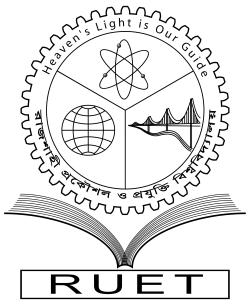
Heaven's Light is Our Guide



DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Rajshahi University of Engineering & Technology, Bangladesh

Advancing Machine Learning-Based Predictive Models for Airline Passenger Satisfaction through Investigating Hyperparameter Optimization and Feature Reduction Techniques

Author

Farjana Islam

Roll No. 1703097

Department of Computer Science & Engineering

Rajshahi University of Engineering & Technology

Supervised by

Md. Farukuzzaman Faruk

Assistant Professor

Department of Computer Science & Engineering

Rajshahi University of Engineering & Technology

ACKNOWLEDGEMENT

First of all, I would like to thank almighty Allah, for his grace and blessings as well as for providing me with the diligence and enthusiasm along the way to accomplishing my thesis work.

I also want to express my sincere gratitude, admiration and heartfelt appreciation to my supervisor **Md. Farukuzzaman Faruk**, Assistant Professor, Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology, Rajshahi. Throughout the year, he has not only provided me with the technical instructions and documentation to complete the work, but he has also continuously encouraged me, offered me advise, assisted me, and cooperated sympathetically whenever he deemed necessary. His constant support was the most successful tool that helped me to achieve my result. Whenever I was stuck in any complex problems or situation he was there for me at any time of the day. Without his sincere care, this work not has been materialized in the final form that it is now at the present.

I am also grateful to respected **Prof. Dr. Md. Al Mehedi Hasan**, Head of the Department of Computer Science & Engineering and all the respective teachers of Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology, Rajshahi for their valuable suggestions and inspirations from time to time.

Finally, I would like to convey my thanks to my parents, friends, and well-wishers for their true motivations and many helpful aids throughout this work.

April 16, 2024 Farjana Islam

RUET, Rajshahi

Heaven's Light is Our Guide

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Rajshahi University of Engineering & Technology, Bangladesh

CERTIFICATE

This is to certify that this thesis report entitled ***"Advancing Machine Learning-Based Predictive Models for Airline Passenger Satisfaction through Investigating Hyperparameter Optimization and Feature Reduction Techniques "*** *submitted by* ***Farjana Islam, Roll:1703097*** *in partial fulfillment of the requirement for the award of the degree of Bachelor of Science in Department of Computer Science & Engineering of Rajshahi University of Engineering & Technology, Bangladesh is a record of the candidate own work carried out by him under my supervision. This thesis has not been submitted for the award of any other degree.*

Supervisor

_____

Md. Farukuzzaman Faruk
Assistant Professor
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi-6204

External Examiner

_____

Abu Sayeed
Assistant Professor
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi-6204

ABSTRACT

In the aviation industry, ensuring passenger satisfaction is of utmost importance as it holds a pivotal position in improving airline services. Machine learning models have become essential in various industries, including aviation, allowing for comprehensive analysis and informed decision-making. The central goal of this study was to determine the most accurate model from six alternatives, primarily focusing on pinpointing the features that hold the greatest appeal for customers. The proposed machine learning model employed three distinct feature reduction techniques and integrated a hyperparameter tuning method known as GridSearch to achieve optimal results. Six different models were tested to determine the one that could accurately predict customer preferences based on various features. To identify the best model, various combinations of settings were explored using GridSearch, which systematically evaluated hyperparameter configurations and selected the one with the highest performance score. After exploring various feature reduction methods, the study ultimately concluded that the Random Forest model emerged as the clear frontrunner. The model achieved an outstanding accuracy score of 97%, accompanied by remarkable values of precision, recall, and f1-score, all of which also reached 97%. These exceptional performance metrics surpassed those of most recent similar studies. The study's outcomes offer airlines with valuable insights to prioritize features for enhancing passenger satisfaction and fostering loyalty.

CONTENTS

ACKNOWLEDGEMENT

CERTIFICATE

ABSTRACT

CHAPTER 1

Introduction 1

1.1 Introduction 1

1.2 Pandemic Impact on Airline 2

1.3 Customer Satisfaction 3

1.4 Overview 3

1.5 Motivation 4

1.6 Objective of the Thesis 5

1.7 Challenges of Machine Learning Prediction Without Hyperparameter Tuning 6

1.8 Thesis Organization 6

1.9 Conclusion 7

CHAPTER 2

Machine Learning 8

2.1 Introduction 8

2.2 Business Potential with Machine Learning 8

2.3 Machine Learning for Customer Satisfaction Prediction 11

2.4 Significance of Feature Reduction in Machine Learning 13

2.5 Conclusion 14

CHAPTER 3

Literature Review 15

3.1 Introduction 15

3.2 Related Works 15
3.3 Conclusion 22

CHAPTER 4
Proposed Methodology & Implementation 23
4.1 Introduction 23
4.2 Methodology 23
4.3 Dataset Preprocessing 25
4.4 Feature Reduction 26
4.5 Machine Learning Models-Based Framework 28
4.6 Training Set Testing Set 33
4.7 Hyperparameter Tuning with GridSearch 33
4.8 Conclusion 34

CHAPTER 5
Result & Performance Analysis 35
5.1 Introduction 35
5.2 Evaluation Metrics 35
5.3 Model Performance 39
5.4 Conclusion 51

CHAPTER 6
Conclusion & Future Works 52
6.1 Introduction 52
6.2 Summary 52
6.3 Conclusion 52
6.4 Limitation 53
6.5 Future Works 54

REFERENCES 55

LIST OF TABLES

4.1	Feature Reduction Techniques with the Number of Features	26
5.1	Best parameters of machine learning models using GridSearch	40
5.2	Model Performance - Feature Collection 1	41
5.3	Model Performance - Feature Collection 2	44
5.4	Model Performance - Feature Collection 3	47
5.5	A Comparative analysis of different works with the proposed Machine learning model. ACC: ACCURACY; PREC: PRECISION; REC: RECALL	51

LIST OF FIGURES

1.1 The flights monitored by Flightradar24 in 2020 showed a difference compared to the recorded in 2019. [1] 2

4.1 Proposed Machine learning Based Architecture 24
4.2 Features Visualization from Airline Passenger Satisfaction Dataset 25
4.3 Dataset after label encoding 26
4.4 Mutual Information values of all features with a bar plot 27
4.5 Linear Regression vs Logistic Regression [2] 29
4.6 SVM with Potential Hyperplane [3] 29
4.7 Gaussian Naive Bayes Classifier[4] 32
4.8 Find k-nearest neighbors[5] 32

5.1 Confusion Matrix of Random Forest from Feature Reduction 2 for Proposed Architecture 37
5.2 Confusion Matrix of Random Forest from Feature Reduction 1 for Proposed Architecture 42
5.3 Confusion Matrix of Decision Tree from Feature Reduction 1 for Proposed Architecture 42
5.4 Confusion Matrix of Support Vector Machine from Feature Reduction 1 for Proposed Architecture 42
5.5 Confusion Matrix of Logistic Regression from Feature Reduction 1 for Proposed Architecture 43
5.6 Confusion Matrix of K Nearest Neighbor from Feature Reduction 1 for Proposed Architecture 43
5.7 Confusion Matrix of Gaussian Naïve Bayes from Feature Reduction 1 for Proposed Architecture 43

5.8 Confusion Matrix of Random Forest from Feature Reduction 2 for Proposed Architecture 45
5.9 Confusion Matrix of Decision Tree from Feature Reduction 2 for Proposed Architecture 45
5.10 Confusion Matrix of Support Vector Machine from Feature Reduction 2 for Proposed Architecture 45
5.11 Confusion Matrix of Logistic Regression from Feature Reduction 2 for Proposed Architecture 46
5.12 Confusion Matrix of K Nearest Neighbor from Feature Reduction 2 for Proposed Architecture 46
5.13 Confusion Matrix of Gaussian Naïve Bayes from Feature Reduction 2 for Proposed Architecture 46
5.14 Confusion Matrix of Random Forest from Feature Reduction 3 for Proposed Architecture 48
5.15 Confusion Matrix of Decision Tree from Feature Reduction 3 for Proposed Architecture 48
5.16 Confusion Matrix of Support Vector Machine from Feature Reduction 3 for Proposed Architecture 48
5.17 Confusion Matrix of Logistic Regression from Feature Reduction 3 for Proposed Architecture 49
5.18 Confusion Matrix of K Nearest Neighbor from Feature Reduction 3 for Proposed Architecture 49
5.19 Confusion Matrix of Gaussian Naïve Bayes from Feature Reduction 3 for Proposed Architecture 49

Chapter 1

Introduction

1.1 Introduction

Tourism is a dominant mode of human movement in the contemporary world[6], with air transport playing a crucial role in facilitating this global phenomenon. The interconnection between air travel and tourism is undeniable[7], as the growth of the international tourism industry heavily relies on the availability and efficiency of air transportation. As of 2020, statistics from ATAG reveal that a significant 58% of international tourists opted for air travel in 2018, a substantial increase compared to the 35% recorded in 1980 before the onset of the COVID-19 pandemic[8] With over half of international tourists choosing air as their preferred mode of travel[9], air travel allows tourists to reach their desired destinations conveniently and efficiently. It is important to acknowledge that the COVID-19 pandemic has undoubtedly impacted air travel, leading to temporary setbacks and disruptions. During the COVID-19 pandemic, various countries implemented movement restrictions, leading to a significant decrease in the number of passengers traveling. This reduction in travel resulted in the cancellation of numerous flights and other transportation services worldwide. Consequently, these cancellations caused service failures, as the expected services could not be fulfilled, leading to customer dissatisfaction and disappointment[10]. To mitigate customer disappointment, developing a machine learning model for aviation customer satisfaction involves data-driven analysis of influential factors, predicting satisfaction levels, and gaining insights to improve service quality.

1.2 Pandemic Impact on Airline

The COVID-19 pandemic had a profound impact on the aviation industry as a whole. Governments around the world implemented strict travel restrictions to control the spread of the virus, leading to an unprecedented decrease in air travel demand. Fear of infection, quarantine measures, and uncertainties surrounding travel regulations further discouraged people from undertaking non-essential trips. As a result, airlines experienced an unparalleled decrease in passenger numbers, leading to financial losses and challenges in sustaining their operations.

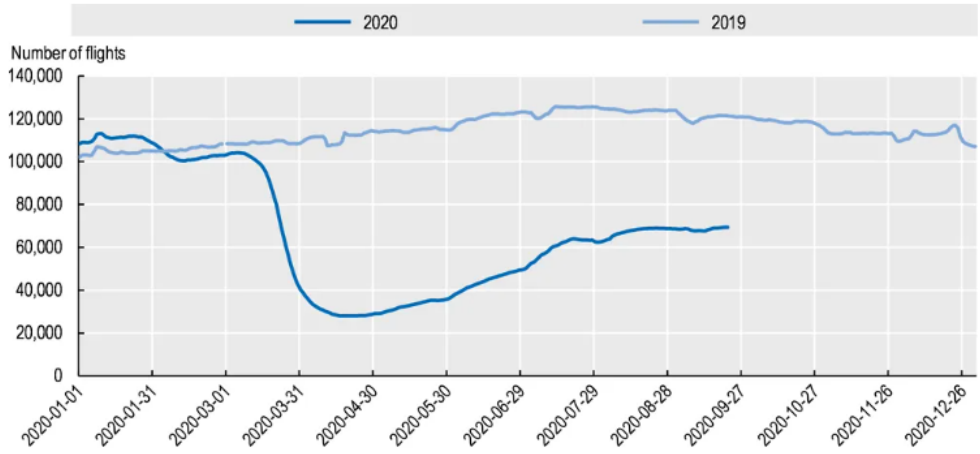


Figure 1.1: The flights monitored by Flightradar24 in 2020 showed a difference compared to the recorded in 2019. [1]

Figure.1.1 demonstrates a decline in the growth of commercial air traffic in 2020, which can be attributed to the impact of the pandemic. The drastic decline in passenger air transport was also influenced by the widespread economic crisis triggered by the pandemic. Businesses faced disruptions, unemployment rates surged, and consumer spending plummeted. Many people postponed or canceled their travel plans, both for leisure and business purposes, as financial concerns took precedence. Moreover, international travel restrictions and border closures severely impacted the tourism industry, as countries sought to protect their citizens and contain the virus. As reported by IATA, the revenue passenger kilometers, a measure of passenger air transport, experienced a staggering 90% year-on-year decline in April 2020[11], and despite some improvement, it was still down by 75% in August of the same year[12]. Additionally, the collapse in economic activities and global trade also impacted the freight sector, which saw a nearly 30% decrease in April compared to the previous year[13] and remained approximately 12% lower in August[14].

1.3 Customer Satisfaction

Customer satisfaction measures the level of contentment and fulfillment customers experience with a business's products, services, or overall interactions[15]. It is a critical component of a business's success and is often used as a key performance indicator to assess how well a company meets its customers' needs and expectations. The concept of customer satisfaction revolves around the idea that a satisfied customer is more likely to be loyal, make repeat purchases, and speak positively about the business to others, thus driving word-of-mouth referrals. Satisfied customers are also less likely to switch to competitors, contributing to higher customer retention rates. Aviation companies faced significant challenges during the COVID-19 pandemic, and some of the ways they might have failed in their services include Flight cancellations and disruptions, Poor communication, Refund, and compensation delays, Health and safety measures, Limited route options, Workforce issues, Inflexible policies. Many airlines had to cancel or reschedule flights due to travel restrictions, lockdowns, and declining demand. However, some airlines faced criticism for providing insufficient notice to passengers about cancellations or offering limited alternative options, leading to inconvenience and frustration for travelers[16]. Due to travel restrictions and reduced demand, airlines suspended or reduced routes to various destinations. This limited travelers' options and caused inconvenience for those who needed to travel during the pandemic. Conduct customer surveys and feedback sessions to understand their post-pandemic expectations and concerns. Use this data to make informed decisions and improvements based on customer preferences.

1.4 Overview

Due to overcoming the challenges brought by the COVID-19 pandemic, aviation companies are taking proactive steps to increase their business and regain customer satisfaction. One of the key approaches is enhancing communication with customers through multiple channels, providing them with timely updates, and addressing their concerns effectively. Additionally, offering flexible booking options and relaxed refund policies to accommodate changing travel plans has proven beneficial in building trust and loyalty among customers. To achieve higher service quality and informed decision-making, aviation companies are leveraging machine learning models. The primary goal of recent study in this area was to identify the most accurate machine learning model among six options, with a particular focus on determining the features that contribute

most to customer appeal. The proposed model integrated three different feature reduction techniques and GridSearch, a hyperparameter tuning method, to optimize performance. Throughout the study, six machine learning models were rigorously tested to predict customer preferences based on various features. Utilizing GridSearch allowed for systematic exploration of hyperparameter configurations, identifying the settings that yielded the highest performance scores. Among the models, the Random Forest model emerged as the top performer, achieving an exceptional accuracy score of 97%, with impressive precision, recall, and f1-score values, all at 97% as well. These remarkable results surpassed those reported in recent similar studies. The insights gained from this study provide valuable guidance for airlines in understanding customer preferences and prioritizing features to enhance passenger satisfaction and foster customer loyalty. By making data-driven decisions, airlines can tailor their services and offerings, creating a more satisfying and personalized travel experience for passengers. The integration of machine learning in the aviation industry presents a significant opportunity for improving service quality and meeting customer expectations effectively.

1.5 Motivation

The motivation behind conducting a study on fine-tuning machine learning models for airline passenger satisfaction prediction lies in addressing the pressing need of the aviation industry to enhance customer experience. As the world emerges from the challenges posed by the COVID-19 pandemic, airlines are keen to regain customer satisfaction and loyalty. Understanding and predicting passenger preferences and satisfaction levels are critical for airlines to tailor their services effectively, offer personalized experiences, and meet customer expectations. Through fine-tuning machine learning models, the goal is to identify the most accurate predictive model that can effectively analyze large amounts of data and provide valuable insights. Such insights can help airlines prioritize features, services, and amenities that appeal most to passengers, thereby improving their overall satisfaction. Furthermore, the study explores hyperparameter optimization and feature reduction techniques to ensure that the machine learning models are efficient and precise in their predictions. Integrating machine learning within the aviation industry has the capability to reshape airlines' operations, granting them the ability to formulate decisions guided by data, efficiently manage resources, and deliver exceptional services. By integrating state-of-the-art technologies and methodologies, this research seeks to offer novel

and impactful solutions to real-world challenges faced by the aviation industry.

1.6 Objective of the Thesis

The multifaceted nature of the travel process, which encompasses a multitude of factors such as airport operations, customs procedures, and immigration protocols, coupled with the wide-ranging and unique preferences of travelers, has presented a formidable hurdle in the pursuit of consistent customer satisfaction. The main Objectives of this work are as follows:

- Prediction of Passenger Satisfaction: The primary goal is to develop machine learning models that can accurately predict passenger satisfaction levels based on various features and factors. By fine-tuning these models, the study aims to achieve high accuracy in predicting whether passengers are satisfied with their travel experience.
- Identification of Key Factors: By analyzing different features and preferences, the study aims to pinpoint the key aspects that significantly influence a passenger's overall satisfaction with the airline's services
- Hyperparameter Optimization: Another objective is to explore various hyperparameter configurations using the GridSearch technique. By systematically evaluating different settings, the study aims to identify the optimal hyperparameters that maximize the model's predictive performance.
- Feature Reduction Techniques: The study aims to employ three distinct feature reduction techniques to streamline the machine learning process and enhance model performance.
- Model Comparison: The research aims to test and compare six different machine learning models to determine which one provides the most accurate predictions of passenger satisfaction. The study will assess the models based on various performance metrics to identify the best-performing model.
- Advancing Airline Services: By fine-tuning machine learning models for passenger satisfaction prediction, the study aims to contribute to the advancement of airline services. The research strives to offer airlines valuable information to prioritize and tailor their offerings, leading to a more personalized and satisfactory travel experience for passengers.

1.7 Challenges of Machine Learning Prediction Without Hyperparameter Tuning

The difficulties of using a machine learning model for prediction without hyperparameter tuning can include :

- Poor Model Performance.
- Inefficient Resource Usage.
- Lack of Model Robustness.
- Difficulty in Model Selection.
- Limited Generalization.
- Lack of Adaptability.

These difficulties were overcome by proposed machine learning model, which resulted in an efficient model. It is discussed in the following chapters, along with a description of how it was handled.

1.8 Thesis Organization

The report is organized into 6 chapters including this chapter: ***Introduction*** where all the related topics are discussed which are needed for understanding the research work. The outline of rest of the works are organized as follows:

Chapter 2

Topic - Machine Learning

This chapter presents the utilization of machine learning in business contexts, outlines the application of machine learning for forecasting customer satisfaction, and delves into the advantages of feature reduction techniques.

Chapter 3

Topic - Literature Review

This chapter covers some works related to predicting customer satisfaction with their contributions and limitations.

Chapter 4

Topic - Materials & Methodology

This chapter discusses the dataset and the proposed methodology. It also includes a detailed description of data pre-processing, proposed architecture, and model training with the help of GridSearch.

Chapter 5

Topic - Result & Performance Analysis

This chapter analyses the experimental result and performance of the proposed architecture along with the comparison with related works. The metrics which were used to evaluate our model is also described here.

Chapter 6

Topic - Conclusion

Through this chapter, the research work has been concluded. This article gives a summary of my research's findings. I have also tried to make an effort to highlight limitations and potential future work areas of my work.

1.9 Conclusion

In this chapter, we were provided with an overview of the upcoming study and a glimpse into the work that lies ahead. The discussion included insights into the inspiration, objectives, and research challenges that will be further elaborated in subsequent chapters.

Chapter 2

Machine Learning

2.1 Introduction

Machine learning, a subset of artificial intelligence, involves crafting algorithms and models that empower computers to learn from data, thus enabling them to predict outcomes and make decisions without needing direct, task-specific programming. The foundational concept driving machine learning is to grant computers the ability to recognize patterns in data, draw valuable conclusions, and enhance their efficacy as they accumulate experiential knowledge. This iterative learning process transforms machines into proficient pattern recognizers and decision-makers, with applications spanning diverse fields.

2.2 Business Potential with Machine Learning

Machine learning can significantly enhance business analysis by providing sophisticated tools to process and extract insights from large volumes of data. Here's how machine learning helps in analyzing business operations and decision-making-

- Machine learning enables businesses to gain deep insights into customer behavior and preferences by analyzing vast amounts of data. This information allows companies to deliver personalized experiences, tailor marketing campaigns, and recommend products or services that resonate with individual customers, fostering stronger relationships and increasing customer satisfaction.

- Machine learning models can predict future trends, market shifts, and customer demands based on historical data. By accurately forecasting these aspects, businesses can make proactive decisions about resource allocation, inventory management, and product development, ultimately improving operational efficiency.

- Machine learning algorithms excel at identifying unusual patterns and anomalies within large datasets. Businesses can leverage this capability to detect fraudulent activities in real-time, whether it's financial transactions, user behaviors, or security breaches, thus safeguarding their assets and maintaining trust with customers.

- By automating repetitive and manual tasks, machine learning streamlines business processes and enhances efficiency. This allows employees to focus on more strategic and creative tasks, leading to increased productivity and reduced operational costs.

- Machine learning empowers businesses to optimize their marketing efforts by analyzing customer interactions and campaign outcomes. Insights gained from these analyses help refine marketing strategies, target audiences more effectively, and allocate resources where they're most impactful.

- In finance and lending, machine learning evaluates credit risks by assessing borrowers' creditworthiness based on historical data and behavioral patterns. This assists financial institutions in making informed lending decisions and managing potential risks.

- Machine learning optimizes supply chain operations by analyzing data related to demand, inventory levels, and transportation routes. Businesses can ensure the right products are available at the right time and minimize delays in the distribution process.

- Machine learning plays a pivotal role in healthcare by analyzing medical data and images to assist in disease diagnosis and treatment recommendations. It also aids in drug discov-

ery by identifying potential drug candidates and predicting their effectiveness.

- In industries focused on energy consumption, such as utilities and manufacturing, machine learning optimizes energy usage by analyzing patterns and recommending strategies to reduce costs and environmental impact.

- Machine learning processes and analyzes data, offering businesses real-time insights into changing market conditions, consumer behaviors, and operational performance, facilitating timely decision-making.

- Machine learning assists in talent acquisition by analyzing resumes, identifying suitable candidates, and predicting employee attrition risks. It also helps in assessing employee performance and tailoring professional development strategies.

- In manufacturing and industrial sectors, machine learning identifies defects in products by analyzing sensor data and equipment conditions. This aids in maintaining product quality and minimizing downtime through timely maintenance.

- Machine learning processes vast amounts of data to provide insights into market trends, consumer sentiment, and competitor strategies, enabling businesses to make more informed strategic decisions.

- Across all business functions, from finance and operations to marketing and sales, machine learning provides data-driven insights that guide effective decision-making, leading to optimized processes and outcomes.

In essence, machine learning empowers businesses to harness the full potential of their data, providing deeper and more accurate insights that drive informed decision-making, operational efficiency, and strategic innovation.

2.3 Machine Learning for Customer Satisfaction Prediction

Machine learning can predict customer satisfaction by analyzing various data points and patterns associated with customer interactions, behaviors, and feedback.

- Gather relevant data from different sources, including customer interactions, purchase history, survey responses, social media engagement, and more. This data provides a comprehensive view of customer behavior and sentiment.

- Identify key features or attributes that could influence customer satisfaction. These could include factors like purchase frequency, browsing history, product ratings, time spent on the website, demographic information, and more.

- Clean and preprocess the data to handle missing values, outliers, and inconsistencies. This step ensures that the data is suitable for analysis and model training.

- Create labels that represent different levels of customer satisfaction. For example, you might label customers as ”Satisfied,” ”Neutral,” or ”Dissatisfied” based on their feedback or purchase history.

- Choose an appropriate model for classification or regression tasks, depending on whether you’re predicting discrete satisfaction levels or a continuous satisfaction score.

- Create new features or transform existing ones to enhance the predictive power of the model. This could involve combining variables, creating interaction terms, or applying dimensionality reduction techniques.

- Train the machine learning model using historical data that includes both the selected features and the corresponding customer satisfaction labels. The model learns the relationships between the features and customer satisfaction levels.

- Split the dataset into training and testing subsets to evaluate the model's performance. Techniques like cross-validation help assess how well the model generalizes to unseen data.

- Measure the model's accuracy, precision, recall, F1-score, or other relevant metrics to assess its ability to predict customer satisfaction correctly.

- Once the model is trained and validated, use it to predict customer satisfaction for new data points. The model analyzes the features of these new data points and assigns a predicted satisfaction level.

- Continuously update the model with new data to improve its accuracy over time. Customer behaviors and preferences can change, and the model should adapt accordingly.

- Depending on the algorithm used, it's important to interpret the model's predictions. This helps businesses understand the factors driving customer satisfaction and make informed decisions based on these insights.

- The model's predictions can guide businesses in taking specific actions to improve customer satisfaction. For instance, if the model predicts dissatisfaction among a certain customer segment, businesses can tailor their offerings or customer support to address the issues.

- Incorporate customer feedback and outcomes into the model's training process. If the model's predictions align with actual satisfaction levels, it validates the model's effectiveness. If not, the model might need further refinement.

- Continuously monitor the model's performance and recalibrate it as necessary to ensure accurate predictions as customer behavior evolves.

analyzing historical data and identifying patterns, machine learning models can provide businesses with valuable insights into the factors that influence customer satisfaction. This enables proactive efforts to enhance customer experiences and drive long-term faithfulness.

2.4 Significance of Feature Reduction in Machine Learning

Feature reduction is a critical process in machine learning that involves choosing the most relevant and informative features or attributes from a given dataset. The goal is to enhance model performance, reduce complexity, and improve interpretability by selecting a subset of features that contribute the most to the model's predictive accuracy.

- Opting for the most pertinent features enhances the model's capacity to apply its findings to fresh, unfamiliar data. Features that are unimportant or laden with noise can result in overfitting, wherein the model memorizes the training data but struggles to excel when confronted with novel information.

- Including unnecessary features in a model can increase computational complexity and training time. Feature selection helps streamline the model's processing and reduces resource requirements.

- Models with fewer features are easier to interpret and understand. This is especially important in business settings where stakeholders need to comprehend the factors influencing predictions and decisions.

- With an escalation in the count of features, the volume of requisite data to adequately train a model also rises. Employing feature selection assists in alleviating the complexities presented by high-dimensional data, where the proportion of samples to features is limited.

- Models built with a reduced set of features are less likely to be influenced by noise or irrelevant variations in the dataset, leading to improved extrapolation when faced with

novel data points.

- Feature selection can reveal the most significant variables driving certain outcomes, providing valuable insights for decision-making and strategy development.

2.5 Conclusion

Machine learning stands as a transformative force in the business landscape, offering multifaceted advantages that propel organizations toward enhanced performance, informed decision-making, and unprecedented innovation. By harnessing algorithms and models, businesses can decipher intricate patterns within vast datasets, predict future trends, and optimize operations, thereby reshaping traditional practices and fostering sustainable growth. Machine learning's role in predicting customer satisfaction exemplifies its potential. Through meticulous data analysis and predictive modeling, businesses gain insights into customer behaviors, preferences, and sentiments. This empowers them to tailor their offerings, refine strategies, and proactively address concerns, ultimately nurturing lasting customer relationships and driving competitive advantage. In the heart of these endeavors lies the pivotal concept of feature selection. By sifting through attributes to identify the most influential, businesses can streamline models, boost predictive accuracy, and unveil actionable insights. This spotlight on feature selection underscores its indispensable role in enhancing model performance, interpretability, and generalization, illustrating its value in steering organizations toward more effective data-driven strategies.

Chapter 3

Literature Review

3.1 Introduction

The literature review examines studies concerning airline customer satisfaction, where data mining and machine learning techniques are employed to forecast satisfaction levels. Additionally, the review delves into the repercussions of flight delays on passenger demand and airfares. Within this context, multiple models like Naïve Bayes, PSO, GA, Random Forest, and XGBoost are assessed for their predictive accuracy. The review underscores the significance of suitable predictor variables and engages in discussions regarding potential enhancements for the airline sector, encompassing the augmentation of customer experience and the optimization of operational procedures.

3.2 Related Works

In the study conducted by Religia et al. [17], data mining techniques were employed to forecast airline customer satisfaction, utilizing the Naive Bayes Algorithm as a classification model. This algorithm exhibited noteworthy classification accuracy, particularly when augmented with two optimization strategies: particle swarm optimization (PSO) and genetic algorithm (GA). PSO entailed identifying optimal parameter values within system constraints, and ensuring adherence to design prerequisites [18]. Conversely, GA found near-optimal solutions for intricate challenges that would otherwise demand extensive computational time [19]. These optimization approaches were harnessed for selecting pertinent features in the classification model. Post comprehensive experimentation, researchers determined that the optimal performance emerged

from the Naive Bayes algorithm, synergized with PSO optimization, for categorizing Airline Passenger Satisfaction data. Outcomes unveiled an accuracy of 86.13%, precision at 87.90%, recall reaching 87.29%, along with an AUC value of 0.923.

In a study carried out by Cobb et al. [20] an analysis was conducted using cumulative confirmed COVID-19 cases data encompassing the period from January 21, 2020, to March 31, 2020, encompassing all 3139 counties across the United States. Compound growth rates were computed to portray the pace of virus propagation during specific time intervals, given its suitability for machine learning analysis as a singular metric. Employing both statistical methodologies and a random forest machine learning model, the researchers examined disparities between counties with and without shelter-in-place (SIP) directives. The statistical evaluations unveiled that the presidential recommendation on March 16, imposing limitations on gatherings of <=10 people, yielded a 6.6% reduction in the compound growth rate of COVID-19 across all US counties. Additionally, counties that enforced SIP orders post-March 16 observed an additional decrease of 7.8% in contrast to counties that abstained from implementing SIP orders after that date. In terms of predicting the compound growth rate following a SIP order, a random forest machine learning model was developed, achieving a notable accuracy level of 92.3%. This random forest analysis highlighted factors like population, longitude, and population per square mile as the most pivotal elements influencing the efficacy of SIP measures.

Walker et al.[21] introduce a replicable implementation of an adaptive boosting (AdaBoost) model specifically designed to forecast the probability of inducing purchases for titles under Demand-driven acquisition (DDA). The study involved a comparative analysis of the predictive performance between this model and a conventional logistic regression approach, employing the same predictive variables. The results underscore that the AdaBoost model surpasses the regression-based alternative considerably in terms of its predictive prowess. Once trained using localized DDA data, the AdaBoost algorithm achieves precise predictions in 82% of instances. This emphasizes the superior efficacy of the AdaBoost model in foreseeing purchase triggers for DDA titles in contrast to the traditional logistic regression methodology.

Chen et al.[22] presented a sophisticated pattern classification approach aimed at discerning the behavior of the primary user within a network. They achieved this by combining a robust machine learning classifier (MLC) and decision stumps (DS) in an adaptive boosting (AdaBoost)

framework. In contrast to the typical AdaBoost method that amalgamates various sub-classifiers based on their weights, their innovative hybrid AdaBoost algorithms integrate MLC and DS to attain an elevated detection probability compared to conventional machine learning-based spectrum sensing techniques. To validate the efficacy of their strategy, they conducted simulations.

Wie et al.[23] developed a logical framework for evaluating customer satisfaction in China Southern Airlines, based on China's service industry customer satisfaction index model. They introduced ISO9000 quality management principles into the airline's service quality system, creating a new process-based service quality management model. Additionally, they analyzed service remediation in the aviation industry within the context of the airline's value chain and suggested its potential for generating benefits.

Tsafarakis et al.[24] demonstrated the use of a multicriteria satisfaction analysis technique to assess passenger satisfaction across various service dimensions and identify areas for improvement. They applied this method to Aegean Airlines as an illustration of its effectiveness in measuring and analyzing passenger satisfaction. The research provided valuable revelations regarding the criteria and subcriteria for satisfaction that hold the highest significance among passengers of a full-service airline. Additionally, it sheds light on intriguing patterns observed within various segmentation strategies.

Sajadi et al.[25] have contributed to the existing body of literature by illustrating how airlines can enhance passenger satisfaction and loyalty through prioritizing the enhancement of specific components within the pre-flight and in-flight service encounters that fall directly within their purview. The outcomes of their investigation unveiled that perceived quality of services before the flight and during the flight represent distinct elements within the overall airline service quality. Importantly, each of these aspects independently and positively impacts passenger satisfaction. Furthermore, their study underscored the notable and positive influence of perceived pre-flight service quality on passenger loyalty. These findings underscore the importance of pre-flight service quality and emphasize the essential role that customer perceptions of service quality before the actual experience—encompassing communication, procedures, and interactions—play in enhancing customer satisfaction and cultivating enduring customer loyalty. The study illuminates the pivotal role of airlines' control over both pre-flight and in-flight service elements in shaping passenger satisfaction and loyalty. By directing efforts towards

enhancing these areas, airlines can effectively shape customer perceptions and ultimately elevate the overall air travel experience. This research provides airlines with valuable insights to formulate strategies prioritizing customer-centric service enhancements, ultimately driving elevated customer satisfaction and fostering long-term loyalty.

In a groundbreaking study, Chen et al. [26] delved into the realm of feature selection, a widely employed technique aimed at streamlining predictors by identifying the most informative attributes from the original feature set. They centered their focus on Recursive Feature Elimination (RFE), a prominent method used for reducing data dimensionality and enhancing efficiency. RFE generates a prioritized roster of features and associated subsets, each with accuracy values. Traditionally, the subset with the highest accuracy or a predetermined number of features is chosen as the final subset. However, this approach can lead to the selection of an excessive number of features or might be subjective when the preset count is uncertain. To tackle this issue, the researchers explored alternative decision-making approaches to automatically ascertain the optimal feature subset once candidate subsets were derived from RFE. They introduced an algorithm termed Random Forest (RF) Recursive Feature Elimination (RF-RFE), coupled with a voting strategy to serve this purpose. These variations were meticulously scrutinized and compared using two distinct molecular biology datasets—one geared towards toxicogenomic study and the other centered on protein sequence analysis. This study introduces an automated and objective avenue for selecting the optimal feature subset through RF-RFE, offering insightful methods to enhance the efficiency and efficacy of feature selection across diverse applications.

Ontivero-Ortega et al.[27] introduced massive-GNB, a faster and efficient implementation of the Gaussian Naive Bayes classifier for the searchlight technique in multivariate pattern analysis (MVPA). This new classifier allows simultaneous classification in all searchlights, outperforming previous implementations of GNB and other complex classifiers like SVM. The research juxtaposed the effectiveness of massive-GNB and SVM in identifying the lateral occipital complex (LOC) within an fMRI localizer experiment involving 26 subjects. While SVM showed slightly higher accuracy in individual searchlights and better selectivity for LOC, both classifiers performed equally well in cluster-level analysis with multiple comparison correction. The findings suggest that massive-GNB provides comparable accuracy to sophisticated classifiers but with significantly faster processing, making it a valuable tool for broader use in neuroscience research.

Gregorutti et al.[28] studied variable selection using random forests in the presence of correlated predictors, particularly in high-dimensional regression or classification scenarios. They proposed the recursive feature elimination (RFE) algorithm, which efficiently selects relevant variables by recursively eliminating them based on permutation importance measures. Through simulation experiments, the effectiveness of RFE in choosing a limited set of variables with minimal prediction error was showcased. The algorithm's validation was extended to actual Landsat Satellite data, underscoring its usefulness in addressing the complexities of variable selection arising from correlated predictors.

Naik et al.[29] undertook a thorough exploration of classification algorithms by employing various freely accessible tools for data mining and knowledge discovery. These encompassed WEKA, RapidMiner, Tanagra, Orange, and Knime. Developed at the University of Waikato, WEKA is a widely-used Java-based toolkit catering to machine learning and data mining, offering an array of algorithms for classification, regression, clustering, association rules, data visualization, and preprocessing. Tanagra finds its footing in supporting a multitude of data mining tasks, striking a balance between statistical methodologies, multivariate analysis techniques, and machine learning approaches. Orange, on the other hand, is a component-based visual programming software crafted for data mining, data analysis, and machine learning, while Knime stands as an open-source graphical workbench that enjoys extensive use in data mining, reporting, and visualization. The research centered around assessing the accuracy of classification algorithms such as Decision Trees, Decision Stumps, K-Nearest Neighbors, and Naïve Bayes across all five tools, leveraging the Indian Liver Patient DataSet. This particular dataset was employed for classifying individuals with and without liver disorders.

Kumar et al.[30] conducted a study focusing on improving customer experience in the airline industry by analyzing tweets posted on Twitter. They adopted a machine learning technique that encompassed the utilization of word embedding through the Glove dictionary and n-gram methodologies to derive features from the tweets. SVM and various ANN architectures were employed to create a classification model that categorized tweets as positive or negative. Additionally, CNN models were developed and compared with SVM and ANN models, with CNN showing superior performance. Association rule mining was then used to identify interesting relationships between different tweet categories and sentiment categories, providing valuable insights to enhance airline industries' customer experience.

Moreira et al.[31] delved into the prevalent issue of flight delays, which holds substantial impli-

cations for airlines, airports, and passengers alike. Drawing on data from the Brazilian National Civil Aviation Agency (ANAC), it was revealed that around 22% of domestic flights in Brazil encountered delays exceeding 15 minutes between 2009 and 2015. The anticipation of these delays is pivotal in ameliorating their repercussions and refining decision-making within the aviation system. The study places a spotlight on the uneven distribution of delay classifications (presence and absence) and addresses this through the evaluation of diverse preprocessing techniques to formulate machine-learning models for classifying flight delays. The dataset utilized amalgamated national flight operations data with meteorological conditions at airports. The outcomes demonstrated that models employing balancing methodologies exhibited notably enhanced performance in predicting instances of delay, attaining an accuracy of approximately 60%.

Kuhn et al.[32] harnessed the capabilities of machine learning algorithms, including decision trees, logistic regression, and neural networks classifiers, to forecast the likelihood of flight arrival delays. The classifiers underwent training and testing using a dataset comprising 100,000 samples, adhering to a recommended split of 70-30. For decision tree and neural network classifiers, cross-validation was executed, with utilization of the scikit and keras APIs when deemed necessary. The research spotlighted that straightforward classifiers, such as decision trees and logistic regression, can adeptly predict whether a flight's arrival will be subject to delay. In their forthcoming endeavors, the researchers intend to augment their models by incorporating more extensive training data or exploring more intricate neural networks. Moreover, they have aspirations to extend their investigations to forecasting taxi delays, taking into account aspects like airport runway and taxiway configurations, a domain where current research remains limited.

Gopalakrishnan et al.[33] undertook a comparative exploration of diverse strategies for forecasting delays within air traffic networks. These strategies encompassed the Markov Jump Linear System (MJLS), conventional machine learning techniques such as Classification and Regression Trees (CART), as well as architectures of Artificial Neural Networks (ANN). Their findings underscored the substantial influence of model selection and prediction type on performance outcomes. In terms of classification, ANN showcased remarkable performance, attaining an accuracy of almost 94% in predicting delays surpassing 60 minutes. Conversely, the MJLS model outshone ANN in predicting actual delay levels across different connections, with an average error of 4.7 minutes for a 2-hour forecast horizon. The study emphasized the significance of judiciously selecting predictor variables and engaged in discussions about the

balance between model simplicity and prediction precision. Particularly, the MJLS model, tailored to capture the aggregate dynamics of air traffic, exhibited superiority in predicting the forthcoming spatial distribution of delays.

Britto et al.[34] delved into the repercussions of flight delays on passenger demand and airfares. Their investigation involved evaluating delays in relation to planned block times and more optimal feasible flight times, coupled with assessing the associated welfare impacts through econometric assessments. The study unearthed that flight delays along a specific route translated to diminished passenger demand and escalated airfares, consequently inducing notable declines in both consumer and producer welfare. Notably, the estimated effects on producer welfare significantly surpassed those on consumer welfare, being threefold in magnitude. In light of these findings, the authors suggest that, from an economic efficiency standpoint, airlines should take the primary responsibility for implementing measures to alleviate flight delays.

Baswardono et al.[35] performed a comparison between the Random Forest and C4.5 machine learning models, evaluating them based on accuracy, precision, F1 score, and recall as assessment criteria. Random Forest demonstrated superior prediction accuracy, prompting them to recommend dataset improvement through a modified method for better predictions.

Homaid et al.[36] prepared a dataset by eliminating stopwords and applying TF-IDF for numerical conversion of review text. They compared five machine-learning models and found XGBoost to be the most accurate for predictions. The authors suggested sharing their model with airline operators to prioritize passengers' challenges automatically.

Suwanto et al.[37] found Random Forest to be effective and efficient in classifying passenger ratings and determining satisfaction with airline services. The tree construction process required fewer parameters. AlHabbal et al.[38] reviewed classification algorithms and optimized them for better results, using LDA for attribute reduction and implementing a Big Data pipeline for passenger feedback analysis.

Hatıpoğlu et al.[39]proposed estimations using modern techniques like XGBoost, LightGBM, and CatBoost, utilizing Gradient Boosting. Bayesian technique was employed to establish the best hyper-parameters for predicting airline flight delays, including alternatives for cargo flights requiring urgent transportation.

Samah et al.[40] designed a web-based dashboard for bilingual sentiment analysis of Twitter content, employing the Naïve Bayes algorithm. The evaluation of the dashboard's user-friendliness was conducted using the System Usability Scale, resulting in a score of 94.7%.

The acceptability assessment indicated a positive outcome, suggesting that the study provides a valuable tool that contributes to the comprehension of public sentiments regarding airline companies in Malaysia.

3.3 Conclusion

The majority of previous literature reviews have centered on customer satisfaction, with many studies utilizing various machine learning models and technologies to predict satisfaction levels. However, most of these models fall short of achieving an accuracy above 95%. In response to this limitation, I have developed and implemented my proposed methodology to address the issue.

Chapter 4

Proposed Methodology & Implementation

4.1 Introduction

Machine learning can significantly help predict customer satisfaction efficiently by analyzing and understanding patterns in customer data. Feature reduction techniques and grid search also help improve the performance of a machine learning model in predicting customer satisfaction. When dealing with many features, feature reduction techniques help simplify the model, reduce computational complexity, and improve generalization. Grid search helps optimize these hyperparameters for the best possible model performance. By combining feature reduction techniques with grid search, you can identify the most relevant features and fine-tune the model's hyperparameters, leading to a more efficient and accurate customer satisfaction prediction model. This, in turn, helps businesses make better decisions and improve customer experiences.

4.2 Methodology

In the proposed architecture, a comprehensive approach was adopted to predict customer satisfaction efficiently. Three distinct feature reduction methods were employed to streamline the model and enhance its performance. Each of these methods played a crucial role in identifying the most relevant features and mitigating computational complexity. To ensure optimal model performance, GridSearch was thoughtfully integrated with each feature reduction method. This strategic combination allowed for an exhaustive exploration of various hyperparameter configurations, enabling the identification of the most effective model for each feature reduction technique. The intricate workflow, beautifully depicted in Figure. 4.1, showcases the seamless

execution of these steps.

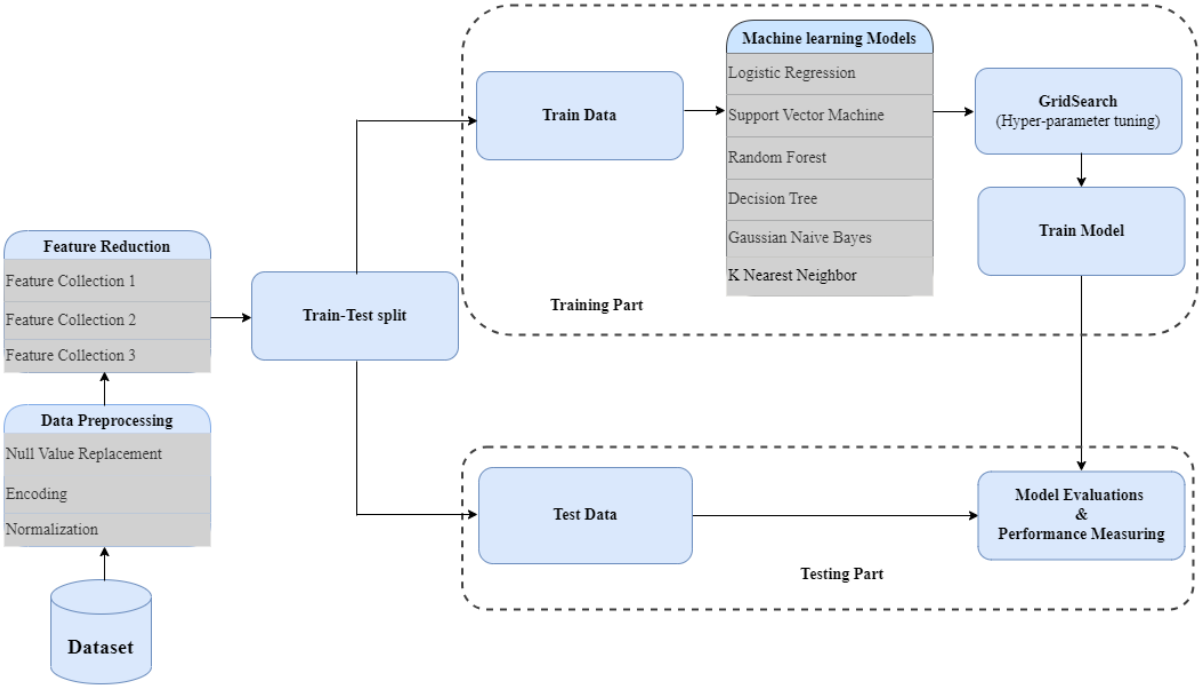


Figure 4.1: Proposed Machine learning Based Architecture

4.2.1 Dataset Descriptions

This study utilized the Airline Passenger Satisfaction dataset, which was obtained from a reliable source, Kaggle[41]. This dataset comprises survey data collected from passengers of an airline. This dataset consisted of 24 features and includes a total of 103,904 rows of data. The dataset also contained two target classes: "satisfied" and "neutral or dissatisfied". The textual data type features included Gender, Customer type, Type of travel, Class. Passenger satisfaction for various services such as Inflight wiFi service, Departure/Arrival time convenience, Online booking ease, Gate location, Food and drink, Online boarding, Seat comfort, Inflight entertainment, On-board service, Leg room service, Baggage handling, Checkin service, Inflight service, and Cleanliness were represented in the dataset. The satisfaction ratings for these services ranged from 0 to 5. The Arrival Delay in Minutes feature was float type, while the remaining features were integer datatype. In Figure. 4.2, a visualization of certain features taken from the Airline Passenger Satisfaction dataset is displayed.

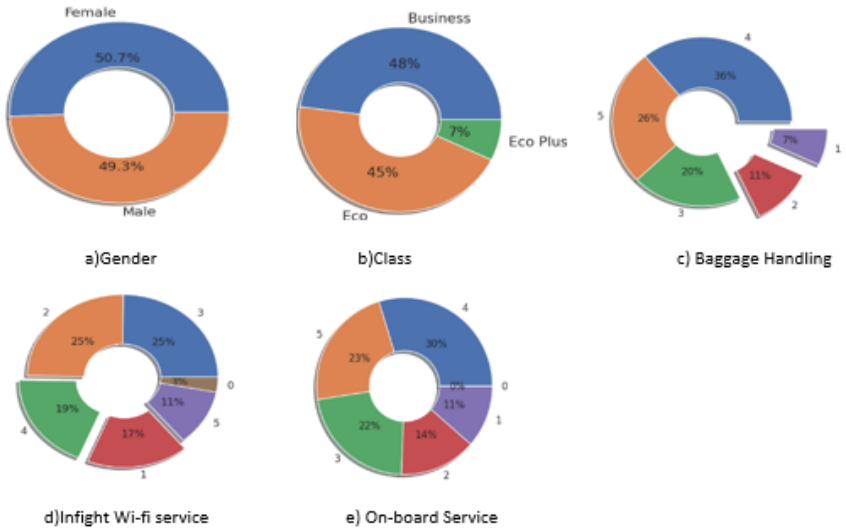


Figure 4.2: Features Visualization from Airline Passenger Satisfaction Dataset

4.3 Dataset Preprocessing

Data preprocessing involved cleaning, transforming, and integrating data to enhance its quality and suitability for analyzing specific machine learning models. During the cleaning process of the airline passenger dataset, which contained 310 null values, the null values were identified and subsequently filled with the mean values. In the dataset, the features of Gender, Customer Type, Type of Travel, and Class were initially classified as object data types. During the preprocessing phase, label encoding was performed on these features. This process assigned a unique numerical value to each distinct category within these features, which made it possible to perform subsequent analysis and modeling tasks. After label encoding, the dataset is transformed into a format that is presented in the Figure. 4.3. This conversion allows machine learning algorithms to process the data more effectively, as they typically require numerical inputs. The Figure. 4.3 likely displays the dataset after label encoding, showing the numerical representation of the categorical features.

MinMaxScaler is utilized to transform numerical features, ensuring they are scaled to a specific range, thereby promoting uniformity and facilitating various analytical tasks.

$$X_{\text{std}} = \frac{X - X_{\text{min(axis=0)}}}{X_{\text{max(axis=0)}} - X_{\text{min(axis=0)}}} \tag{4.1}$$

$$X_{\text{scaled}} = X_{\text{std}} \times (\text{max} - \text{min}) + \text{min} \tag{4.2}$$

Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	...	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	Inflight service	Cleanliness
1	0	13	1	2	460	3	4	...	5	4	3	4	4	5	5
1	1	25	0	0	235	3	2	...	1	1	5	3	1	4	1
0	0	26	0	0	1142	2	2	...	5	4	3	4	4	4	5
0	0	25	0	0	562	2	5	...	2	2	5	3	1	4	2
1	0	61	0	0	214	3	3	...	3	3	4	4	3	3	3
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
0	1	23	0	1	192	2	1	...	2	3	1	4	2	3	2
1	0	49	0	0	2347	4	4	...	5	5	5	5	5	5	4
1	1	30	0	0	1995	1	1	...	4	3	2	4	5	5	4
0	1	22	0	1	1000	1	1	...	1	4	5	1	5	4	1
1	0	27	0	0	1723	1	3	...	1	1	1	4	4	3	1

Figure 4.3: Dataset after label encoding

4.4 Feature Reduction

This study employed three distinct feature reduction techniques, referred to as Feature Collection 1, Feature Collection 2, and Feature Collection 3.

Table 4.1: Feature Reduction Techniques with the Number of Features

Feature Reduction Technique	No of features
Feature Collection 1	23 features
Feature Collection 2	20 features
Feature Collection 3	13 Features

4.4.1 Feature Collection 1

In this study, a feature reduction technique was employed to select a subset of features. The method involved identifying and eliminating highly correlated features. The dataset and a threshold value were used to determine the level of correlation at which features were considered highly correlated.

$$r = \frac{\sum((X_i - \bar{X}) \cdot (Y_i - \bar{Y}))}{\sqrt{\sum(X_i - \bar{X})^2} \cdot \sqrt{\sum(Y_i - \bar{Y})^2}} \tag{4.3}$$

When the absolute correlation coefficient exceeded the threshold, it indicated a strong correlation between the two features. In such cases, one of the correlated features was selected for removal. After eliminating the correlated feature, a total of 23 features were selected in this particular study.

4.4.2 Feature Collection 2

In this technique, feature reduction was performed using mutual information. Mutual information measured the dependence between the training data and the target variable. Sorting the mutual information scores in descending order helps identify the features with the highest mutual information regarding the target variable. The top 20 features with the highest mutual information scores were selected for further analysis. In the transformation process, the training and test datasets were modified to include only the selected features.

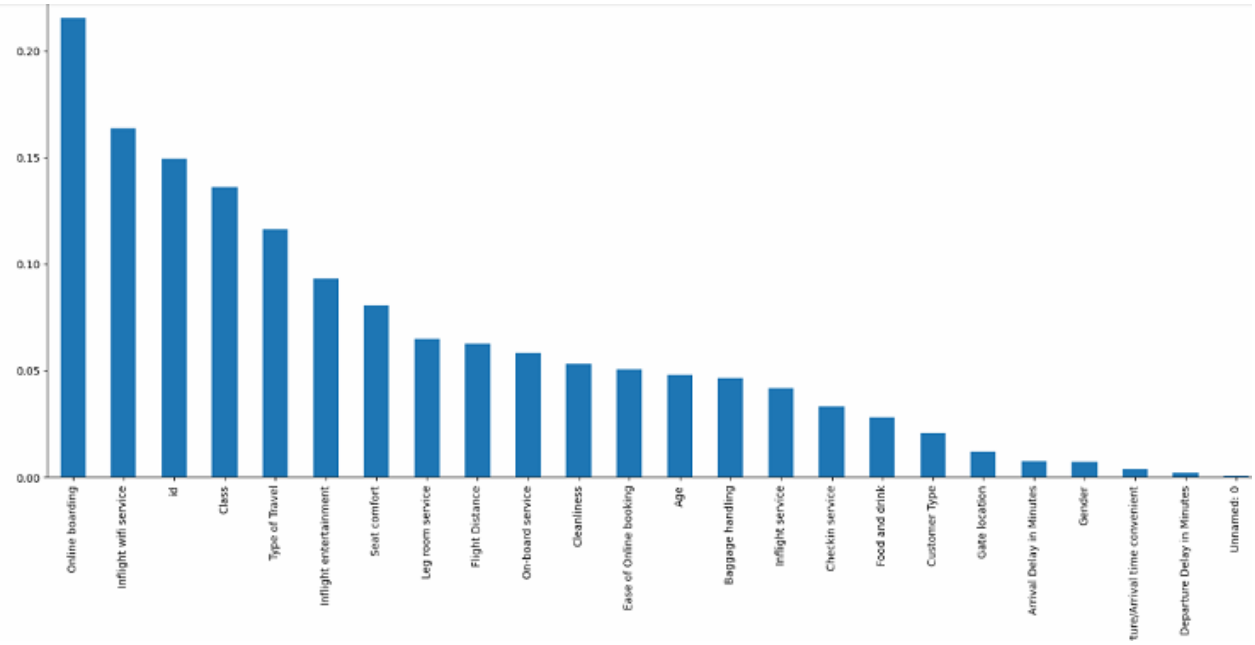


Figure 4.4: Mutual Information values of all features with a bar plot

4.4.3 Feature Collection 3

In this case, a Tree-based method, specifically Random Forest, was employed to perform feature reduction. Random Forest, being highly versatile, considers feature interactions and non-linear relationships, rendering it well-suited for a diverse range of data types and structures. One of its key advantages lies in its ability to mitigate overfitting, achieved through evaluating feature importance during the training process. The higher the importance score assigned to a feature, the greater its influence in making accurate predictions. Upon completion of the Random Forest training, the feature importance scores for each feature were obtained. Subsequently, a feature selection process was executed, sorting the features based on their importance scores in descending order. By retaining the top 13 features with the highest importance scores, a reduced

feature dataset was created, encompassing the most informative features.

4.5 Machine Learning Models-Based Framework

In this section, six different machine learning models and their functionalities are discussed. Each model serves a specific purpose.

4.5.1 Logistic Regression

Logistic regression stands out as a straightforward and highly efficient technique for solving binary and linear classification tasks. As a classification model, it offers ease of implementation while delivering impressive performance, particularly when dealing with classes that can be separated linearly[42].

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}} \tag{4.4}$$

Logistic Regression offers the advantage of producing well-calibrated probabilities in addition to classification outcomes, setting it apart from models that solely provide final class labels as outputs[43]. Logistic Regression is easy to understand and implement, making it a suitable choice for machine learning and data science. The model's coefficients represent the impact of each feature on the probability of the positive class, allowing for easy interpretation and understanding of the relationships between variables. Logistic Regression is computationally efficient and can handle large datasets with relatively low memory and processing requirements. When classes exhibit linear separability, Logistic Regression has the potential to yield outcomes of considerable accuracy. The sigmoid function transforms a straight line into a characteristic S-shaped curve[44].

4.5.2 Support Vector Machine

The Support Vector Machine (SVM) is a machine learning model tailored for the classification of data into two distinct classes, relying on a labeled training set. Its fundamental objective is to detect a hyperplane that can effectively segregate the two classes. While numerous hyperplanes can achieve this segregation, the SVM strives to identify the hyperplane boasting the broadest

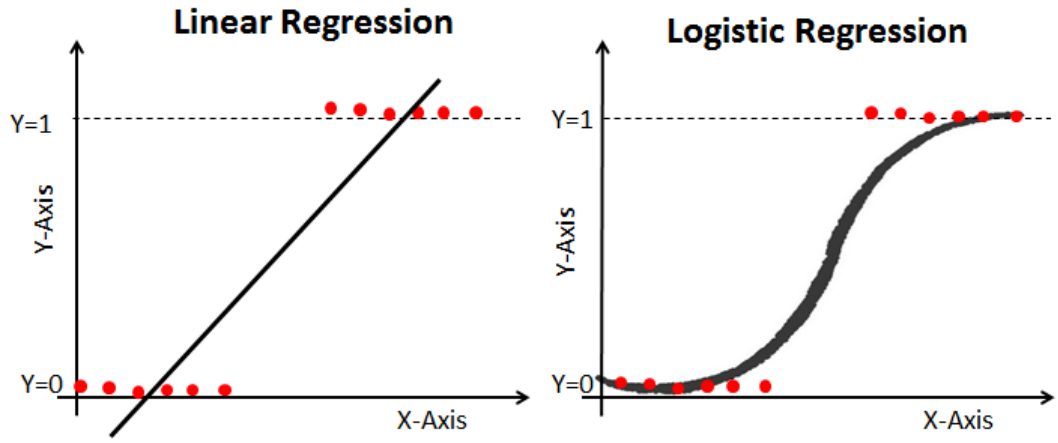


Figure 4.5: Linear Regression vs Logistic Regression [2]

margin, signifying the maximal gap between the two classes. The overarching aim is to enhance generalization, enabling the SVM to more accurately classify novel data points in subsequent instances[45].

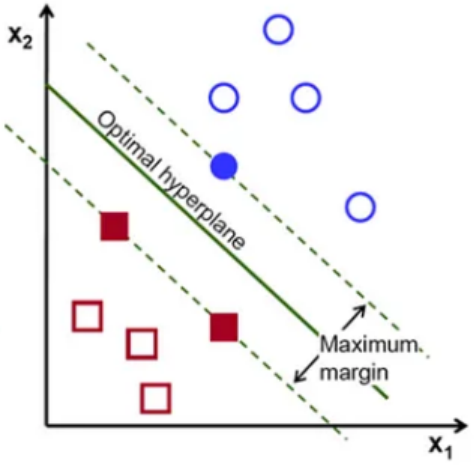


Figure 4.6: SVM with Potential Hyperplane [3]

The regularization parameter controls the penalty for misclassifying data points in the training set. A smaller value of the regularization parameter imposes a higher penalty for misclassification, encouraging the SVM to prioritize a larger margin even if it results in misclassifying a few training points. On the other hand, a larger value of regularization parameter allows the SVM to have fewer misclassifications on the training data, but this might lead to a smaller margin and potentially overfitting.

4.5.3 Random Forest

Random Forest stands as an ensemble classifier comprising several decision trees trained on distinct segments of the original dataset[46]. Falling within the realm of ensemble learning, Random Forest amalgamates multiple models to arrive at decisions collectively. Unlike relying solely on a solitary decision tree, this approach assembles a multitude of decision trees. Each tree is constructed independently, utilizing a unique subset of training data along with a random assortment of features. This divergence guarantees that the trees encapsulate diverse facets of the data, effectively curbing overfitting. When confronted with a novel data point for prediction, each decision tree within the Random Forest generates an individual prediction. The ultimate prediction is subsequently determined by consolidating outcomes through voting or averaging the predictions of individual trees.

By amalgamating forecasts from a multitude of decision trees, the Random Forest method mitigates the risk of overfitting, elevating the overall accuracy and resilience of the model. Notably, Random Forest adeptly manages datasets featuring numerous features, thanks to its mechanism of randomly selecting subsets of features for each tree, which serves to mitigate challenges stemming from high dimensionality. Furthermore, Random Forest yields insights into feature importance, highlighting which attributes wield the most substantial influence on the model's predictions. Notably, the training process of individual decision trees can be parallelized, rendering Random Forest training efficient and well-suited for expansive datasets.

4.5.4 Decision Tree

A Decision Tree constitutes a machine learning model applied to create a predictive framework through the assimilation of uncomplicated decision rules drawn from historical data (training data)[47]. These rules empower the model to prognosticate the class or value of the target variable for novel, uncharted data instances. An advantage of Decision Trees is their ability to handle non-linear data without necessitating the preprocessing of features. Unlike methods that entail the concurrent application of weighted combinations of multiple features, decision trees approach individual attributes in isolation, thereby negating the demand for such transformations.

$$H(X) = -\sum[p(x_i) \cdot \log_2(p(x_i))] \tag{4.5}$$

Entropy is a measure of impurity or uncertainty in a dataset[48]. In decision trees, it is used to evaluate how well a particular split separates the data into different classes. By minimizing entropy, decision trees strive to create pure and homogeneous subsets, which leads to more accurate and effective decision-making. Decision trees using entropy tend to be less sensitive to irrelevant features, as entropy focuses on the importance of attributes in reducing uncertainty rather than their absolute values. When selecting the best attribute to split the data, the attribute with the lowest entropy is chosen, indicating that it provides the most useful information for classifying the data. Calculating entropy is computationally efficient, making decision trees based on entropy a practical choice for various machine learning tasks.

4.5.5 Gussian Navie Bayes

Gaussian Naive Bayes is suitable for handling continuous-valued features and assumes that each feature follows a Gaussian (normal) distribution. To create a simple model, the approach assumes that the data conforms to a Gaussian distribution with no covariance (independent dimensions) between the features. This model is fitted by finding the mean and standard deviation of the data points within each label, which fully defines the Gaussian distribution for each class[4].

Gaussian Naive Bayes is straightforward and relatively easy to implement, making it accessible to beginners in machine learning and data science. The model's simplicity allows it to scale well with large datasets, making it computationally efficient and suitable for real-world applications with substantial data[49].The training process of Gaussian Naive Bayes is fast and efficient, particularly due to its independence assumption, which allows the model to calculate probabilities for each feature independently[49]. Gaussian Naive Bayes can effectively handle continuous-valued features, making it appropriate for datasets containing numeric data without requiring feature discretization. The algorithm can handle missing data during the training and prediction phases by ignoring the missing feature when estimating probabilities.

4.5.6 K Nearest Neighbor

The K-nearest neighbors (KNN) algorithm utilizes ”feature similarity” to make predictions for new data points, which involves assigning a value to the new data point based on its proximity

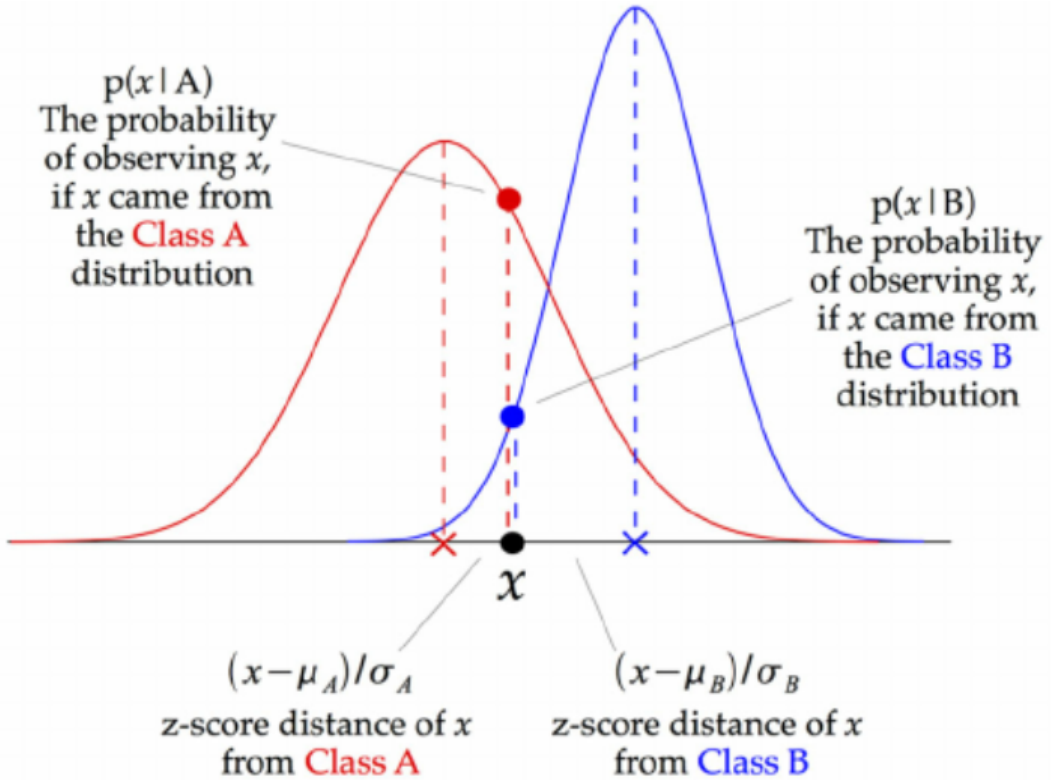


Figure 4.7: Gaussian Naive Bayes Classifier[4]

to the points in the training set[50].

KNN is an instance-based learning algorithm, which means it does not have a traditional

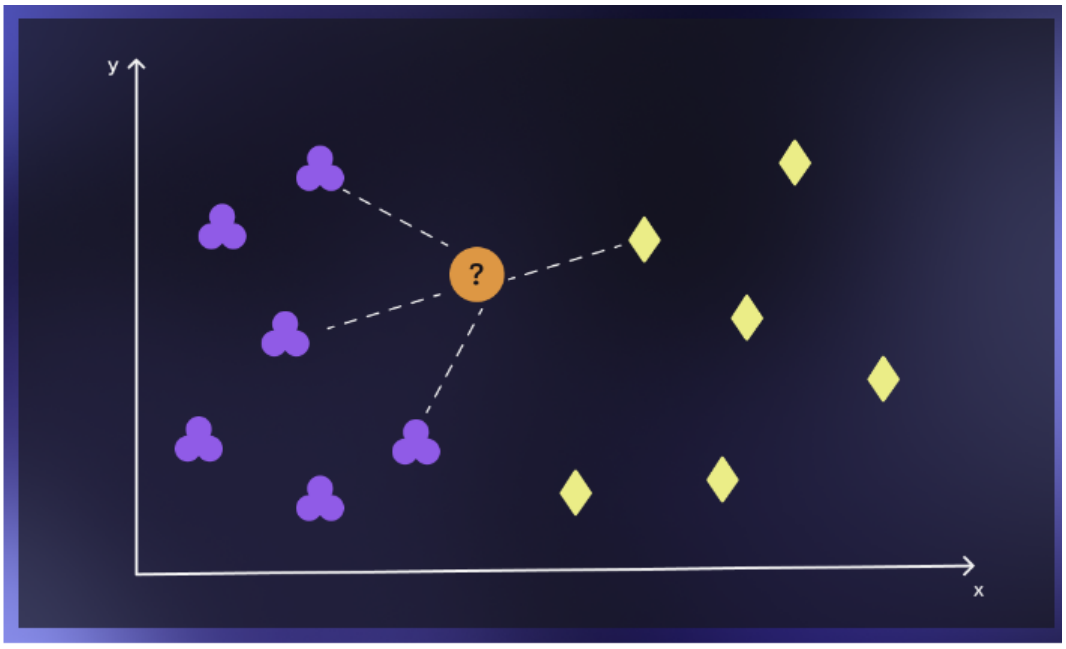


Figure 4.8: Find k-nearest neighbors[5]

training phase. The model is built directly from the training data, leading to a swift and efficient

training process. KNN can handle complex and non-linear relationships between features and the target variable, making it suitable for datasets with intricate decision boundaries. As a lazy learning algorithm, KNN does not build an explicit model during training. Instead, it directly stores the training data, making it easy to adapt to new data points as they become available.The K-nearest neighbors (KNN) model doesn't impose any presumptions regarding the inherent distribution of the data. Instead of learning explicit parameters from the training data, non-parametric KNN models rely on the data to make predictions[51].

4.6 Training Set Testing Set

The dataset was split into two subsets: the training and test subsets. The training subset comprises 80% of the entire dataset, while the test subset accounts for the remaining 20%. This division ensures that the model is trained on a majority of the data, allowing it to learn patterns and relationships present in the dataset. During the training process, the model uses the training data to adjust its parameters. The objective is to optimize the model's performance metric, which could be accuracy, precision, recall, F1-score, or any other relevant metric, depending on the specific problem. By fitting the training data into the model and adjusting its parameters, the model aims to find the best configuration that maximizes the chosen performance metric. The optimization process involves iteratively fine-tuning the model's parameters to achieve the highest possible performance on the training data. After the training phase, the model is ready for evaluation using the test data. The test data is data that the model has not seen during training, and its purpose is to assess the model's generalization ability.

4.7 Hyperparameter Tuning with GridSearch

Hyperparameter tuning with GridSearch is a systematic and exhaustive approach to finding the best machine learning model's best hyperparameters. Hyperparameters are settings or configurations that are not learned from the data during the training process but need to be specified by the user before training the model. They can significantly impact the model's performance and generalization.

GridSearch is a hyperparameter tuning technique that involves creating a grid of possible hyperparameter values and evaluating the model's performance for each combination of hyperpa-

rameters. The grid consists of different values for each hyperparameter that the user wants to tune. In GridSearch, the selection of hyperparameters is driven by a performance metric, usually evaluated through cross-validation on the training set or assessment on a separate validation set[52]. GridSearch processes:

- Choose Machine learning Model.
- Identify Hyperparameter.
- Construct a grid containing all possible combinations of hyperparameter values.
- For each combination of hyperparameters, use cross-validation to evaluate the model's performance
- The hyperparameter combination that yields the best performance metric on the cross-validation is selected as the optimal set of hyperparameters.
- Once the best hyperparameters are determined, the final model is trained using the entire training data with the chosen hyperparameters.

4.8 Conclusion

This chapter discusses the elements crucial for the suggested method in the machine learning-based architecture, including dataset, data pre-processing, feature reduction methods, machine learning models-based framework, training and testing set, and hyperparameter tuning with GridSearch.

Chapter 5

Result & Performance Analysis

5.1 Introduction

In this study, three distinct feature reduction methodologies were investigated to improve the performance of machine learning models. These methodologies were applied to six diverse models, enabling a comprehensive evaluation. The outcomes of each feature reduction technique were meticulously analyzed and thoroughly discussed in this section. To assess the impact of these techniques on model performance, various performance metrics, including accuracy, precision, f1-score, and recall, were utilized on the test data. This thorough evaluation allowed for direct comparisons between the different feature reduction methods, providing insights into their influence on the model's predictive capabilities. Valuable findings emerged, identifying the most promising methodologies for future applications in similar tasks or domains. The study's outcomes make valuable contributions to the field of feature reduction in machine learning and offer guidance for optimizing model performance in diverse real-world scenarios.

5.2 Evaluation Metrics

To assess the performance of models, several evaluation metrics are commonly used, including Precision, Recall, Accuracy, F1-score, and the confusion matrix. Before delving into the explanation of these evaluation metrics, it's crucial to understand the terms True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) in the context of classification.

True Positives (TP): These are instances where the model correctly predicts the positive class (1) and the actual value of the data point is also positive (1).

True Negatives (TN): These are instances where the model correctly predicts the negative class (0) and the actual value of the data point is also negative (0).

False Positives (FP): False positives occur when the model incorrectly predicts the positive class (1) while the actual value of the data point is negative (0). The term "false" is used because the model made an incorrect prediction, and "positive" indicates that the model predicted the positive class.

False Negatives (FN): False negatives occur when the model incorrectly predicts the negative class (0) while the actual value of the data point is positive (1). Similar to false positives, the term "false" is used because the model made an incorrect prediction, and "negative" indicates that the model predicted the negative class.

In summary, these metrics are used to quantify the performance of classification models by comparing their predictions with the actual labels of the data points. By understanding these metrics, we can now proceed to explain each evaluation metric and how they contribute to assessing the model's effectiveness.

5.2.1 Confusion Matrix

The confusion matrix is a vital evaluation tool for binary classification models, offering a compact overview of the model's predictions compared to the actual data labels. It comprises four essential elements: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). From these values, critical performance metrics such as Precision, Recall, Accuracy, and F1-score can be computed, enabling a comprehensive and quantitative assessment of the model's efficacy in distinguishing between the two classes. Precision quantifies the accuracy of positive predictions, Recall measures the model's ability to identify positive instances correctly, Accuracy provides an overall correctness measure, and F1-score balances Precision and Recall. By analyzing the confusion matrix and these derived metrics, researchers and practitioners gain valuable insights into the model's strengths and weaknesses, aiding informed decision-making and optimization strategies for achieving superior classification performance.

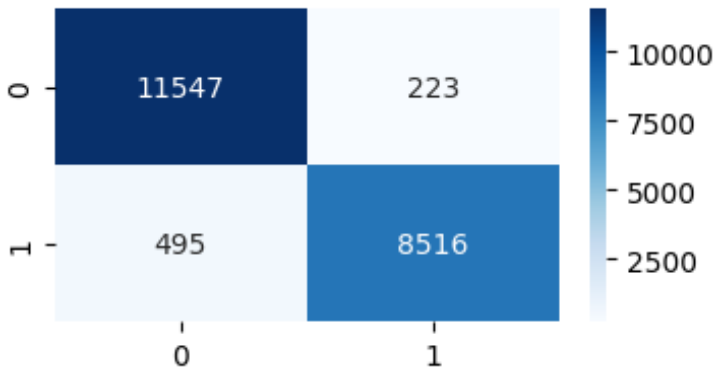


Figure 5.1: Confusion Matrix of Random Forest from Feature Reduction 2 for Proposed Architecture

5.2.2 Accuracy

Accuracy is a crucial metric used to evaluate the performance of machine learning algorithms, particularly in classification tasks. It measures the proportion of correct predictions made by the algorithm compared to the total number of predictions. It's essential to keep in mind that accuracy alone may not provide a complete picture of the model's performance, especially when dealing with imbalanced datasets or when the costs of false positives and false negatives are significantly different. Additionally, while accuracy is useful for evaluating the model's performance on the existing dataset. Therefore, it's essential to use techniques like cross-validation and hold-out testing to assess the model's generalization capability and avoid overfitting to the training data.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

5.2.3 Precision

Precision is a performance metric used in binary classification tasks to evaluate the model's ability to correctly identify positive instances (true positives) among all instances that the model predicted as positive (both true positives and false positives). In other words, precision measures the accuracy of positive predictions made by the model.

$$Precision = \frac{TP}{TP + FP} \tag{5.2}$$

A high precision score indicates that the model makes fewer false positive predictions and is more reliable in identifying positive instances correctly. It means that when the model predicts

a data point as positive, it is more likely to be correct. On the other hand, a low precision score suggests that the model tends to make more false positive predictions and may be less reliable in distinguishing between positive and negative instances.

5.2.4 Recall

Recall, also known as sensitivity or true positive rate, is another essential performance metric used in binary classification tasks. It measures the model's ability to correctly identify positive instances (true positives) out of all the instances that are actually positive (both true positives and false negatives).

$$Recall = \frac{TP}{TP + FN} \tag{5.3}$$

A high recall value indicates that the model is proficient in correctly capturing positive instances from the dataset. In other words, when the model encounters a positive data point, it is more likely to classify it as positive correctly. On the other hand, a low recall value suggests that the model tends to miss positive instances and is less effective in identifying all the positive cases accurately.

5.2.5 F1-Score

The F1-score is a performance metric used in binary classification tasks to strike a balance between precision and recall. It is particularly useful when the dataset is imbalanced, meaning that one class is much more prevalent than the other. The F1-score provides a single value that considers both precision and recall, giving a more comprehensive evaluation of the model's performance.

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5.4}$$

When precision and recall have similar values, the F1-score will be close to 1, indicating a balanced model performance in terms of both positive predictions' accuracy and the ability to capture all positive instances. However, if precision and recall differ significantly, the F1-score will be lower, reflecting the model's inability to achieve a balance between the two metrics.

5.3 Model Performance

Three distinct techniques for feature reduction were applied to six different models, resulting in valuable insights and noteworthy outcomes. Applying these feature reduction techniques across six diverse models means that the methods were utilized on different machine learning algorithms, such as Decision Trees, Random Forests, Logistic Regression, Support Vector Machines (SVM), and others. Each model may have specific requirements and perform differently based on the reduced feature set. The results of these feature reduction methodologies would have provided valuable insights into the impact of feature selection on model performance, interpretability, and generalization to new data. Additionally, it could highlight the trade-offs between model complexity and predictive accuracy, allowing for a deeper understanding of the underlying data patterns and the model's behavior.

5.3.1 Best Parameters Found Using GridSearch

This study utilized GridSearch for optimizing machine learning models. Hyperparameters were tuned through param-grid to train the models, and the outcomes were derived accordingly. The best-performing parameter combinations were then listed in a table, showcasing the optimal configurations that yielded the highest performance metrics. GridSearch systematically explores the hyperparameter space to find the most suitable values, leading to improved model performance. The table presents the results, allowing readers to identify the parameter settings that produce the best results for the given machine learning tasks.

Table 5.1: Best parameters of machine learning models using GridSearch

Model	Best-Parameter(GridSearchCV)
Logistic Regression	$C = 0.008885$, Solver = saga
KNN	Metric = manhattan, N_neighbors = 9, Weights = uniform
Random Forest	max_features = sqrt, max_depth = None, min_samples_leaf = 2, min_samples_split = 2, n_estimators = 4000
Decision Tree	Criterion = entropy, max_depth = 9, min_samples_leaf = 4, min_samples_split = 5
Gaussian Naïve Bayes	Priors = None, Var_smoothing = 0.01
Support Vector Machine	C = 10, Degree = 3, Gamma = scale, Kernel = rbf

5.3.2 Model Performance after Feature Reduction 1 with Accuracy Precision, Recall, F1-Score

One of the features that showed a high correlation with another feature was removed from the dataset, and the subsequent impact on model performance was analyzed.

Table 5.2: Model Performance - Feature Collection 1

Model	Performance	Feature Collection 1
Random Forest	Accuracy	0.96
	Precision	0.96
	Recall	0.96
	F1-score	0.96
Support Vector Machine	Accuracy	0.95
	Precision	0.95
	Recall	0.95
	F1-score	0.95
Decision Tree	Accuracy	0.95
	Precision	0.95
	Recall	0.95
	F1-score	0.95
K Nearest Neighbor	Accuracy	0.92
	Precision	0.92
	Recall	0.92
	F1-score	0.92
Gaussian Naïve Bayes	Accuracy	0.87
	Precision	0.87
	Recall	0.87
	F1-score	0.87
Logistic Regression	Accuracy	0.88
	Precision	0.88
	Recall	0.88
	F1-score	0.88

5.3.3 Model Performance after Feature Reduction 1 with Confusion Matrix

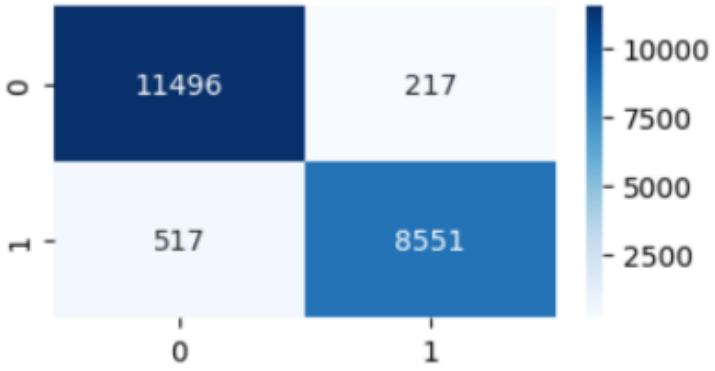


Figure 5.2: Confusion Matrix of Random Forest from Feature Reduction 1 for Proposed Architecture

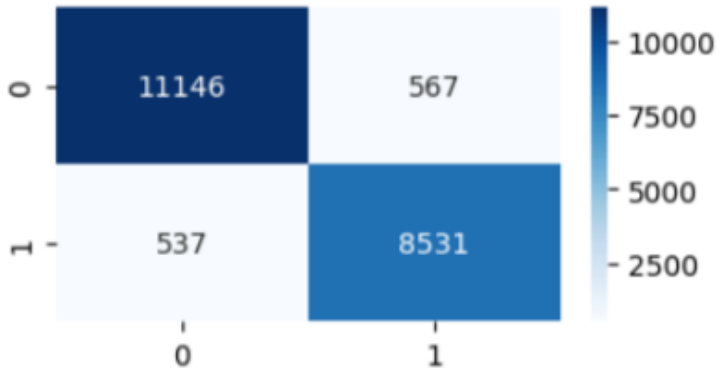


Figure 5.3: Confusion Matrix of Decision Tree from Feature Reduction 1 for Proposed Architecture

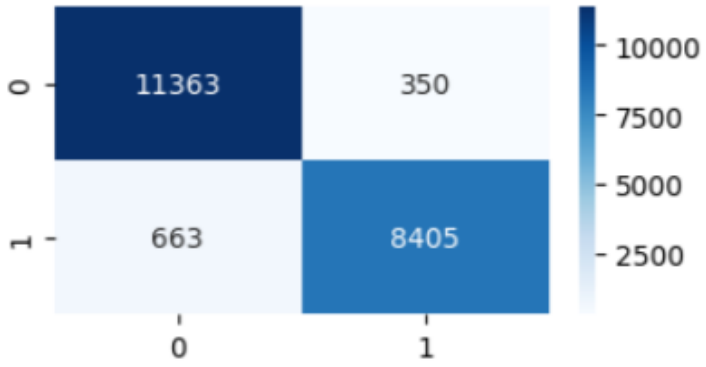


Figure 5.4: Confusion Matrix of Support Vector Machine from Feature Reduction 1 for Proposed Architecture

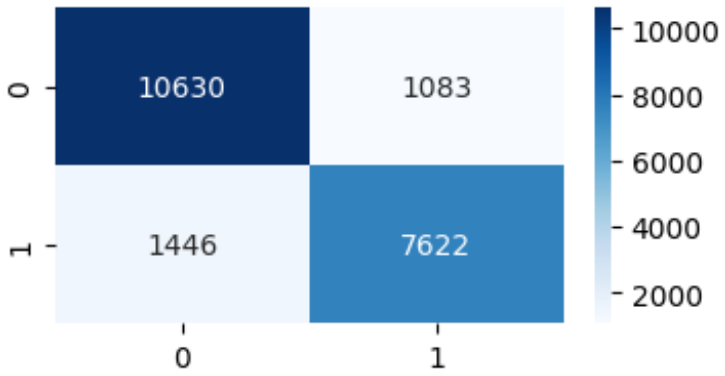


Figure 5.5: Confusion Matrix of Logistic Regression from Feature Reduction 1 for Proposed Architecture

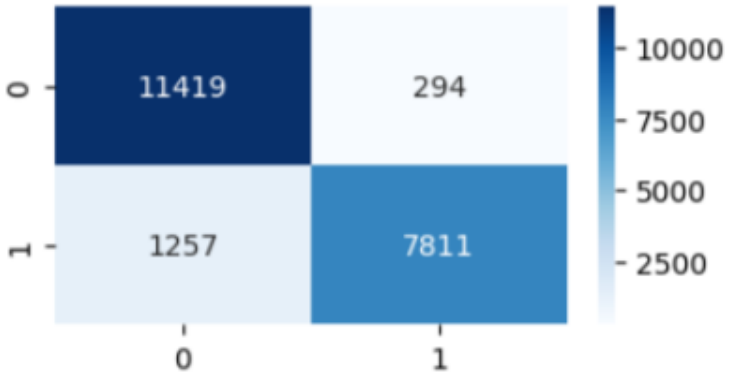


Figure 5.6: Confusion Matrix of K Nearest Neighbor from Feature Reduction 1 for Proposed Architecture

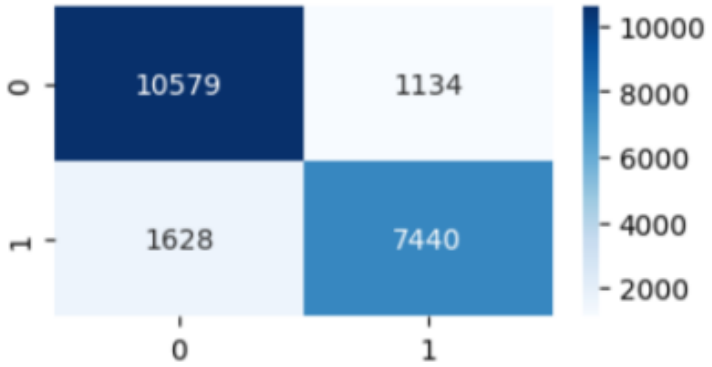


Figure 5.7: Confusion Matrix of Gaussian Naïve Bayes from Feature Reduction 1 for Proposed Architecture

5.3.4 Model Performance after Feature Reduction 2 with Accuracy Precision, Recall, F1-Score

Mutual Information is employed as a technique to decrease the number of features in the dataset.

Table 5.3: Model Performance - Feature Collection 2

Model	Performance	Feature Collection 2
Random Forest	Accuracy	0.97
	Precision	0.97
	Recall	0.97
	F1-score	0.97
Support Vector Machine	Accuracy	0.65
	Precision	0.66
	Recall	0.65
	F1-score	0.62
Decision Tree	Accuracy	0.95
	Precision	0.95
	Recall	0.95
	F1-score	0.95
K Nearest Neighbor	Accuracy	0.68
	Precision	0.68
	Recall	0.68
	F1-score	0.68
Gaussian Naïve Bayes	Accuracy	0.82
	Precision	0.82
	Recall	0.82
	F1-score	0.82
Logistic Regression	Accuracy	0.79
	Precision	0.79
	Recall	0.79
	F1-score	0.79

5.3.5 Model Performance after Feature Reduction 2 with Confusion Matrix

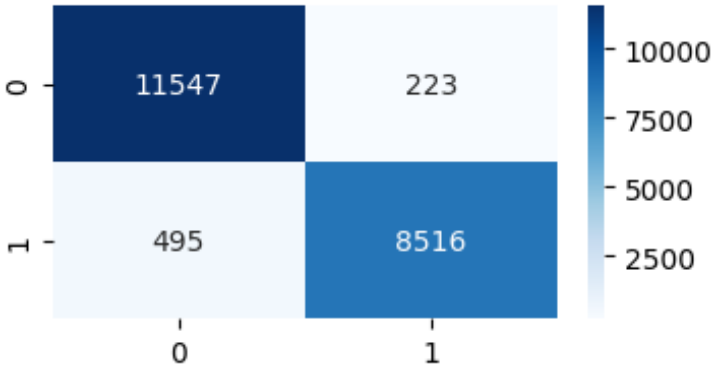


Figure 5.8: Confusion Matrix of Random Forest from Feature Reduction 2 for Proposed Architecture

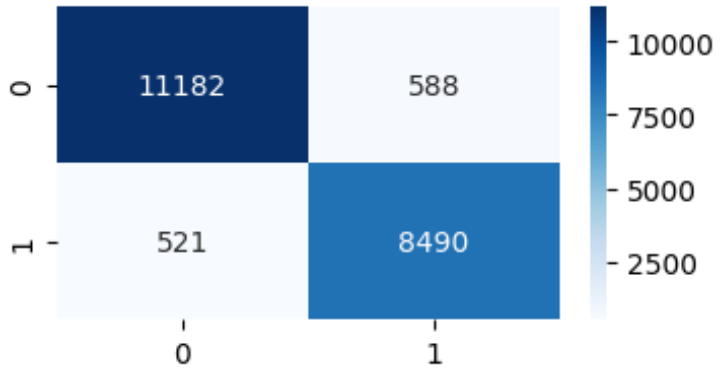


Figure 5.9: Confusion Matrix of Decision Tree from Feature Reduction 2 for Proposed Architecture

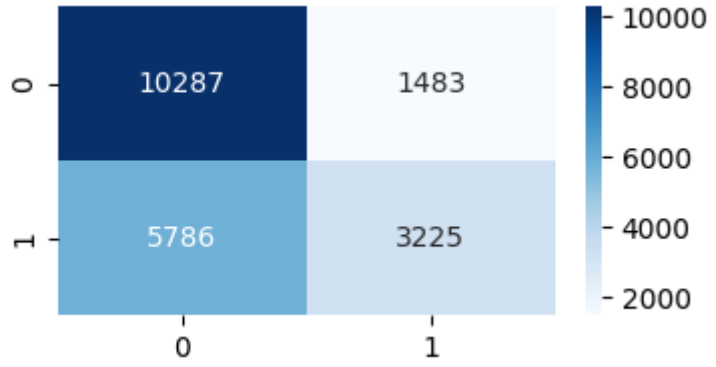


Figure 5.10: Confusion Matrix of Support Vector Machine from Feature Reduction 2 for Proposed Architecture

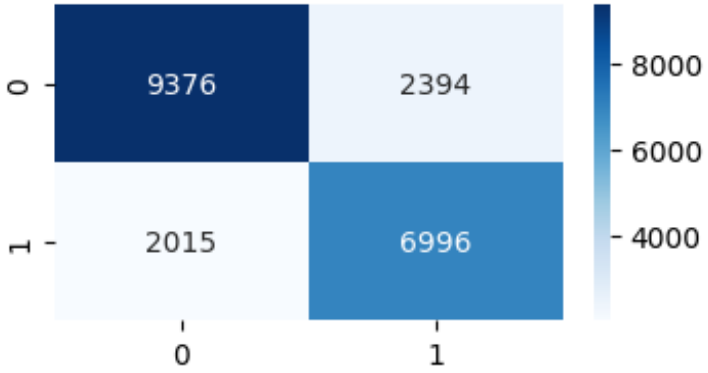


Figure 5.11: Confusion Matrix of Logistic Regression from Feature Reduction 2 for Proposed Architecture

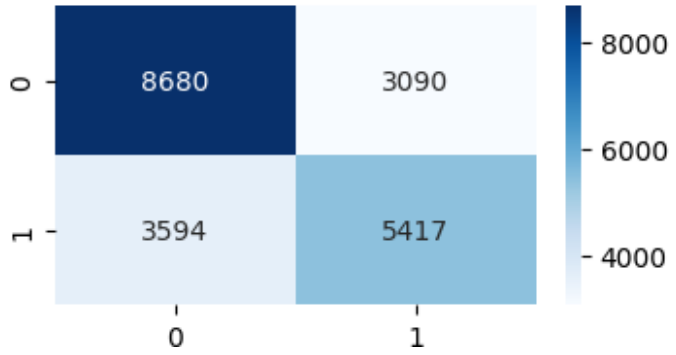


Figure 5.12: Confusion Matrix of K Nearest Neighbor from Feature Reduction 2 for Proposed Architecture

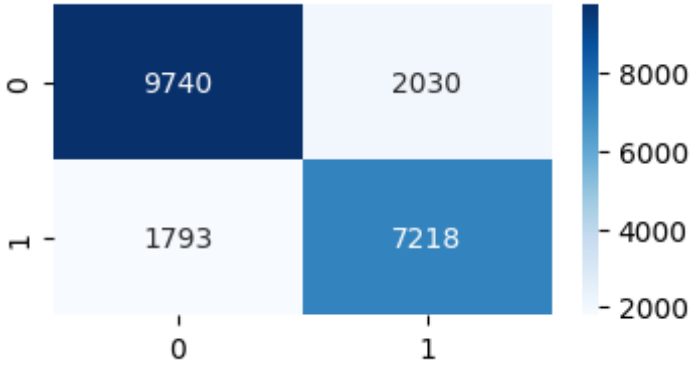


Figure 5.13: Confusion Matrix of Gaussian Naïve Bayes from Feature Reduction 2 for Proposed Architecture

5.3.6 Model Performance after Feature Reduction 3 with Accuracy Precision, Recall, F1-Score

A tree-based feature reduction method was implemented to decrease the number of features in the dataset. This method uses Random Forests to identify the most relevant features that

contribute significantly to the model's predictive power.

Table 5.4: Model Performance - Feature Collection 3

Model	Performance	Feature Collection 3
Random Forest	Accuracy	0.95
	Precision	0.96
	Recall	0.95
	F1-score	0.95
Support Vector Machine	Accuracy	0.94
	Precision	0.94
	Recall	0.94
	F1-score	0.94
Decision Tree	Accuracy	0.93
	Precision	0.93
	Recall	0.93
	F1-score	0.93
K Nearest Neighbor	Accuracy	0.94
	Precision	0.94
	Recall	0.94
	F1-score	0.94
Gaussian Naïve Bayes	Accuracy	0.87
	Precision	0.87
	Recall	0.87
	F1-score	0.87
Logistic Regression	Accuracy	0.87
	Precision	0.87
	Recall	0.87
	F1-score	0.87

5.3.7 Model Performance after Feature Reduction 3 with Confusion Matrix

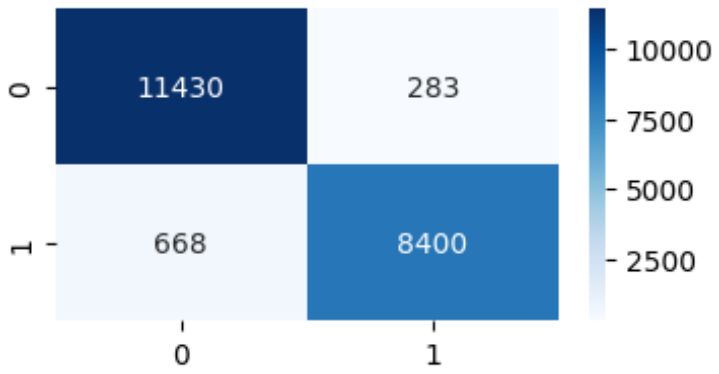


Figure 5.14: Confusion Matrix of Random Forest from Feature Reduction 3 for Proposed Architecture

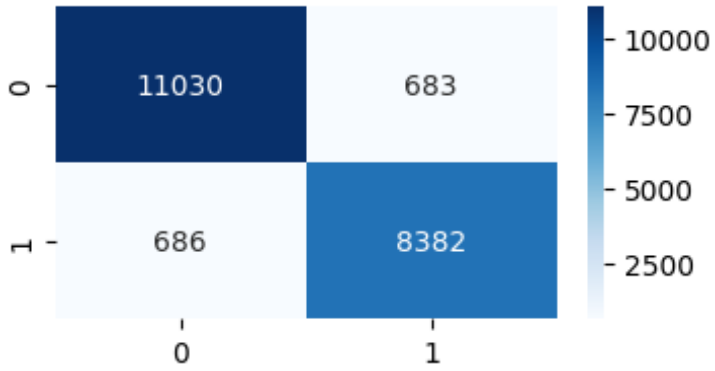


Figure 5.15: Confusion Matrix of Decision Tree from Feature Reduction 3 for Proposed Architecture

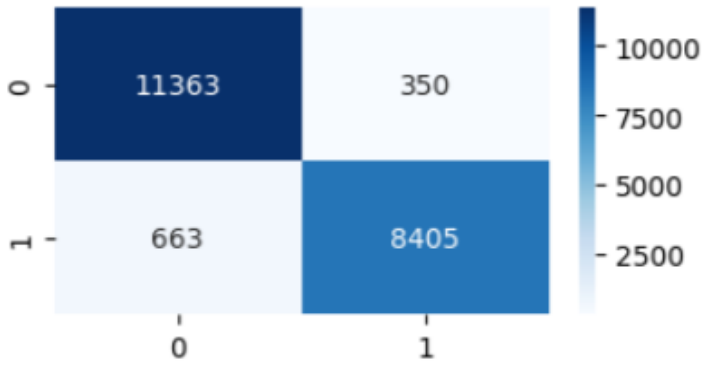


Figure 5.16: Confusion Matrix of Support Vector Machine from Feature Reduction 3 for Proposed Architecture

The results indicate that both Random Forest and Decision Tree models yield superior outcomes. These favorable outcomes were obtained through the three feature collection methods. It is worth noting that these models achieved impressive accuracy rates of 97% and 95%, re-

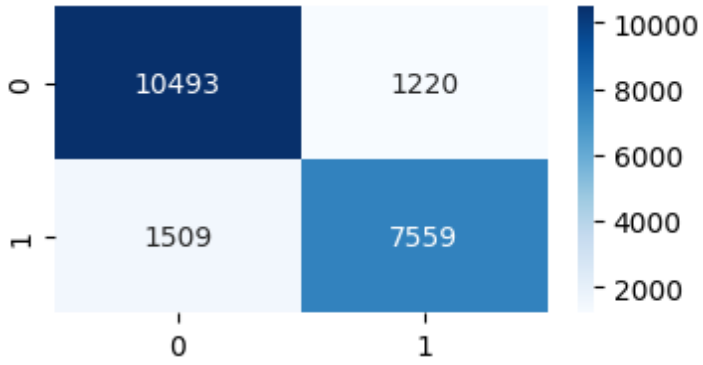


Figure 5.17: Confusion Matrix of Logistic Regression from Feature Reduction 3 for Proposed Architecture

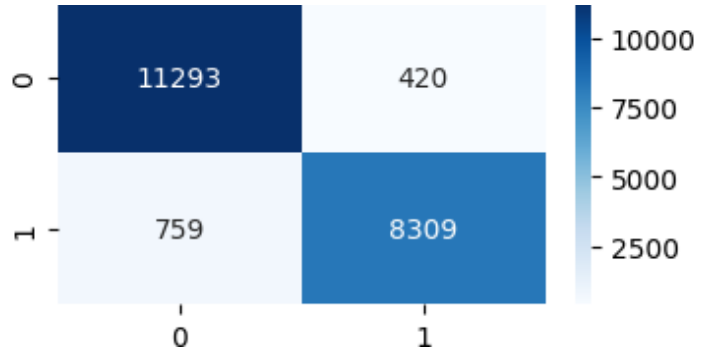


Figure 5.18: Confusion Matrix of K Nearest Neighbor from Feature Reduction 3 for Proposed Architecture

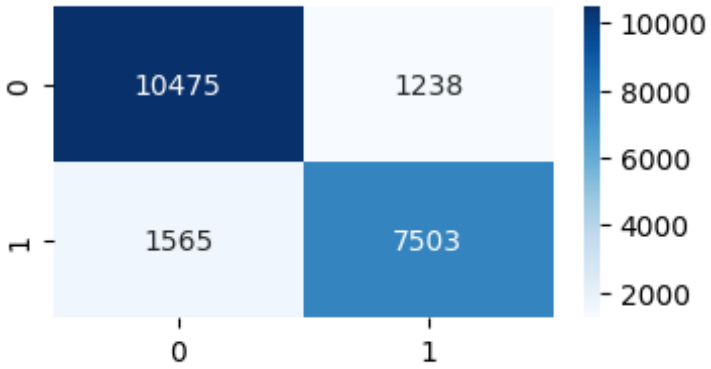


Figure 5.19: Confusion Matrix of Gaussian Naïve Bayes from Feature Reduction 3 for Proposed Architecture

spectively. Furthermore, their precision scores were also commendable, standing at 97% for Random Forest and 95% for Decision Tree. Additionally, the recall rates were found to be 97% and 95% for Random Forest and Decision Tree, respectively, further demonstrating their strong performance. The F1 scores, which signify a balance between precision and recall, were also

notably high, reaching 97% for Random Forest and 95% for Decision Tree.

5.3.8 Comparison of the Proposed Model

The proposed model's performance is compared with other models, either similar models with the same dataset or different models with similar and different datasets. The goal is to identify the model that achieves the highest accuracy, precision, recall, f1-score, or any other relevant performance metric. The idea is to understand how the proposed model's performance stacks up against other algorithms when dealing with the same type of data. This type of comparison helps assess the model's ability to generalize and perform well on unseen data that might have different characteristics. Table. 5.5 presents a comparison.

Table 5.5: A Comparative analysis of different works with the proposed Machine learning model. ACC: ACCURACY; PREC: PRECISION; REC: RECALL

Authors	Datasets	Models used	Major Findings (only best results)
Baswardono et al. [35]	US Airline Passenger Satisfaction [53]	Random Forest and C4.5	Random Forest Acc: 0.9332; Prec: 0.9282; Rec: 0.9559
Suwanto et al. [37]	Airline Passenger Satisfaction [41]	Random Forest, AdaBoost, XGBoost	Random Forest Acc: 0.894; Prec: 0.90208; Rec: 0.89413
AlHabbal et al. [38]	US Airline passenger satisfaction survey [53]	Decision Tree, Random Forest, Logistic Regression, KNN, ANN	Random Forest Acc: 0.9595; Prec: 0.9525; Rec: 0.9728
Hatıpoğlu et al. [39]	Turkish airline company flight data	XGBoost, LightGBM, CatBoost	XGBoost Acc: 0.96; Rec: 0.903
Homaid et al. [36]	SKYTRAX website data [54]	Logistic Regression, XGBoost, SVM, Random Forest, Naïve Bayes	XGBoost Acc: 0.88; Prec: 0.85; Rec: 0.83
Proposed Machine Learning Model (Feature Collection 2)	Airline Passenger Satisfaction [41]	Logistic Regression, Gaussian Naïve Bayes, K Nearest Neighbor, Decision Tree, Random Forest, Support Vector Machine	Random Forest Acc: 0.97; Prec: 0.97; Rec: 0.97; F1 score: 0.97

5.4 Conclusion

This chapter incorporates the utilization of the confusion matrix and various evaluation metrics to measure the model performance for all the models. By comparing the performance of the suggested machine learning approach with relevant architectures, the findings demonstrate that the proposed approach outperformed all the other methods.

Chapter 6

Conclusion & Future Works

6.1 Introduction

In this chapter, a concise overview of the entire research is presented, covering the problem domain, previous studies, the novel contributions made, experimental analysis, and the conclusions drawn. Additionally, a brief glimpse into potential future directions for further research is provided.

6.2 Summary

Earlier research on this topic achieved a Random Forest accuracy of 95%. However, this paper proposes novel methods, including feature reduction and hyperparameter tuning, which lead to an improved Random Forest accuracy of 97%. Moreover, the paper introduces a unique feature reduction technique different from the methodologies used in other papers. Through the implementation of these methods, the proposed approach achieves an enhanced Random Forest accuracy of 97%, surpassing the results of previous research. The findings indicate that the novel feature reduction method and the optimized hyperparameter settings contribute significantly to the model's improved performance.

6.3 Conclusion

According to the findings from the comparative analysis, the Random Forest model, when coupled with the proposed feature reduction technique and GridSearch, demonstrates superior per-

formance. The accuracy achieved by the Random Forest model is an impressive 97%. GridSearch plays a crucial role in identifying the most optimal configuration for the Random Forest model, leading to enhanced accuracy in predicting customer satisfaction. As a result, this optimized Random Forest model becomes a valuable tool in accurately predicting customer satisfaction levels in the aviation industry. By employing the optimized Random Forest model, the aviation company will possess a powerful tool that enables more precise predictions of customer satisfaction. The feature reduction technique enhances the model's ability to focus on the most relevant and influential aspects of the data, thereby improving its overall performance. Through this investigation, the company can pinpoint the features that have the most significant impact on customer satisfaction levels. This knowledge is invaluable in making informed decisions to prioritize improvements and allocate resources strategically. By identifying which features are more precise indicators of customer satisfaction, the aviation company can tailor its services to meet customer expectations better and ultimately enhance overall customer experience.

6.4 Limitation

Though the proposed model outperformed, it also possesses some limitations. For hyperparameter tuning, this paper used GridSearch, which has several limitations:

- GridSearch operates on a predefined grid of hyperparameter values. If the optimal hyperparameter values lie outside the specified grid, GridSearch might miss them, leading to suboptimal model performance.

- Defining the grid of hyperparameter values requires manual input from the user. Selecting appropriate ranges for hyperparameters might require domain knowledge or experimentation, which can be time-consuming.

- GridSearch is not efficient for continuous hyperparameters, as it can only explore a limited set of discrete values.

6.5 Future Works

There are numerous possibilities to expand the scope of this study in the future. Here are a few examples of potential tasks:

- The main focus will be on improving the model's accuracy through extensive exploration of various hyperparameter configurations using GridSearch.

- The goal is to identify the best combination that maximizes accuracy.

- By continuously aiming to enhance accuracy, organizations can make well-informed decisions, enhance customer experiences, and offer improved services.

- Can be considered other approaches for better performance.

If circumstances allow, I aspire to make a contribution to these challenges in the coming time, with the goal of enhancing the model to be more versatile and widely applicable than ever before.

REFERENCES

[1] "Flightradar24." `https://www.flightradar24.com`. Accessed on August 6, 2023.

[2] "DataCamp." Accessed on: August 6, 2023.

[3] immohann, "Support Vector Machine Explained and Implemented," 2021. Accessed on August 6, 2023.

[4] OpenGenus IQ, "Gaussian Naive Bayes." Accessed on: August 6, 2023.

[5] Serokell, "K-Nearest Neighbors (KNN) Algorithm in Machine Learning." Accessed on: August 6, 2023.

[6] S. Kumar and S. M. Nafi, "Impact of covid-19 pandemic on tourism: Perceptions from bangladesh," *Available at SSRN 3632798*, 2020.

[7] D. T. Duval, "Critical issues in air transport and tourism," *Tourism geographies*, vol. 15, no. 3, pp. 494–510, 2013.

[8] M. Gill, "Aviation benefits beyond borders," *Air Transport Action Group (ATAG)*, vol. 1, no. 1, pp. 1–76, 2016.

[9] A. B. B. Borders, "Tourism enabler." Accessed on August 6, 2023.

[10] Organisation for Economic Co-operation and Development, "COVID-19 and the Aviation Industry: Impact and Policy Responses," 2020. Accessed on: August 6, 2023.

[11] International Air Transport Association, "Air Passenger Monthly Analysis - April 2020," 2020. Accessed on: August 6, 2023.

[12] International Air Transport Association, "Air Passenger Monthly Analysis - August 2020," 2020. Accessed on August 6, 2023.

[13] International Air Transport Association, "Air Freight Monthly Analysis - April 2020," 2020. Accessed on August 6, 2023.

[14] International Air Transport Association, "Air Freight Monthly Analysis - August 2020," 2020. Accessed on August 6, 2023.

[15] Survicate, "The Importance of Customer Satisfaction," 2023. Accessed on August 6, 2023.

[16] I.-J. Park, J. Kim, S. S. Kim, J. C. Lee, and M. Giroux, "Impact of the covid-19 pandemic on travelers' preference for crowded versus non-crowded options," *Tourism Management*, vol. 87, p. 104398, 2021.

[17] Y. Religia, A. Amali, *et al.*, "Perbandingan optimasi feature selection pada naïve bayes untuk klasifikasi kepuasan airline passenger," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 3, pp. 527–533, 2021.

[18] GeeksforGeeks, "Particle Swarm Optimization (PSO) - An Overview," 2023. Accessed on August 6, 2023.

[19] TutorialsPoint, "Genetic Algorithms - Introduction." Accessed on August 6, 2023.

[20] J. S. Cobb and M. A. Seale, "Examining the effect of social distancing on the compound growth rate of covid-19 at the county level (united states) using statistical analyses and a random forest machine learning model," *Public health*, vol. 185, pp. 27–29, 2020.

[21] K. W. Walker and Z. Jiang, "Application of adaptive boosting (adaboost) in demand-driven acquisition (dda) prediction: A machine-learning approach," *The Journal of Academic Librarianship*, vol. 45, no. 3, pp. 203–212, 2019.

[22] S. Chen, B. Shen, X. Wang, and S.-J. Yoo, "A strong machine learning classifier and decision stumps based hybrid adaboost classification algorithm for cognitive radios," *Sensors*, vol. 19, no. 23, p. 5077, 2019.

[23] Y. Wei, L. Shutao, and T. Mingkui, "Gene selection method based on svm-rfe-sfs," *Chin. J. Biomed. Eng*, vol. 29, no. 1, pp. 93–99, 2010.

[24] S. Tsafarakis, T. Kokotas, and A. Pantouvakis, "A multiple criteria approach for airline passenger satisfaction measurement and service quality improvement," *Journal of air transport management*, vol. 68, pp. 61–75, 2018.

[25] R. Etemad-Sajadi, S. A. Way, and L. Bohrer, "Airline passenger loyalty: The distinct effects of airline passenger perceived pre-flight and in-flight service quality," *Cornell Hospitality Quarterly*, vol. 57, no. 2, pp. 219–225, 2016.

[26] Q. Chen, Z. Meng, X. Liu, Q. Jin, and R. Su, "Decision variants for the automatic determination of optimal feature subset in rf-rfe," *Genes*, vol. 9, no. 6, p. 301, 2018.

[27] M. Ontivero-Ortega, A. Lage-Castellanos, G. Valente, R. Goebel, and M. Valdes-Sosa, "Fast gaussian naïve bayes for searchlight classification analysis," *Neuroimage*, vol. 163, pp. 471–479, 2017.

[28] B. Gregorutti, B. Michel, and P. Saint-Pierre, "Correlation and variable importance in random forests," *Statistics and Computing*, vol. 27, pp. 659–678, 2017.

[29] A. Naik and L. Samant, "Correlation review of classification algorithm using data mining tool: Weka, rapidminer, tanagra, orange and knime," *Procedia Computer Science*, vol. 85, pp. 662–668, 2016.

[30] S. Kumar and M. Zymbler, "A machine learning approach to analyze customer satisfaction from airline tweets," *Journal of Big Data*, vol. 6, no. 1, pp. 1–16, 2019.

[31] L. Moreira, C. Dantas, L. Oliveira, J. Soares, and E. Ogasawara, "On evaluating data preprocessing methods for machine learning models for flight delays," in *2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2018.

[32] N. Kuhn and N. Jamadagni, "Application of machine learning algorithms to predict flight arrival delays," *CS229*, pp. 2326–9865, 2017.

[33] K. Gopalakrishnan and H. Balakrishnan, "A comparative analysis of models for predicting delays in air traffic networks," ATM Seminar, 2017.

[34] R. Britto, M. Dresner, and A. Voltes, "The impact of flight delays on passenger demand and societal welfare," *Transportation Research Part E: Logistics and Transportation Review*, vol. 48, no. 2, pp. 460–469, 2012.

[35] W. Baswardono, D. Kurniadi, A. Mulyani, and D. Arifin, "Comparative analysis of decision tree algorithms: Random forest and c4. 5 for airlines customer satisfaction classification," in *Journal of Physics: Conference Series*, vol. 1402, p. 066055, IOP Publishing, 2019.

[36] M. S. Homaid and I. Moulitsas, "Measuring airport service quality using machine learning algorithms," in *Proceedings of the 6th International Conference on Advances in Artificial Intelligence*, pp. 8–14, 2022.

[37] E. Indra, J. Suwanto, D. R. H. Sitompul, S. H. Sinurat, A. Situmorang, R. Ruben, D. J. Ziegel, *et al.*, "Comparison of classification algorithm in classifying airline passenger satisfaction," *Jurnal Sistem Informasi dan Ilmu Komputer Prima (JUSIKOM PRIMA)*, vol. 6, no. 1, pp. 92–98, 2022.

[38] M. R. AlHabbal, "Predicting & optimizing airlines customer satisfaction using classification," 2022.

[39] I. Hatıpoğlu, Ö. Tosun, and N. Tosun, "Flight delay prediction based with machine learning," *LogForum*, vol. 18, no. 1, 2022.

[40] K. A. F. A. Samah, N. F. A. Misdan, M. N. H. H. Jono, and L. S. Riza, "The best malaysian airline companies visualization through bilingual twitter sentiment analysis: A machine learning classification," *JOIV: International Journal on Informatics Visualization*, vol. 6, no. 1, pp. 130–137, 2022.

[41] T. KLEIN, "Airline Passenger Satisfaction." [Online], 2020.

[42] ScienceDirect, "Logistic Regression," 2017. Accessed on August 6, 2023.

[43] OpenGenus Foundation, "Advantages and Disadvantages of Logistic Regression." Accessed on August 6, 2023.

[44] Analytics Vidhya, "Conceptual Understanding of Logistic Regression for Data Science Beginners." Accessed on: August 6, 2023.

[45] Analytics Steps, "How Does the Support Vector Machine Algorithm Work in Machine Learning?," 2020. Accessed on: August 6, 2023.

[46] JavaTpoint, "Machine Learning - Random Forest Algorithm." Accessed on August 6, 2023.

[47] KDnuggets, "Decision Tree Algorithm Explained." Accessed on August 6, 2023.

[48] Towards Data Science, “Decision Trees Explained: Entropy, Information Gain, Gini Index, CCP Pruning.” Accessed on August 6, 2023.

[49] “Gaussian Naive Bayes: Advantages and Disadvantages.” Accessed on: August 6, 2023.

[50] TutorialsPoint, “Machine Learning with Python - K-Nearest Neighbors Algorithm: Finding Nearest Neighbors.” Accessed on August 6, 2023.

[51] MyGreatLearning, “KNN Algorithm - An Introduction.” Accessed on August 6, 2023.

[52] Wikipedia, “Hyperparameter optimization.” Accessed on August 6, 2023.

[53] “US Airline Passenger Satisfaction Survey.” [Online].

[54] “A-Z Airport Reviews - SKYTRAX.” [Online].