, and analytics", International journal of information management, 35(2), 137-144, 2015.

[87] Alasadi, S. A., & Bhaya, W. S. "Review of data preprocessing techniques in data mining", Journal of Engineering and Applied Sciences, 12(16), 4102-4107, 2017.

[88] Ramírez-Gallego, S., García, S., Mouriño-Talín, H., Martínez-Rego, D., Bolón-Canedo, V., Alonso-Betanzos, A., ... & Herrera, F. "Data discretization:

taxonomy and big data challenge", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 6(1), 5-21, 2016.

[89] World Bank Open Data. "World Bank Open Data", https://data.worldbank.org/indicator/NE.EXP.GNFS.CD, n.d.

[90] Marshall, A., Mueck, S., & Shockley, R. "How leading organizations use big data and analytics to innovate", Strategy & Leadership, 43(5), 32-39, 2015.

[91] The Economist. "The world's most valuable resource is no longer oil, but data", The Economist. Retrieved from https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data, 2017, May 6.

[92] Statista. "Global big data industry market size 2011-2027", Retrieved from https://www.statista.com/statistics/254266/global-big-data-market-forecast/, 2024, February 13.

[93] Provost, F., & Fawcett, T. "Data science and its relationship to big data and data-driven decision making", Big data, 1(1), 51-59, 2013.

[94] Uzun, A. M. "Deregülasyon ve Ekonomik Etkileri: Abd ve Türkiye Havayolu Taşımacılığı Deneyimi Üzerine Bir İnceleme", Kafkas Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 10(19), 474-489, 2019.

[95] United States Congress. "S.2493- 95th Congress (1977-1978): Airline Deregulation Act", Retrieved from https://www.congress.gov/bill/95th-congress/senate-bill/2493, 1978, October 24.

[96] Mevzuat Bilgi Sistemi. "Turkish Civil Aviation Act No. 2920", Retrieved from https://www.mevzuat.gov.tr/mevzuat?MevzuatNo=2920&MevzuatTur=1&MevzuatTertip=5, n.d.

[97] European Commission. "Speech/96/50: European liberalisation", Retrieved from https://ec.europa.eu/commission/presscorner/detail/en/SPEECH_96_50, 1996, May 20.

[98] Butcher, L. "Aviation: European liberalisation, 1986-2002 - House of Commons Library", Retrieved from https://commonslibrary.parliament.uk/research-briefings/sn00182/, 2024, July 3.

[99] Doganis, R. "Flying off course: The economics of international airlines", Routledge, 2013.

[100] Vikipedi. "Hava korsanlığı", Retrieved from https://tr.wikipedia.org/wiki/Hava_korsanl%C4%B1%C4%9F%C4%B1, 2007, January 22.

[101] Davenport, T. H., & Dyché, J. "Big data in big companies", International Institute for Analytics, 3(1-31), 2013.

[102] Gosavi, A. "A reinforcement learning algorithm based on policy iteration for average reward: Empirical results with yield management and convergence analysis", Machine Learning, 55, 5-29, 2004.

[103] Gosavi, A., Bandla, N., & Das, T. K. "A reinforcement learning approach to a single leg airline revenue management problem with multiple fare classes and overbooking", IIE transactions, 34(9), 729-742, 2002.

[104] Belobaba, P., Odoni, A., & Barnhart, C. (Eds.). "The global airline industry", John Wiley & Sons, 2015.

[105] Phillips, R. L. "Pricing and revenue optimization", Stanford University Press, 2021.

[106] Mumbower, S., Garrow, L. A., & Higgins, M. J. "Estimating flight-level price elasticities using online airline data: A first step toward integrating pricing, demand, and revenue optimization", Transportation Research Part A: Policy and Practice, 66, 196-212, 2014.

[107] Srinidhi, S., & Manrai, A. K. "A conceptual model for demand forecasting and service positioning in the airline industry", Journal of Modelling in Management, 8(1), 123-139, 2013.

[108] Sivrikaya, O., & Tunç, E. "Demand forecasting for domestic air transportation in Turkey", The Open Transportation Journal, 7(1), 2013.

[109] Li, M. Z., Zhang, A., & Zhang, Y. "Airline seat allocation competition", International transactions in operational research, 15(4), 439-459, 2008.

[110] Cizaire, C., & Belobaba, P. "Joint optimization of airline pricing and fare class seat allocation", Journal of Revenue and Pricing Management, 12, 83-93, 2013.

[111] Raza, A. S., & Akgunduz, A. "An airline revenue management pricing game with seat allocation", International Journal of Revenue Management, 2(1), 42-62, 2008.

[112] Kyparisis, G. J., & Koulamas, C. "Optimal pricing and seat allocation for a two-cabin airline revenue management problem", International Journal of Production Economics, 201, 18-25, 2018.

[113] Banerjee, N., Morton, A., & Akartunalı, K. "Passenger demand forecasting in scheduled transportation", European Journal of Operational Research, 286(3), 797-810, 2020.

[114] Yu, B., Guo, Z., Asian, S., Wang, H., & Chen, G. "Flight delay prediction for commercial air transport: A deep learning approach", Transportation Research Part E: Logistics and Transportation Review, 125, 203-221, 2019.

[115] Yazdi, M. F., Kamel, S. R., Chabok, S. J. M., & Kheirabadi, M. "Flight delay prediction based on deep learning and Levenberg-Marquart algorithm", Journal of Big Data, 7, 1-28, 2020.

[116] Khanmohammadi, S., Tutun, S., & Kucuk, Y. "A new multilevel input layer artificial neural network for predicting flight delays at JFK airport", Procedia Computer Science, 95, 237-244, 2016.

[117] Belcastro, L., Marozzo, F., Talia, D., & Trunfio, P. "Using scalable data mining for predicting flight delays", ACM Transactions on Intelligent Systems and Technology (TIST), 8(1), 1-20, 2016.

[118] Lai, K. K., & Ng, W. L. "A stochastic approach to hotel revenue optimization", Computers & Operations Research, 32(5), 1059-1072, 2005.

[119] Phumchusri, N., & Maneesophon, P. "Optimal overbooking decision for hotel rooms revenue management", Journal of Hospitality and Tourism Technology, 5(3), 261-277, 2014.

[120] Fouad, A. M., Atiya, A. F., Saleh, M., & Bayoumi, A. E. M. M. "A simulation-based overbooking approach for hotel revenue management", 2014 10th international computer engineering conference (ICENCO) (pp. 61-69). IEEE, 2014, December.

[121] Saito, T., Takahashi, A., Koide, N., & Ichifuji, Y. "Application of online booking data to hotel revenue management", International Journal of Information Management, 46, 37-53, 2019.

[122] Toh, R. S., & Dekay, F. "Hotel room-inventory management: an overbooking model", Cornell Hotel and Restaurant Administration Quarterly, 43(4), 79-90, 2002.

[123] Tse, T. S., & Poon, Y. T. "Modeling no-shows, cancellations, overbooking, and walk-ins in restaurant revenue management", Journal of foodservice business research, 20(2), 127-145, 2017.

[124] Al-Mashraie, M. "Simulation-based evaluation of no-show prediction using deep learning neural networks", State University of New York at Binghamton, 2016.

[125] Carreras-García, D., Delgado-Gómez, D., Llorente-Fernández, F., & Arribas-Gil, A. "Patient no-show prediction: A systematic literature review", Entropy, 22(6), 675, 2020.

[126] Huang, Y., & Hanauer, D. A. "Patient no-show predictive model development using multiple data sources for an effective overbooking approach", Applied clinical informatics, 5(03), 836-860, 2014.

[127] Rothstein, M. "An airline overbooking model", Transportation Science, 5(2), 180-192, 1971.

[128] Garrow, L. A., & Koppelman, F. S. "Multinomial and nested logit models of airline passengers' no-show and standby behaviour", Journal of Revenue and Pricing Management, 3, 237-253, 2004.

[129] Klophaus, R., & Pölt, S. "Airline overbooking with dynamic spoilage costs", Journal of Revenue and Pricing Management, 6, 9-18, 2007.

[130] Subramanian, J., Stidham Jr, S., & Lautenbacher, C. J. "Airline yield management with overbooking, cancellations, and no-shows", Transportation science, 33(2), 147-167, 1999.

[131] Bodily, S. E., & Pfeifer, P. E. "Overbooking decision rules", Omega, 20(1), 129-133, 1992.

[132] Suzuki, Y. "An empirical analysis of the optimal overbooking policies for US major airlines", Transportation Research Part E: Logistics and Transportation Review, 38(2), 135-149, 2002.

[133] Chatwin, R. E. "Multiperiod airline overbooking with a single fare class", Operations Research, 46(6), 805-819, 1998.

[134] Feng, Y., & Xiao, B. "A continuous-time seat control model for single-leg flights with no-shows and optimal overbooking upper bound", European Journal of Operational Research, 174(2), 1298-1316, 2006.

[135] Alstrup, J., Boas, S., Madsen, O. B., & Vidal, R. V. "Booking policy for flights with two types of passengers", European Journal of Operational Research, 27(3), 274-288, 1986.

[136] Zhang, C., Guo, C., & Yi, S. "Airline overbooking problem with uncertain no-shows", Journal of Applied Mathematics, 2014, 2014.

[137] Klophaus, R., & Pölt, S. "Overbooking: A sacred Cow ripe for slaughter?", 9th Annual Air Transport Research Society Conference Papers (ATRS, Rio de Janeiro, 3-7 July, 2005)./Air Transport Research Society (ed.), 2005.

[138] Alpaydın, E. "Yapay öğrenme: yeni yapay zeka", 2020.

[139] Christian, B., & Griffiths, T. "Algorithms to live by: the computer science of human decisions", First international edition. New York, Henry Holt and Company, 2016.

[140] Sarker, I. H. "Machine learning: Algorithms, real-world applications and research directions", SN computer science, 2(3), 160, 2021.

[141] El Naqa, I., & Murphy, M. J. "What is machine learning?", Springer International Publishing (pp. 3-11), 2015.

[142] Ij, H. "Statistics versus machine learning", Nat Methods, 15(4), 233, 2018.

[143] Zhou, Z. H. "Three perspectives of data mining", 2003.

[144] Muhamedyev, R. "Machine learning methods: An overview", Computer modelling & new technologies, 19(6), 14-29, 2015.

[145] Zhang, R., & Xi, Z. "Research on Anomaly Identification and Screening and Metallogenic Prediction Based on Semisupervised Neural Network", Computational Intelligence and Neuroscience, 2022(1), 8745036, 2022.

[146] Schmidhuber, J. "Deep learning in neural networks: An overview", Neural networks, 61, 85-117, 2015.

[147] Janiesch, C., Zschech, P., & Heinrich, K. "Machine learning and deep learning", Electronic Markets, 31(3), 685-695, 2021.

[148] Rokach, L., & Maimon, O. "Clustering methods", Data mining and knowledge discovery handbook, 321-352, 2005.

[149] Sykes, A. O. "An introduction to regression analysis", 1993.

[150] Joseph, V. R., & Vakayil, A. "SPlit: An optimal method for data splitting", Technometrics, 64(2), 166-176, 2022.

[151] Raschka, S. "Model evaluation, model selection, and algorithm selection in machine learning", arXiv preprint arXiv:1811.12808, 2018.

[152] Hawkins, D. M. "The problem of overfitting", Journal of chemical information and computer sciences, 44(1), 1-12, 2004.

[153] Ying, X. "An overview of overfitting and its solutions", Journal of physics: Conference series (Vol. 1168, p. 022022). IOP Publishing, 2019, February.

[154] Jabbar, H., & Khan, R. Z. "Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study)", Computer Science, Communication and Instrumentation Devices, 70(10.3850), 978-981, 2015.

[155] Domingos, P. "A unified bias-variance decomposition", Proceedings of 17th international conference on machine learning (pp. 231-238). Morgan Kaufmann Stanford, 2000, June.

[156] Kelleher, J. D., & Tierney, B. "Veri bilimi", 2020.

[157] Moor, J. H. "An analysis of the Turing test", Philosophical Studies, 30(4), 249-257, 1976.

[158] McCarthy, J., Samuel, A. L., Feigenbaum, E., & Lederberg, J. "Project technical report (p. 0084)", Stanford University, Computer Science Department; [available from the Clearinghouse for Federal Scientific and Technical Information, Springfield, Va.], 1971.

[159] Stanton, J. M. "Galton, Pearson, and the peas: A brief history of linear regression for statistics instructors", Journal of Statistics Education, 9(3), 2001.

[160] Nasteski, V. "An overview of the supervised machine learning methods", Horizons. b, 4(51-62), 56, 2017.

[161] Czepiel, S. A. "Maximum likelihood estimation of logistic regression models: theory and implementation", Available at czep. net/stat/mlelr. pdf, 83, 2002.

[162] Sur, P., & Candès, E. J. "A modern maximum-likelihood theory for high-dimensional logistic regression", Proceedings of the National Academy of Sciences, 116(29), 14516-14525, 2019.

[163] Patle, A., & Chouhan, D. S. "SVM kernel functions for classification", 2013 International conference on advances in technology and engineering (ICATE) (pp. 1-9). IEEE, 2013, January.

[164] Dayan, P., Sahani, M., & Deback, G. "Unsupervised learning", The MIT encyclopedia of the cognitive sciences, 857-859, 1999.

[165] Jain, A. K., Murty, M. N., & Flynn, P. J. "Data clustering: a review", ACM computing surveys (CSUR), 31(3), 264-323, 1999.

[166] Sapkal, S. D., Kakarwal, S. N., & Revankar, P. S. "Analysis of classification by supervised and unsupervised learning", International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007) (Vol. 1, pp. 280-284). IEEE, 2007, December.

[167] Solorio-Fernández, S., Carrasco-Ochoa, J. A., & Martínez-Trinidad, J. F. "A review of unsupervised feature selection methods", Artificial Intelligence Review, 53(2), 907-948, 2020.

[168] Grira, N., Crucianu, M., & Boujemaa, N. "Unsupervised and semi-supervised clustering: a brief survey", A review of machine learning techniques for processing multimedia content, 1(2004), 9-16, 2004.

[169] Ang, J. C., Mirzal, A., Haron, H., & Hamed, H. N. A. "Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection", IEEE/ACM transactions on computational biology and bioinformatics, 13(5), 971-989, 2015.

[170] Toldo, M., Maracani, A., Michieli, U., & Zanuttigh, P. "Unsupervised domain adaptation in semantic segmentation: a review", Technologies, 8(2), 35, 2020.

[171] Swenson, E. R., Bastian, N. D., & Nembhard, H. B. "Healthcare market segmentation and data mining: A systematic review", Health marketing quarterly, 35(3), 186-208, 2018.

[172] Das, S., & Nayak, J. "Customer segmentation via data mining techniques: state-of-the-art review", Computational Intelligence in Data Mining: Proceedings of ICCIDM 2021, 489-507, 2022.

[173] Ganguli, R., Mehta, A., & Sen, S. "A survey on machine learning methodologies in social network analysis", 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO) (pp. 484-489). IEEE, 2020, June.

[174] Alghamdi, H. M., & Selamat, A. "Arabic Web page clustering: A review", Journal of King Saud University-Computer and Information Sciences, 31(1), 1-14, 2019.

[175] Ray, P., Reddy, S. S., & Banerjee, T. "Various dimension reduction techniques for high dimensional data analysis: a review", Artificial Intelligence Review, 54(5), 3473-3515, 2021.

[176] Saini, O., & Sharma, S. "A review on dimension reduction techniques in data mining", Comput. Eng. Intell. Syst, 9(1), 7-14, 2018.

[177] Nassif, A. B., Talib, M. A., Nasir, Q., & Dakalbab, F. M. "Machine learning for anomaly detection: A systematic review", Ieee Access, 9, 78658-78700, 2021.

[178] Ravinder, M., & Kulkarni, V. "A review on cyber security and anomaly detection perspectives of smart grid", 2023 5th international conference on smart systems and inventive technology (ICSSIT) (pp. 692-697). IEEE, 2023, January.

[179] Hilal, W., Gadsden, S. A., & Yawney, J. "Financial fraud: a review of anomaly detection techniques and recent advances", Expert systems With applications, 193, 116429, 2022.

[180] Jain, A. K. "Data clustering: 50 years beyond K-means", Pattern recognition letters, 31(8), 651-666, 2010.

[181] Romanuke, V. "Random centroid initialization for improving centroid-based clustering", Decision Making: Applications in Management and Engineering, 6(2), 734-746, 2023.

[182] Saputra, D. M., Saputra, D., & Oswari, L. D. "Effect of distance metrics in determining k-value in k-means clustering using elbow and silhouette method", Sriwijaya international conference on information technology and its applications (SICONIAN 2019) (pp. 341-346). Atlantis Press, 2020, May.

[183] Arthur, D., & Vassilvitskii, S. "k-means++: The advantages of careful seeding", Soda (Vol. 7, pp. 1027-1035), 2007, January.

[184] Suwanda, R., Syahputra, Z., & Zamzami, E. M. "Analysis of euclidean distance and manhattan distance in the K-means algorithm for variations number of

centroid K", Journal of Physics: Conference Series (Vol. 1566, No. 1, p. 012058). IOP Publishing, 2020, June.

[185] Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Tibshirani, R., & Friedman, J. "Unsupervised learning", The elements of statistical learning: Data mining, inference, and prediction, 485-585, 2009.

[186] Kurita, T. "Principal component analysis (PCA)", Computer Vision: A Reference Guide, 1-4, 2019.

[187] Jolliffe, I. T., & Cadima, J. "Principal component analysis: a review and recent developments", Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202, 2016.

[188] Mishra, S. P., Sarkar, U., Taraphder, S., Datta, S., Swain, D., Saikhom, R., ... & Laishram, M. "Multivariate statistical data analysis-principal component analysis (PCA)", International Journal of Livestock Research, 7(5), 60-78, 2017.

[189] Abdi, H., & Williams, L. J. "Principal component analysis", Wiley interdisciplinary reviews: computational statistics, 2(4), 433-459, 2010.

[190] Yuan, X. "An improved Apriori algorithm for mining association rules", AIP conference proceedings (Vol. 1820, No. 1). AIP Publishing, 2017, March.

[191] Agrawal, R., & Srikant, R. "Fast algorithms for mining association rules", Proc. 20th int. conf. very large data bases, VLDB (Vol. 1215, pp. 487-499), 1994, September.

[192] Al-Maolegi, M., & Arkok, B. "An improved Apriori algorithm for association rules", arXiv preprint arXiv:1403.3948, 2014.

[193] Rana, S., Ali, W., & Khan, G. "Effect of Association Rule in Data Mining", Resurging India—Myths & Realities, 385, 2015.

[194] Hady, M. F. A., & Schwenker, F. "Semi-supervised learning", Handbook on Neural Information Processing, 215-239, 2013.

[195] Van Engelen, J. E., & Hoos, H. H. "A survey on semi-supervised learning", Machine learning, 109(2), 373-440, 2020.

[196] Reddy, Y. C. A. P., Viswanath, P., & Reddy, B. E. "Semi-supervised learning: A brief review", Int. J. Eng. Technol, 7(1.8), 81-89, 2018.

[197] Amini, M. R., Feofanov, V., Pauletto, L., Hadjadj, L., Devijver, E., & Maximov, Y. "Self-training: A survey", arXiv preprint arXiv:2202.12040, 2022.

[198] Haffari, G. R., & Sarkar, A. "Analysis of semi-supervised learning with the yarowsky algorithm", arXiv preprint arXiv:1206.5240, 2012.

[199] Kingma, D. P., Mohamed, S., Jimenez Rezende, D., & Welling, M. "Semi-supervised learning with deep generative models", Advances in neural information processing systems, 27, 2014.

[200] Ning, X., Wang, X., Xu, S., Cai, W., Zhang, L., Yu, L., & Li, W. "A review of research on co-training", Concurrency and computation: practice and experience, 35(18), e6276, 2023.

[201] Wang, W., & Zhou, Z. H. "A New Analysis of Co-Training", ICML (Vol. 2, p. 3), 2010, June.

[202] Kiritchenko, S., & Matwin, S. "Email classification with co-training", Proceedings of the 2001 conference of the Centre for Advanced Studies on Collaborative research (p. 8), 2001, November.

[203] Ding, S., Zhu, Z., & Zhang, X. "An overview on semi-supervised support vector machine", Neural Computing and Applications, 28, 969-978, 2017.

[204] Li, Y. F., & Zhou, Z. H. "Improving semi-supervised support vector machines through unlabeled instances selection", Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 25, No. 1, pp. 386-391), 2011, August.

[205] Wiering, M. A., & Van Otterlo, M. "Reinforcement learning", Adaptation, learning, and optimization, 12(3), 729-752, 2012.

[206] Kaelbling, L. P., Littman, M. L., & Moore, A. W. "Reinforcement learning: A survey", Journal of artificial intelligence research, 4, 237-285, 1996.

[207] Li, Y. "Reinforcement learning applications", arXiv preprint arXiv:1908.06973, 2019.

[208] Berrar, D. "Cross-validation", 2019.

[209] Diamantidis, N. A., Karlis, D., & Giakoumakis, E. A. "Unsupervised stratification of cross-validation for accuracy estimation", Artificial Intelligence, 116(1-2), 1-16, 2000.

[210] Yadav, S., & Shukla, S. "Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification", 2016 IEEE 6th International conference on advanced computing (IACC) (pp. 78-83). IEEE, 2016, February.

[211] Wong, T. T. "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation", Pattern recognition, 48(9), 2839-2846, 2015.

[212] Burkov, A. "The hundred-page machine learning book", 2019.

[213] Abdualgalil, B., & Abraham, S. "Applications of machine learning algorithms and performance comparison: a review", 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) (pp. 1-6). IEEE, 2020, February.

[214] Mehta, P., Bukov, M., Wang, C. H., Day, A. G., Richardson, C., Fisher, C. K., & Schwab, D. J. "A high-bias, low-variance introduction to machine learning for physicists", Physics reports, 810, 1-124, 2019.

[215] Seliya, N., Khoshgoftaar, T. M., & Van Hulse, J. "A study on the relationships of classifier performance metrics", 2009 21st IEEE international conference on tools with artificial intelligence (pp. 59-66). IEEE, 2009, November.

[216] Hossin, M., & Sulaiman, M. N. "A review on evaluation metrics for data classification evaluations", International journal of data mining & knowledge management process, 5(2), 1-14, 2015.

[217] Ferri, C., Hernández-Orallo, J., & Modroiu, R. "An experimental comparison of performance measures for classification", Pattern recognition letters, 30(1), 27-38, 2009.

[218] Gonçalves, L., Subtil, A., Oliveira, M. R., & de Zea Bermudez, P. "ROC curve estimation: An overview", REVSTAT-Statistical journal, 12(1), 1-20, 2014.

[219] Boyd, K., Eng, K. H., & Page, C. D. "Area under the precision-recall curve: point estimates and confidence intervals", Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13 (pp. 451-466). Springer Berlin Heidelberg, 2013.

[220] Vovk, V. "The fundamental nature of the log loss function", Fields of logic and computation II: Essays dedicated To Yuri Gurevich on the Occasion of His 75th Birthday, 307-318, 2015.

[221] Adeodato, P. J., & Melo, S. B. "On the equivalence between Kolmogorov-Smirnov and ROC curve metrics for binary classification", arXiv preprint arXiv:1606.00496, 2016.

[222] Turkish Airlines. "Traffic Results", Retrieved from https://investor.turkishairlines.com/tr/mali-ve-operasyonel-veriler/trafik, n.d.

[223] Weather Underground. "Local weather forecast, news and conditions", Retrieved from https://www.wunderground.com/, n.d.

[224] Ul Haq, I., Gondal, I., Vamplew, P., Brown, S., "Categorical features transformation with compact one-hot encoder for fraud detection in distributed environment", Data Mining: 16th Australasian Conference, AusDM 2018, Bahrurst, NSW, Australia, November 28–30, 2018, Revised Selected Papers, Vol. 16, pp. 69-80, Springer Singapore, 2019.

[225] Moubayed, A., Injadat, M., Nassif, A. B., Lutfiyya, H., Shami, A., "E-learning: Challenges and research opportunities using machine learning & data analytics", IEEE Access, Vol. 6, pp. 39117-39138, 2018.

[226] Manjrekar, O. N., Dudukovic, M. P., "Identification of flow regime in a bubble column reactor with a combination of optical probe data and machine learning technique", Chemical Engineering Science: X, Vol. 2, p. 100023, 2019.

[227] Bre, F., Gimenez, J. M., & Fachinotti, V. D. "Prediction of wind pressure coefficients on building surfaces using artificial neural networks", Energy and Buildings, 158, 1429-1441, 2018.

[228] Reader, C. "Decision Tree Regression Explained with Implementation in Python", Medium. Retrieved from https://medium.com, 2022, January 4.

[229] Hou, S. "Research on the Application of Data Mining Technology in the Analysis of College Students' Sports Psychology", Mobile Information Systems, 2021(1), 6529174, 2021.

[230] Rodríguez-Hernández, M. M., Pruneda, R. E., & Rodríguez-Díaz, J. M. "Statistical analysis of the evolutive effects of language development in the resolution of mathematical problems in primary school education", Mathematics, 9(10), 1081, 2021.