

**Republic of Iraq**  
**Ministry of Higher Education and Scientific**  
**Research University of Babylon**  
**College of Information Technology**  
**Software Department**



**Predicting customers' satisfaction with airline  
companies: The application of machine and deep  
learning techniques**

A Thesis

Submitted to the Council of the College of Information Technology for  
Postgraduate Studies of the University of Babylon in Partial Fulfillment of the  
Requirements for the Degree of Master in Information Technology – Software

*By*

**Shahad Hussein Ali Hussein**

*Supervised by*

**Assist. Prof. Dr. Ahmed Habeeb Said Al-Azawei**

**2023 A.D.**

**1445 A.H.**

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

﴿إِنِّي جَزَيْتُهُمُ الْيَوْمَ بِمَا صَبَرُوا أَنَّهُمْ هُمُ  
الْفَائِزُونَ﴾

بِسْمِ اللَّهِ  
الرَّحْمَنِ الرَّحِيمِ

سورة المؤمنون

## **Dedications**

**To the savior of mankind ...**

**Al-Imam Al-Mehdi** (peace be upon him) who will fill the earth with justice and equity, after it has been filled with injustice and oppression.

**The survival of Allah on the earth.**

### **To my father**

**My moonlight, the place I go when there is no one to run to, I'll always cherish the times I spent with you.**

### **To my mother**

**The one and only, she deserves the world for all the support she gave me throughout this journey.**

### **To my sisters**

**You were my candles in the long dark times.**

*Shahad Ewadh*

## **Declaration**

I hereby declare that this dissertation entitled “**Predicting customers' satisfaction with airline companies: The application of machine and deep learning techniques**”, submitted to the University of Babylon in partial fulfillment of requirements for the degree of Master in Information Technology \ Software, has not been submitted as an exercise for a similar degree at any other University. I also certify that this work described here is entirely my own except for experts and summaries whose source is appropriately cited in the references.

Signature:

Name: **Shahad Hussein Ali**

Date: / / 2023

## Supervisor Certification

I certify that the thesis entitled “**Predicting customers' satisfaction with airline companies: The application of machine and deep learning techniques**” was prepared under my supervision at the department of Software / College of Information Technology / the University of Babylon as partial fulfillment of the requirements of the degree of Master in Information Technology - Software.

Signature:

Supervisor Name: **Assist. Prof. Dr. Ahmed Habeeb Said Al-Azawei**

Date: / / 2023

## The Head of the Department Certification

In view of the available recommendations, I forward the thesis entitled “**Predicting customers' satisfaction with airline companies: The application of machine and deep learning techniques**” for debate by the examination committee.

Signature:

**Assist. Prof. Dr. Sura Zaki Alrashid**

Head of Software Department

Date: / / 2023

## **Certification of the Examination Committee**

We, the undersigned, certify that **(Shahad Hussein Ali)** candidate for the degree of Master in Information Technology - Software, has presented his thesis of the following title **(Predicting customers' satisfaction with airline companies: The application of machine and deep learning techniques)** as it appears on the title page and front cover of the thesis that the said thesis is acceptable in form and content and displays a satisfactory knowledge of the field of study as demonstrated by the candidate through an oral examination held on:

Signature:  
Name: Nidaa A. Abbas  
Title: Professor  
Date: / / 2023  
**(Chairman)**

Signature:  
Name: Safa S. Abbas  
Title: Assistant Professor  
Date: / /2023  
**(Member)**

Signature:  
Name: Haider H. Mshali  
Title: Assistant Professor  
Date: / / 2023  
**(Member)**

Signature:  
Name: Ahmed H. Al-Azawei  
Title: Assistant Professor  
Date: / / 2023  
**(Supervisor)**

Approved by the Dean of the College of Information Technology, University of Babylon.

Signature:  
Name: Wesam S. Bhaya  
Title: Professor  
Date: / / 2023  
**(Dean of College of Information Technology)**

## **Acknowledgements**

All praise be to ALLAH Almighty who enabled me to complete this task successfully and my utmost respect to his last Prophet Mohammad PBUH and the hero of Islam born in the heart of Kaaba Amir-Al-Mo'mineen Al-Imam Ali Ibn Abi Talib.

To my supervisor, Dr. Ahmed Habeeb Al-Azawei, many thanks for all advice and comments during my research career. I am grateful for his help to make this research in the best possible form.

Special thanks go to the staff of the College of Information Technology, especially, the staff of Software Department for the great help they have offered me

Last but not least, I would like to thank those who helped me with a word of encouragement or scientific information that benefited me in this thesis: Prof. Dr. Assad Sabah, DR. Tahseen Ali, DR. Ameer Al-Haq Adil and my friends Mais Muhanned and Hiba Hussein

## **Abstract**

Companies compete with each other to provide the best services that could satisfy their customers. Users' satisfaction represents a key aspect that companies aim to achieve as it draws either the success or the failure of a company. However, it is necessary to understand what can lead to customers' satisfaction with a particular company, technology or product. This research aims at (1) predicting customers' satisfaction with airline companies, 2) identifying the most effective features on customers' satisfaction, 3) proposing a modified framework of customers' satisfaction, and 4) enhancing the prediction accuracy using several different techniques.

The dataset of 129880 customers is used in this research. It includes demographic features and airline customers' perceptions. Four data mining techniques are applied to predict customers' satisfaction, namely the Random Forest (RF), K-Nearest Neighbor (KNN), Multi-Layer Perceptron (MLP), and Long Short-Term Memory (LSTM). Many steps are followed to enhance the prediction accuracy of the implemented techniques such as handling missing values and outliers, normalizing the features values, and generating new features. Moreover, feature selection and genetic optimization techniques are applied. To measure the classifiers' performance, three prediction phases are applied which are based on: 1) original twenty-two features of the dataset, 2) twenty-five features which include the original twenty-two and the generated features namely service quality, information quality and engagement, and 3) twelve features which include the three generated features and nine of the original features in order to reduce the model's complexity. The results reveal that from the newly generated features, service quality and engagement are significant predictors of customers' satisfaction, whereas information quality had low impact on customers' satisfaction. In the first prediction scenario, the proposed model uses the original twenty-two features in which the

performance of RF, KNN, MLP, and LSTM was 96%, 78%, 95.2%, and 95.7% respectively. Second, with the original features and the three new generated features, the accuracy of the RF, KNN, MLP, and LSTM techniques was 97%, 96%, 95.2%, and 96% respectively. Finally, based on twelve features only that include the generated features and others that have not been used in the generation process, the model's performance for RF, KNN, MLP, and LSTM techniques was 95%, 93%, 91.2%, and 91.9% respectively. According to such findings, it can be noticed that RF outperformed all other classifiers. It can be concluded that dealing with twelve features can minimize the complexity of the model and result in reasonable prediction accuracy. The research outcomes can help airline companies improve their services and respond to customers' needs.

## Declaration Associated with this Thesis

Some of the works presented in this thesis have been accepted as below.

Title: Enhancing the Prediction of Customers' Satisfaction with Airline Companies Using Data Mining and Genetic Techniques

Author: Shahad Hussein Ewadh and Ahmed Al-Azawei

Journal: Lecture Notes in Networks and Systems

Impact Factor: 0.7

Year: 2023



**Manchester  
Metropolitan  
University**

Centre for Advanced  
Computational Science



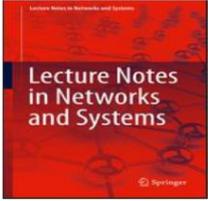
**ICCCNet**



**Springer**

Paper ID: ICCCN<sub>e</sub>t - 202312  
Date: 18-08-2023

### Letter of Acceptance



**Paper Title:** Enhancing the Prediction of Customers' Satisfaction with Airline Companies Using Data Mining and Genetic Techniques

**Author (s):** Shahad Hussein Ewadh, Ahmed Al-Azawei

**Congratulations!**

Based on the recommendations of the Technical Program Committee of (ICCCNet-2023) we are pleased to inform you that your manuscript has been **Accepted as a regular paper** and will be processed for Publication in the Springer Series "**Lecture Notes in Networks and Systems**" [ISSN: 2367-3389; 2367-3370] (**Scopus Indexed**). The paper shall appear in ICCCN<sub>e</sub>t-2023 in Lecture Notes in Networks and Systems.

We will encourage more quality submissions from you and your colleagues in the future.

**TPC Chair /  
Conference Chair / Editor  
ICCCNet-2023-SPRINGER**

General Series Name: Lecture Notes in Networks and Systems  
Series ISSN: 2367-3389; 2367-3370  
<https://www.springer.com/series/15179>

# Table of Content

<b>Chapter One General Introduction</b> .....	<b>1</b>
1.1 Background... ..	1
1.2 Research Problem.....	2
1.3 Research questions .....	2
1.4 Research objectives .....	3
1.5 Related works.....	3
1.6 Thesis organization .....	6
<b>Chapter Two Theoretical Background</b> .....	<b>9</b>
2.1 Overview.....	9
2.2 Customer satisfaction .....	9
2.3 Technology Acceptance Model (TAM) .....	9
2.4 Delone and Mclean model of information systems success (D&M-ISS).....	11
2.5 Dataset.....	13
2.6 Data preprocessing .....	15
2.6.1 Handling missing values .....	15
2.6.2 Handling outliers .....	16
2.6.3 Normalization.....	16
2.7 Feature generation .....	17
2.8 Feature Selection (FS).....	18
2.9 Data mining concepts .....	19
2.10 The Genetic Algorithm.....	25
2.11 Evaluation of the Prediction Models .....	26
2.11.1 The Cross-Validation Method.....	26
2.11.2 Performance Metric.....	27
<b>Chapter Three The Proposed System</b> .....	<b>29</b>
3.1 Overview.....	29
3.2 The Proposed Framework of customers' satisfaction .....	29
3.3 The Proposed System Architecture .....	30
3.3.1 Data preprocessing .....	32
3.3.2 Feature Generation .....	37
3.3.3 Feature Selection .....	38
3.3.4 Optimizer Technique .....	40
3.3.5 The Classification Methods.....	41
3.3.6 Evaluating the Performance of the Proposed Model.....	43

<b>Chapter Four Experimental Results and Discussion .....</b>	<b>44</b>
4.1 Overview.....	44
4.2 Software and Hardware Environment .....	44
4.3 Results of Data Preprocessing.....	44
4.4 Results of Feature Generation .....	49
4.5 The selected features .....	50
4.6 Results of the Classification Methods .....	57
4.6.1 Results of applying machine learning techniques .....	57
4.6.2 Results of applying deep learning techniques .....	66
4.7 Summary.....	72
<b>Chapter Five Conclusions and Future Works.....</b>	<b>73</b>
5.1 Conclusions.....	73
5.2 Theoretical and practical implications .....	73
5.3 Future Work.....	74
<b>References .....</b>	<b>74</b>

## List of Tables

Table No.	Title	Page No.
2.1	The details of the dataset's features	14
2.2	A two-dimensional confusion matrix	28
4.1	Subsets of twenty-five features by using the KNN classifier	50
4.2	Subsets of twenty-five features by using the RF classifier	53
4.3	Subsets of twelve features by using the KNN classifier	55
4.4	Subsets of twelve features by using the RF classifier	56
4.5	A summary of the accuracy with the twenty-two features after handling missing values	58
4.6	A summary of the accuracy and other metrics with the base attributes and the generated features	58
4.7	A comparison of classifiers' accuracy in the case of genetic optimizer and wrapper feature selection techniques based on twenty-five features	61
4.8	A summary of the best hyperparameters by using genetic optimizer with twenty-five features	61
4.9	A summary of the accuracy with twelve features	62
4.10	A comparison of the classifiers' accuracy in the case of wrapper feature selection and the genetic optimizer techniques based on twelve features	65
4.11	A summary of the best hyperparameters by using genetic optimizer with twelve features	66
4.12	A summary of the accuracy of deep learning techniques based on twenty-two features	66
4.13	A summary of the accuracy of deep learning with the twenty-five features	67
4.14	A summary of the accuracy of deep learning techniques based on twelve features	69
4.15	A summary comparison between findings of previous literature and this thesis	70
4.16	The time execution of all cases	71

## List of Figures

Figure No.	Title	Page No.
2.1	The Technology Acceptance Model (TAM)	11
2.2	The updated D&M information systems success (ISS) model	12
2.3	Data mining tasks	20
2.4	The Random Forest Classifier	22
2.5	The structure of Multi-layer perceptron	24
2.6	The LSTM architecture of single neuron	25
3.1	The proposed framework for customers' satisfaction	29
3.2	The Proposed System	31
3.3	The transformation of the dataset from categorical to numerical	34
3.4	The phases of the prediction process	38
3.5	Three-fold cross validation for the airline customers satisfaction dataset	43
4.1	Removing the ID customer feature from a sample dataset	45
4.2	A sample of the dataset after handling the missing values	45
4.3	A sample of the dataset after transformation	46
4.4	A sample of the dataset after unifying the age feature	46
4.5	Outliers before using the capping technique	47
4.6	Handling outliers by using the capping technique	48
4.7	A sample of the dataset after normalization	48
4.8	A sample of the original dataset before the generation of new features	49

Figure No.	Title	Page No.
4.9	A sample of the dataset after the generation of new features	49
4.10	The contribution of features to decision making	57
4.11	The performance accuracy of twenty-five subset of features by using KNN	59
4.12	The performance accuracy of twenty-five subset of features by using RF	60
4.13	The relationship between features	62
4.14	The performance accuracy of twelve features based on RF	64
4.15	The performance accuracy of twelve subsets of features based on KNN	65
4.16	The confusion matrix of MLP and LSTM with original features	67
4.17	The confusion matrix of MLP and LSTM by using twenty-five features	68
4.18	The confusion matrix of MLP and LSTM with twelve features	69

## List of Abbreviations

Abbreviation	Meaning
D&M-ISS	Delone and Mclean's theory of information systems success
DM	Data Mining
FS	Feature Selection
FN	False Negative
FP	False Positive
KNN	K-Nearest Neighbor
LSTM	Long Short-Term Memory
MLP	Multi-Layer Perceptron
RF	Random Forest
SFS	Sequential Feature selection
TAM	Technology Acceptance Model
TN	True Negative
TP	True Positive

## List of Algorithms

Algorithm No.	Title	Page No.
3.1	Data Cleaning	32
3.2	Handling Missing Values	33
3.3	Handling Outliers	35
3.4	Normalization	36
3.5	Feature generation	38
3.6	Wrapper Feature Selection	39
3.7	Genetic optimizer	40
3.8	Random Forest	42

# *Chapter One*

## *General Introduction*

## Chapter One

### General Introduction

#### 1.1 Background

Airline companies are mainly evaluated by the quality of their services according to customers' opinions. Therefore, such companies aim to increase the efficiency and quality of their services to ensure their customers' loyalty and satisfaction [1]. Due to the Covid19 pandemic, airlines around the world were restricted and most flights were stopped. After the pandemic, the demand for air travel abroad has increased [2]. This requires ensuring that users are satisfied with the services provided by such companies.

User satisfaction refers to an overall assessment of the system's user experience, which holds the potential to significantly impact user behavior [3]. To achieve customers' satisfaction, companies should provide a high standard of service quality. If the performance does not reach clients' requirements, quality is considered to be poor, whereas the quality is considered to be high, if it exceeds customers' expectations [4]. A significant relationship between the quality of services of an airline company and the degree of travelers' satisfaction was found in earlier literature [5].

Previous studies used several models to predict the tendency of users to accept or satisfy with a particular technology. The Delone and Mclean's model of information systems success (D&M-ISS) is one of the most used theories to predict technology success [6]. The updated model includes six factors which are system quality, information quality, service quality, system use, user satisfaction and net benefit [7]. Each of these factors has its role in influencing the success of technology [6]. The technology acceptance model (TAM) is another framework used to predict technology adoption which is proposed by Davis in 1986 [8]. This model consists of five factors namely,

perceived ease of use, perceived usefulness, attitudes towards use, behavioral intention and actual usage [9].

This thesis deals with the perceptions of airline customers about the services provided by airline companies. The identification of the most significant services introduced by airline companies may help enhance their markets. This aim is achieved by using a minimum number of features that can impact the perceptions of airline customers and this, in turn, could lead to minimize the complexity of the proposed system.

## **1.2 Research Problem**

Customers' satisfaction is a main concern that organizations aim to achieve. It indicates the success of company. Thus, investigating what can lead to customers' satisfaction is still an open research question that requires further investigation. There is a scarcity of research to identify the most effective features that can influence users' satisfaction with airline companies [1],[10],[11].

## **1.3 Research questions**

This research answers the following questions:

- Is it possible to build a framework to predict users' satisfaction?
- What are the key features that may affect users' satisfaction?
- How can significant features be selected?
- Is it possible to improve the overall classification accuracy?

## 1.4 Research objectives

The key aim of this research is predicting customers' satisfaction with airline companies. To achieve this aim, the following objectives are drawn:

- 1) Proposing an updated framework of customers' satisfaction based on several different features.
- 2) Generating new features based on previous theories of technology acceptance and success that have not been used in previous research.
- 3) Building a classification model to identify the most important features that may have more effect on airline customers' satisfaction.
- 4) Enhancing the accuracy of the classifiers and comparing the research findings with previous literature.

## 1.5 Related works

Several different studies have investigated factors that may contribute to explaining customer satisfaction with airline companies. This section reviews some of the literature that used machine learning in such a process. It is worth mentioning that the majority of literature is based on the original features in the datasets used.

Amalia et.al [1] compared the accuracy of different algorithms using the airline customer satisfaction dataset. The techniques used are Random Forest (RF) classifier, Split Point And Reduced Attribute Classifier (SPAARC) and multilayer perceptron (MLP). All features of the original dataset are used. The findings suggest that the Random Forest (RF) classifier obtained the highest accuracy of 95.83%.

Baswardono et.al [10] used the airline customer satisfaction dataset to compare the efficiency of different algorithms. The Random Forest (RF)

classifier and C4.5 algorithms are used with twenty features after exclusion of two features namely, arrival delay in min and departure delay in min. As in previous studies, the highest accuracy is achieved by the Random Forest (RF) classifier which is 93.32%.

Hoon et.al [11] conducted an airline customer satisfaction analysis by testing the correlation between customer evaluation data. The experiments implemented Random Forest (RF) classifier, deep neural network (DNN) and support vector machine (SVM) techniques. The number of features used in this research is twenty-two with the highest accuracy of 95.7% based on the Random Forest (RF) classifier.

In another direction of predicting user satisfaction in other fields, the technology acceptance model (TAM) and Delone and Mclean theory of information systems success (D&M-ISS) framework are widely used to predict technology success [12],[13]. In [13], the results showed that perceived ease of use had a significant correlation with user satisfaction ( $r = 0.658$ ), whereas perceived usefulness shows a moderate correlation with user satisfaction ( $r = 0.402$ ).

In [14], the impact of leadership and work motivation on job satisfaction is examined. The results of the study show that leadership and work motivation had a positive and significant effect on employee job satisfaction which was represented by path coefficient of 0.263 and 0.171, respectively.

Another research study investigated the degree of client satisfaction with online food delivery services [15]. It was found that information quality, payment system, customer service and quality of food can be considered as predictors of customer satisfaction.

In [16], the influence of food quality and service quality on satisfaction of hot plate restaurants' consumers was investigated. From the output of R-Square, it can be concluded that food quality and service quality had a significant effect on customer satisfaction ( $R^2 = 069.5\%$ ).

In [17], an adaptation of the widely used D&M-ISS model was conducted in the context of hospital information systems in a developing country. It was revealed that system quality, information quality, and service quality significantly influenced user satisfaction based on the values of Beta and probability of significance where both values are ( $\beta = 0.17, p < 0.001$ ), ( $\beta = 0.17, p < 0.001$ ), and ( $\beta = 0.51, p < 0.001$ ), respectively.

In [18], the factors influencing the satisfaction of customer with e-banking services were examined. Four factors that can drive customer satisfaction towards e-banking are cloud services, security, e-learning, and service quality. The results showed that cloud services had a positive and significant influence on customer satisfaction with e-banking services ( $\beta = 0.18, p < 0.001$ ), system security had a positive influencing on customer satisfaction ( $\beta = 0.32, p < 0.001$ ), the e-learning of users contains to use had a positive and significant impact on customer satisfaction with e-banking services ( $\beta = 0.23, p < 0.001$ ), and finally, a positive relationship between service quality and customer satisfaction with e-banking services was provided ( $\beta = 0.25, p < 0.001$ ).

In [19], the influence of service quality on customer satisfaction and customer loyalty was examined in Nepalese restaurants industry. The result revealed that service quality had highest and positive influence on customer satisfaction ( $\beta = 0.593, p < 0.01$ ). Likewise, the result concluded that there is a positive significant impact of service quality toward customer loyalty ( $\beta = 0.645, p < 0.01$ ).

In this thesis, several measures are considered to improve the accuracy of the implemented methods. These include handling missing values and outliers, normalizing the features values, and generating new features. Moreover, feature selection and genetic optimization techniques are applied. To assess the effectiveness of the classifiers, three predictive phases are executed. These phases involve:

1. The original set of twenty-two dataset attributes.
2. A set of twenty-five attributes, encompassing the original twenty-two and the newly generated attributes which are service quality, information quality, and engagement.
3. A subset of twelve attributes attained through the generated features and others that have not been used in the generation process, aiming at reducing the proposed model's complexity.

## 1.6 Thesis organization

After Chapter One, which presents a general introduction, the rest of the thesis is structured as follows:

- **Chapter Two (Theoretical Background):** This chapter presents a general review of the main concepts of data mining, preprocessing techniques, methods of generating new features, feature selection, machine learning algorithms, deep learning algorithms, model evaluation, and prediction methods.
- **Chapter Three (The Proposed System):** This chapter shows the preprocessing steps, features generation, and features identification. Following this, the prediction models are created to predict customer's satisfaction on airline companies.

- **Chapter Four (The Results and Discussion):** This chapter illustrates and discusses the results obtained after implementing the proposed model.
- **Chapter Five (Conclusions and Future Work):** This chapter summarize the key concepts and findings of this thesis and gives suggestions for future work.

# ***Chapter Two***

## ***Theoretical Background***

## **Chapter Two**

### **Theoretical Background**

#### **2.1 Overview**

This Chapter explains the basic concepts related to the theoretical background of customer satisfaction, the technology acceptance model (TAM) and Delone and Mclean's model of information systems success (D&M-ISS). It, then, moves forward to review the main concepts of data preprocessing, feature generation, feature selection, and data mining techniques. It is noteworthy to mention this chapter mainly focuses on the key techniques and methods adopted in this thesis.

#### **2.2 Customer satisfaction**

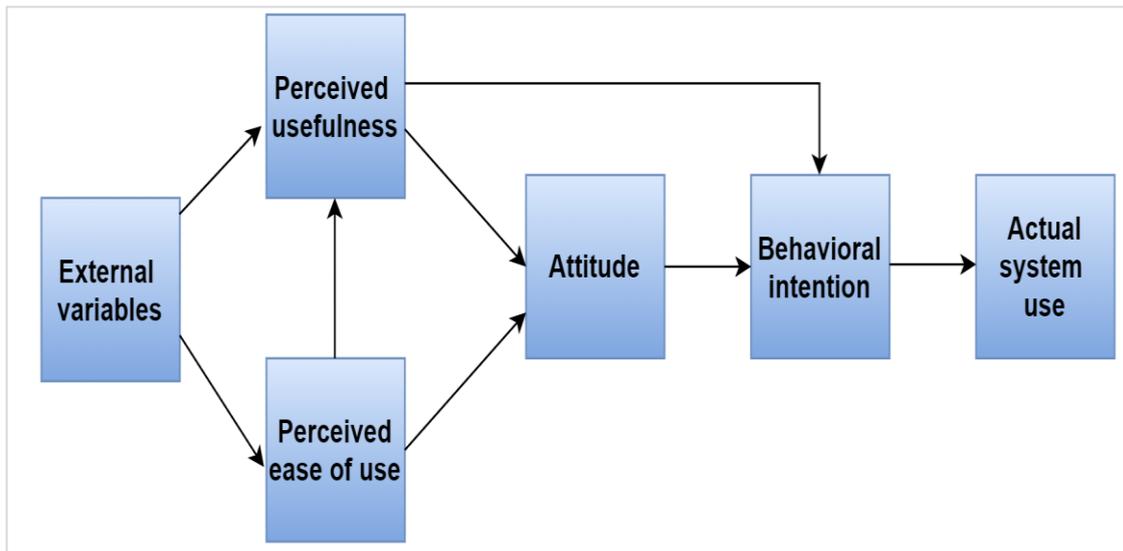
Customer satisfaction can be defined as the degree of contentment experienced by customers with a product or service, which arises from comparing their initial expectations with the actual performance and outcomes they perceive after making a purchase or consuming the offering [20]. Therefore, marketers of products and services attach great significance to satisfaction, as a satisfied customer tends to exhibit positive behaviors such as repeat purchases, positive word-of-mouth recommendations, and long-term loyalty [20]. Customer satisfaction plays a crucial role in fostering customer loyalty, which in turn, contributes to positive growth in market share. Furthermore, customer satisfaction can also serve as a predictor of future market penetration for an airline company [21].

#### **2.3 Technology Acceptance Model (TAM)**

In 1986, Davis proposed the Technology Acceptance Model (TAM) as a way to predict and evaluate users' acceptance of technology. Five factors are suggested in this model which are perceived usefulness, perceived ease of use,

attitude towards use, behavioral intention, and actual system use.

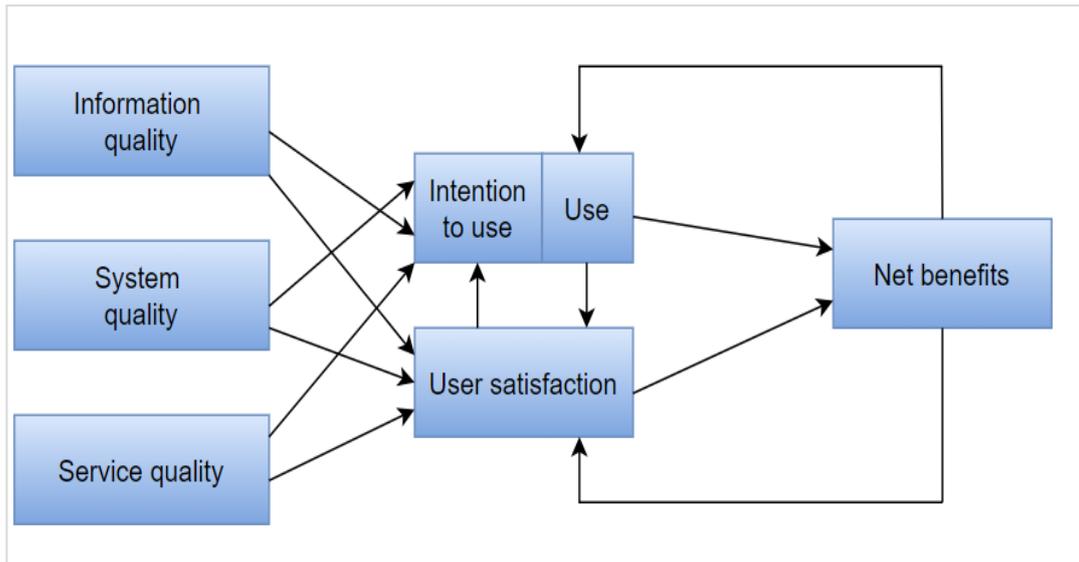
- 1- Perceived usefulness is defined as “ the degree to which an individual believes that using a specific system will enhance their performance” [8].
- 2- Perceived ease of use refers to “the degree to which an individual believes that using a particular system will require minimal physical and mental effort” [8].
- 3- Attitude is defined as a consumer's positive or negative feelings towards a particular product or service [9].
- 4- Behavioral intention represents “a measure of the strength of one's intention to perform a specified behavior” [8].
- 5- Actual use constitutes a response or action in relation to the effective utilization of technology [22]. This signifies the ultimate stage at which individuals employ the technology. Figure 2.1 illustrates TAM.



**Figure 2.1: The Technology Acceptance Model (TAM) [23]**

## **2.4 Delone and Mclean model of information systems success (D&M-ISS)**

The initial proposal of the D&M ISS model was in 1992 [24]. This model suggests a connection between six concepts: information quality, system quality, IS use, perceived satisfaction, user impact, and organization impact. In 2003, Delone and Mclean made an update to their model by introducing a new element called "service quality" [7]. The effectiveness of this new model has been confirmed through successful validation using various web-based technologies such as bank [25] and business-intelligence systems [26]. Figure 2.2 depicts the updated D&M ISS model.



**Figure 2.2: The updated D&M information systems success (ISS) model [7]**

The core categories identified in the D&M model are explained as follows:

- Information quality is defined as a process of "Measuring the information system output" [24]. It is assessed by looking at the system's production of information in terms of reliability, timeliness, accuracy and trustworthiness [27]. Information quality is the extent to which information possesses the content, form, and timing characteristics that make it helpful to specific end users. Previous literature found that information quality had a strong effect on customer satisfaction [15].
- System quality can be identified based on reliability, accessibility, ease of use, and functionality [24]. System quality pertains to the calibre of processing within an information system, encompassing both its software and data constituents [28]. It is found that system quality had a significant impact on user satisfaction [29].
- Service quality is one of the dimensions included in the updated D&M ISS model. Service quality is a subset of satisfaction, which is a long-term overall evaluation of a particular product or service [30]. It refers

to the variance between what a customer anticipates or hopes for and his/her actual perception of the service that is received [31]. It is anticipated that customers will evaluate the quality of service as "low" if the service performance falls under their expectations, while they will consider the quality as "high" if the performance exceeds their expectations [31]. The seven service quality dimensions are tangibility, reliability, responsiveness, assurance, empathy, complaint handling and convenience [32]. Based on experiments conducted by other researchers, service quality is a determinant of customer satisfaction [31], [33].

## **2.5 Dataset**

This thesis uses the airline passengers' satisfaction dataset released in 2019. This dataset can be found on the Kaggle website at <https://www.kaggle.com>. In the dataset, there are 129.880 rows in which each row contains general information about passengers in addition to a range of the rating of services provided. Table 1 shows all the original features of the dataset. The dataset contains twenty-two features, and one class labelled as satisfied and dissatisfied due to customers perception in respect to the corresponding feature. The main purpose of this dataset is to predict whether customers are satisfied with the services provided or not. This topic is important for airline companies, as it focuses on the services that could lead to customers' satisfaction.

**Table 2.1: The details of the dataset's features**

NO	Feature	Description
1	Gender	Gender of passengers (Female, Male).
2	Customer type	The customer type (Loyal customer, disloyal customer).
3	Age	The actual age of the passengers.
4	Type of travel	Purpose of the flight of the passengers (Personal Travel, Business Travel).
5	Class	Travel class of the passengers in the plane (Business Eco, Eco Plus)
6	Distance of travel	The flight distance of this journey
7	Inflight Wi-Fi service	Satisfaction level of the inflight WIFI service
8	Departure/Arrival time convenient	Satisfaction level of Departure/ Arrival time convenient
9	Ease of online booking	Satisfaction level of online booking
10	Gate location	Satisfaction level of Gate location
11	Food and drink	Satisfaction level of food and drink
12	Online boarding	Satisfaction level of Online boarding.
13	Seat comfort	Satisfaction level of seat comfort
14	Inflight entertainment	Satisfaction level of inflight entertainment
15	On-board service	Satisfaction level of on-board service
16	Legroom service	Satisfaction level of legroom service
17	Bagging handling	Satisfaction level of bagging handling
18	Check-in service	Satisfaction level of check-in service
19	Inflight service	Satisfaction level of inflight service
20	Cleanliness	Satisfaction level of cleanliness
21	Departure delay in min.	Minutes delayed when departure

NO	Feature	Description
22	Arrival delay in min	Minutes delayed when the arrival
23	Satisfaction	Airline satisfaction level (satisfaction, neutral or dissatisfaction)

## 2.6 Data preprocessing

Data preprocessing techniques play a crucial role in preparing datasets. Generally, these techniques can be classified into two main categories: selecting data objects and attributes for analysis or creating and modifying attributes [34]. These methods encompass various strategies for addressing common dataset issues such as noise, missing values, and inconsistent data.

Regardless of the specific techniques employed, the objectives of data preprocessing are as follows [35]:

1. Reducing the size of the dataset to enhance the efficiency of analysis in terms of time, cost, and quality.
2. Adapting the dataset to be aligned with the selected analysis method for optimal results.

### 2.6.1 Handling missing values

In many cases, some data objects may have missing attribute values, and these missing values can have a negative impact on the performance of prediction models. Several strategies for handling missing values are discussed in [34], including the following:

- Ignoring missing values during analysis: This approach involves excluding any data objects with missing values from the analysis. However, this can lead to a reduction in the dataset size and potentially overlook valuable information.

- Eliminating data objects: In this strategy, data objects with missing values are completely removed from the dataset. While this eliminates the issue of missing values, it can result in a loss of valuable data and potentially bias the analysis.
- Estimating missing values: This strategy involves estimating the missing values based on the available data. If the attribute with missing values has a small number of missing values that are widely scattered, an estimation method can be used. One approach is to evaluate the missing values using residual values. For categorical predictors, the most frequently occurring predictor value can be used. For continuous predictors, the average predictor value can be utilized [34]. These strategies provide different options for handling missing values, and the choice of strategy depends on the specific characteristics of the dataset and the analysis goals.

### **2.6.2 Handling outliers**

Anomaly detection refers to the process of identifying observations that exhibit significant deviations or differences from the majority of the data. These observations, often referred to as anomalies or outliers, possess characteristics that are distinct from the normal patterns in the dataset. The objective of an anomaly detection algorithm is to accurately identify genuine anomalies while minimizing the occurrence of false positives, where normal objects are mistakenly labeled as anomalous [34].

### **2.6.3 Normalization**

Once data is collected from various sources, it is commonly stored in a data warehouse. However, integrating this data into a single table for preparation in modeling algorithms can present several challenges. One of the challenges is inconsistency. If this issue is addressed, both model accuracy and the

efficiency of the mining process can be enhanced [36]. Its purpose is to ensure that all features within the dataset are represented in the same unit of measurement [34]. Thereby, normalization helps in eliminating discrepancies arising from the varying scales of different features, preventing the dominance of large values over smaller ones, and ensuring a fair comparison among them. Various methods exist for data normalization, including min-max normalization and z-score normalization [34]. In min-max normalization, Equation 2.1 is used to calculate the value [34].

$$X' = \frac{x - \min_x}{\max_x - \min_x} \quad (2.1)$$

where  $X'$  represents the normalized value corresponding to each feature.  $x$  is the original value for a feature,  $\min_x$  refers to the minimum original value for a feature and  $\max_x$  represents the maximum original value for a feature.

## 2.7 Feature generation

Feature generation involves creating new predictors or features from the original ones to capture essential information more effectively [34]. Typically, the number of new predictors is smaller than the number of base features. This allows us to influence the benefits of dimension reduction, such as avoiding the curse of dimensionality, eliminating irrelevant features and reducing noise, creating a more interpretable model, and facilitating data visualization [34]. The following are the most popular methods used for feature generation [34]:

1. Mapping the data to a new space: This approach involves transforming the data into a different space using techniques such as dimensionality reduction or kernel methods.
2. Feature extraction: Feature extraction refers to the process of creating a new set of features from the raw data. It involves transforming the raw data into a higher-level representation that is

more suitable for various types of classification algorithms. An example of feature extraction is extracting edges from images, where the edges serve as informative features.

3. Feature construction: Sometimes, the raw dataset may contain features that possess the necessary information but are not in a suitable form for the modeling algorithm. In such cases, feature construction is employed to address this issue. Feature construction involves creating one or more new features that may be more useful than the original features. For example, dividing the mass by the volume to obtain the density is a form of feature construction. These methods of feature generation play a vital role in enhancing the representation and effectiveness of the dataset for modeling and analysis purposes.

## **2.8 Feature Selection (FS)**

Feature selection (FS) involves the process of identifying the most relevant and influential set of features from the original feature set for a given problem. It plays a crucial role in various aspects such as reducing computational requirements, improving the performance of predictors, enhancing data understanding, and mitigating the challenges posed by the curse of dimensionality [37]. Features can generally be categorized as strongly relevant, weakly relevant, or irrelevant. The primary goal of FS techniques is to identify the strongly relevant features that are most informative for the target class [35]. By selecting these relevant features, the prediction model can achieve improved accuracy while reducing model complexity. FS methods can be classified into different categories, including filter, wrapper and embedded techniques [34]. In the wrapper method, various subsets of features are selected from the search space using a specific search strategy. Wrapper

methods can be further classified into two categories: heuristic search algorithms and sequential selection algorithms. Sequential selection algorithms are so named because of their step-by-step approach to feature addition. The SFS algorithm begins with an empty feature set, in each step, individual features are added to the current subset. The feature that results in the maximum classification accuracy is selected and added to the subset. Thus, one feature is added at a time, and a new subset is formed at each step. This process continues until the desired number of features has been added, and the best accuracy is achieved [37]. The SFS algorithm iteratively evaluates the impact of adding each feature, selecting the most informative one at each step. By gradually building the feature subset, it aims to identify the optimal set of features that maximizes the classification accuracy.

## 2.9 Data mining concepts

Because of the growth of the size of data, researchers need sophisticated methods to analyze it. Therefore, the data mining field was found. Data mining is the process of extracting unseen information and discovering the unrecognized associations between data items along with the hidden patterns in huge data [34]. Data mining techniques have achieved great success in many fields such as business, markets, etc. In general, data mining tasks are divided into two types as shown in Figure 2.3:

- Predictive tasks: The primary aim of these tasks is to anticipate the value of a specific attribute by considering the values of other attributes. The attribute that is to be predicted is often referred to as the target or dependent variable, whereas the attributes utilized for making the prediction are referred to as the explanatory or independent variables [34].
- Descriptive tasks: In this context, the objective is to uncover patterns

such as correlations, trends, clusters, trajectories, and anomalies that provide a summary of the underlying relationships within the data. Descriptive data mining tasks are typically exploratory in nature and often necessitate the use of post-processing techniques to validate and explain the outcomes [34].

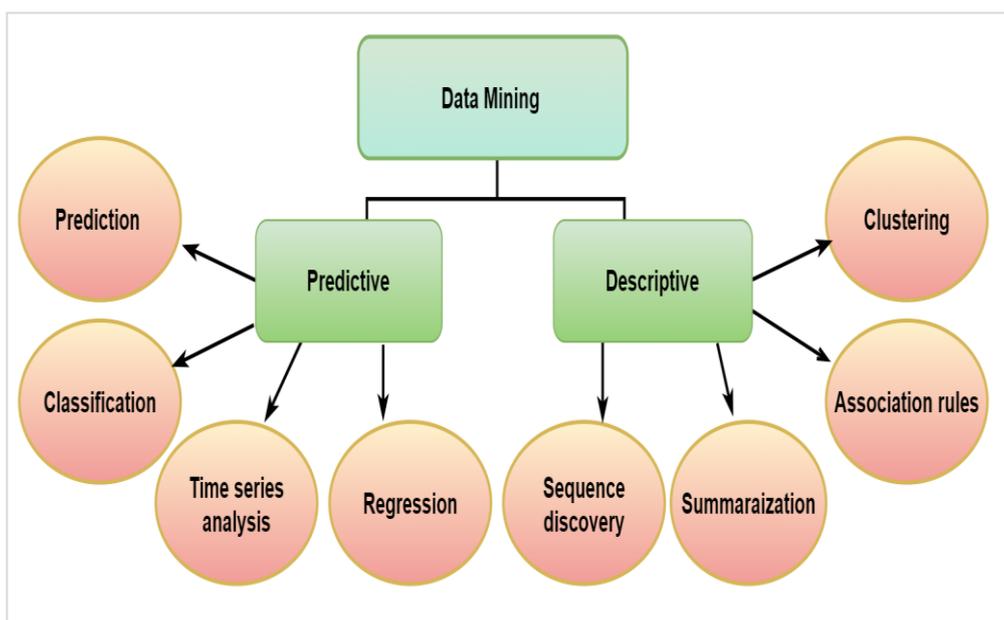


Figure 2.3: Data mining tasks [38]

### • Classification Techniques

The main focus of this thesis is to classify customers' satisfaction as either "satisfy" or "dissatisfy." Since the target class values are discrete, classification techniques are well-suited for this purpose [34]. While numerous machine learning and deep learning algorithms can be used for such predictions with a set of features, the chosen approach revolves around a specific data mining algorithm. Machine learning models can be classified into three major categories based on the nature of the input data provided to the machine learning algorithms. These three categories are:

- Supervised Learning

In supervised learning, the algorithm is trained on a labeled dataset, meaning that the input data is accompanied by corresponding target labels or desired outputs [39].

- Unsupervised Learning

In unsupervised learning, the algorithm is given input data without any explicit target labels [39].

- Semi-Supervised Learning

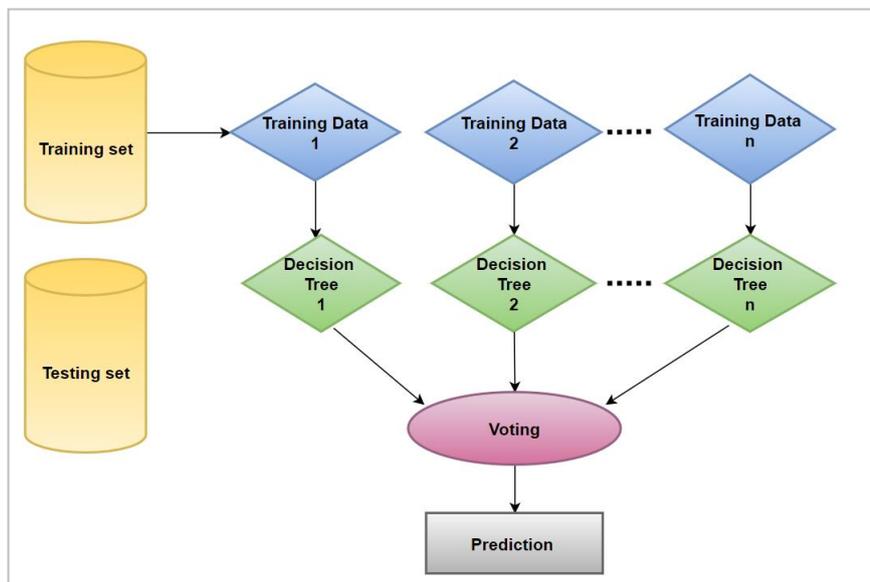
It is a combination of both supervised and unsupervised learning. In this approach, the algorithm is provided with a dataset containing both labeled and unlabeled data [39].

It is intended to use supervised learning model in this thesis as the input data used is accompanied by corresponding target labels. Four algorithms used in this thesis, are explained as follows:

#### **A. Random Forest Classifier (RF)**

Random Forest is a machine learning algorithm that widely uses an ensemble technique for predicting and estimating probabilities. Ensemble methods aim to improve the accuracy of classification by combining multiple classifiers. Random Forest achieves this by creating multiple decision trees where each tree is trained on a subset of data and a subset of features [40]. The predictions from all the trees are then combined to form a final prediction which is often more accurate than the predictions of any single decision tree. The prediction process involves aggregating the predictions made by individual basic classifiers in a voting manner for classification tasks or by taking the average predictions for regression tasks. In classification, multiple basic classifiers are employed to make predictions on the target class. The

predictions from each classifier are collected, and a voting mechanism is used to determine the final prediction. This can be done by selecting the class with the highest number of votes or using more complex methods such as weighted voting. In regression, the predictions from each individual model are averaged to obtain the final prediction. This averaging process takes into account the predictions made by each model and calculates their mean or weighted average, depending on the specific requirements or considerations by combining the predictions of multiple models through voting or averaging, the overall prediction tends to be more robust and accurate, benefiting from the collective wisdom of the individual models. This ensemble approach is a popular technique in machine learning to improve prediction performance. The Random Forest classifier is a popular choice because it is relatively easy to use, can handle large datasets with high-dimensional feature spaces, and is one of the most reliable models for classification [41]. Figure 2.4 illustrates the main notion behind this algorithm.



**Figure 2.4: The Random Forest Classifier [34]**

## B. The K-Nearest Neighbor (KNN) Classifier

The K-Nearest Neighbor (KNN) classifier is a lazy learning method, used to solve classification and regression issues. The KNN algorithm utilizes a parameter called "k" to determine the number of neighbors to select. The choice of k greatly affects the performance of the algorithm in making accurate diagnoses [42]. The input for both classification and regression problems relies on the k-closest training examples in the feature space. The output determines whether the K-Nearest Neighbor (KNN) classifier is applied for regression or categorization [41]. KNN function uses the Euclidean distance as a default and this can be computed using Equation 2.2.

$$D(x, y) = \sqrt{(x_1 - y_1)^2 + \dots + (x_n - y_n)^2} \quad (2.2)$$

where x and y are subjects to be compared with n characteristics [42].

## C. Multiple Layer Perceptron (MLP)

The Multilayer Perceptron (MLP) is a type of artificial neural network that includes one or more hidden layers [43]. Each layer consists of neurons with nonlinear activation functions, typically using the sigmoid function, SoftMax, etc. Figure 2.5 shows the structure of the MLP, which includes an input layer, one or more hidden layers, and an output layer. The nodes in each layer are interconnected in a feed-forward manner, meaning the information flows from the input layer through the hidden layers to the output layer. The connectivity between layers in the MLP is determined by the connection weights. During the initialization of the network, these weights are assigned small random values within the range of -1 to 1 [43]. These random initializations help to introduce diversity and allow the network to learn different features and patterns. The training process of the MLP continues until a specific termination criterion is met. This criterion is typically based on the error of

the network's predictions. For example, training might stop when the error falls below a predefined threshold value, such as 0.01. The error is calculated by comparing the network's output with the desired output for a given set of input data. The training algorithm adjusts the connection weights iteratively using techniques such as backpropagation to minimize this error and improve the network's performance in making accurate predictions [43].

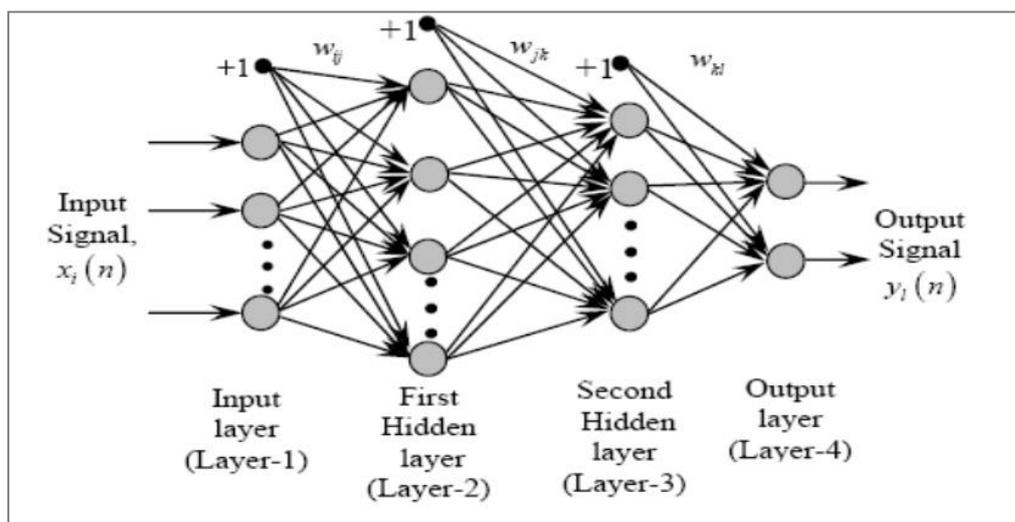


Figure 2.5: The structure of Multi - Layer Perceptron [43]

#### D. Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) is indeed a type of recurrent neural network (RNN) that is specifically designed to capture long-term dependencies in sequential data. The primary objective of constructing LSTM networks is to overcome the vanishing gradient issue commonly encountered in regular RNNs. This challenge arises when the network faces difficulty in acquiring and preserving information from distant time steps in the past, mainly because of the exponential decay that occurs during backpropagation as a result of repeated gradient multiplication. In order to address this problem, LSTM introduces memory cells and gates that regulate the information flow within the network. The essential components include the memory cell, input

gate, forget gate, and output gate in an LSTM unit [44]. Figure 2.6 shows the key concept of LSTM model.

- The memory cell is a storage component that allows the network to retain information across longer sequences.
- Input Gate: determines how much fresh data is stored in the memory cell.
- Output Gate: controls how much information from the current cell state should be output to the next time step.
- Forget Gate: determines how much information from the previous cell state should be discarded or forgotten.

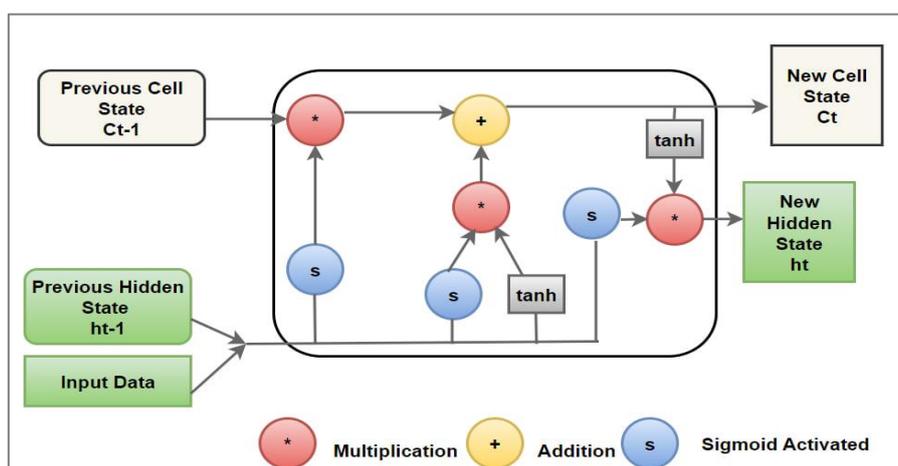


Figure 2.6: The LSTM architecture of single neuron [39]

## 2.10 The Genetic Algorithm

The genetic algorithm is one of the most popular machine-learning techniques for resolving efficiency issues [40]. Genetic algorithms use populations of different solutions to navigate and pass on the best solutions to the following generation and converge on overall solutions [45]. A generation is comprised of a group of people, and each person within that group represents a potential solution in a search space. Each person can be described

as a string of characters, integers, floats, or bits, which is similar to a chromosome [45]. Genetic algorithms mimic the mechanism of natural selection. Such algorithms can adapt to changes in their environment to survive, reproduce, and pass on their genes to the next generation. This iterative process continues until a satisfactory solution is found [45]. This study integrates genetic optimization into the two algorithms of machine learning to select the optimal hyperparameters. Finding the best value of hyperparameters can significantly affect prediction accuracy.

## **2.11 Evaluation of the Prediction Models**

A prediction model is assessed using a variety of resampling techniques, including the Holdout Method, Cross-Validation, and others. The Cross-Validation method is described here in the context of how it is used in this thesis.

### **2.11.1 The Cross-Validation Method**

In this method, each example is used  $K$  times for training and once for testing, where  $K$  is greater than or equal to one. There are three cross-validation methods, which are as follows [34]:

1. **K-fold cross-validation:** This technique splits the dataset into  $K$  equal-sized subsets. There are  $K$  stages to this strategy. All subsets in each stage are used for training, except for one subset that is used for testing. This approach is used, where each partition is tested exactly once.
2. **Leave-one-out cross-validation:** This is a specialized form of  $k$ -fold cross-validation sets in which the dataset is divided into  $K$  portions of equal size, where  $k=N$  denotes the dataset's instance count [34].
3. **Two-fold cross-validation:** The dataset is split into two equal parts in this method. Both training and testing are conducted on the first partition. The partitions' roles are then swapped, making the former

training partition the testing partition and vice versa. Each example is used exactly once for testing and once for training in this method.

### **2.11.2 Performance Metric**

The performance metrics are used to assess the trained model's generalization power and quality when tested on new data. Different metrics can be employed in classification models to assess the effectiveness of a specific classification algorithm. This includes recall, precision, F1-measure, and accuracy. The accuracy of a model is one of the most frequently used metrics to assess its generalizability [46]. The trained model's accuracy is measured by the total number of instances that it correctly predicts when tested against unknown data. To maximize accuracy performance, imbalanced class problems are handled using the recall, precision, and F1-measure metrics [46]. These measurements are computed based on the confusion matrix. This matrix summarizes the number of instances wrongly or rightly predicted by a classification algorithm as shown in Table 2.2 [34].

1. True Positive (TP): it is the true positive actually positive and predict positive.
2. False Negative (FN): it is the false negative actually negative and it is predicted negative.
3. False Positive (FP): it is the false positive actually negative and it is predicted positive.
4. True Negative (TN): it is the true negative actually positive and it is predicted negative.

**Table 2.2: A two-dimensional confusion matrix**

<b>Predict</b> <b>Actual</b>	<b>Positive</b>	<b>Negative</b>
<b>Positive</b>	<b>TP</b>	<b>FN</b>
<b>Negative</b>	<b>FP</b>	<b>TN</b>

The accuracy measure is computed based on the number of correct predictions divided by the total number of predictions as presented in Equation 2.3 [34].

$$\text{Accuracy} = (TP + TN) / TP + TN + \sum FP + \sum FN \quad (2.3)$$

Recall is the number of TP divided by the number of TP and the number of FN. This metric can be computed using Equation 2.4 [34].

$$\text{Recall} = TP / TP + \sum FN \quad (2.4)$$

Precision is the number of TP divided by the number of TP and FP. It can be computed based on Equation 2.5 [34].

$$\text{Precision} = TP / TP + \sum FP \quad (2.5)$$

F1-measure is the  $2 * ((\text{precision} * \text{recall}) / (\text{precision} + \text{recall}))$ . This measure can be computed based on Equation 2.6 [34].

$$\text{F1-measure} = (2 * TP) / (2 * TP + \sum FN + \sum FP) \quad (2.6)$$

# ***Chapter Three***

## ***The Proposed System***

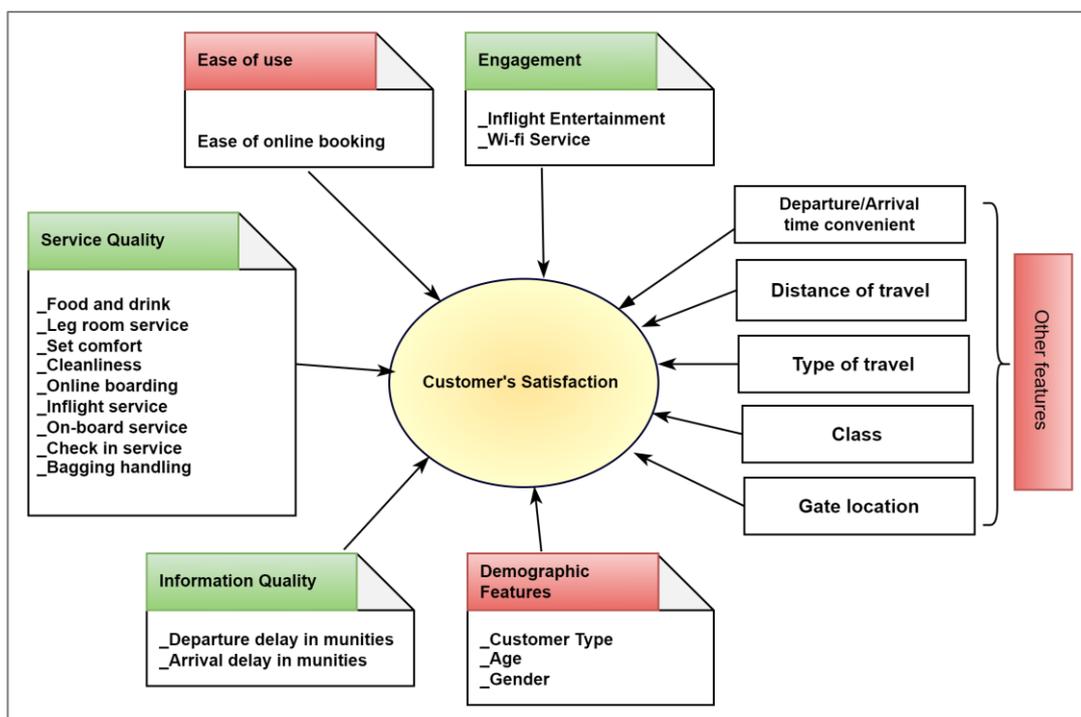
## Chapter Three The Proposed System

### 3.1 Overview

In this Chapter, the steps followed to achieve the key aim of this thesis are described. This includes proposing a framework for customers' satisfaction and the system of predicting that satisfaction. The architecture of this system is depicted first. Then, data preprocessing, feature generation, feature selection, the prediction model, and evaluation strategies are covered.

### 3.2 The Proposed Framework of customers' satisfaction

In this research, a framework is proposed to predict customers' satisfaction as shown in Figure 3.1. It includes six key categories which are, engagement, ease of use, service quality, information quality, demographic, departure\arrival time convenient, distance of travel, type of travel, class and gate location. These categories are explained as follows:



**Figure 3.1: The proposed framework of customers' satisfaction**

### 3.3 The Proposed System Architecture

The proposed system architecture comprises five phases: data preprocessing, feature generation, feature selection, the classification model, and evaluation of the research findings. Each stage consists of sub-steps to address the research problem to answer the research questions and to achieve the research aim and objectives that are discussed in Chapter One.

Figure 3.2 demonstrates a general overview of these phases. The next subsections provide further details. First, the data preprocessing phase includes data cleaning, handling missing values, handling outliers and normalization. Second, feature generation involves of use feature construction. Three new features are constructed. The third stage involves the use of the wrapper method. This method helps reduce the number of features and identify the most important features used in the classification stage. In the final phase, classification techniques are implemented. This comprises the application of Random Forest (RF), K-Nearest Neighbor (KNN), Multilayer Perceptron (MLP), and Long Short-Term Memory (LSTM).

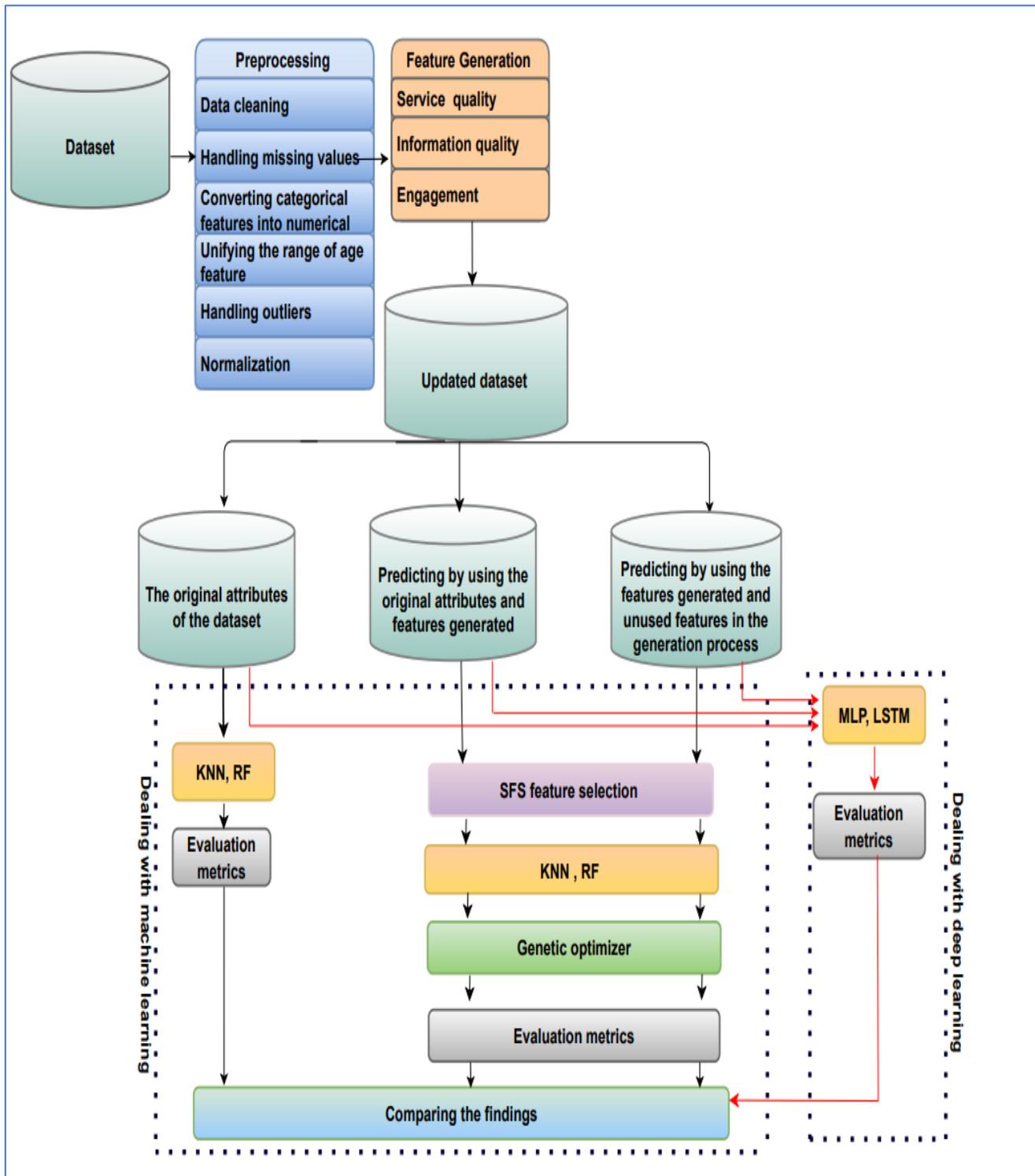


Figure 3.2: The proposed system

### 3.3.1 Data preprocessing

To make the dataset suitable for machine learning and deep learning techniques (prediction models), various preprocessing steps are implemented. Specifically, this procedure encompasses four distinct stages:

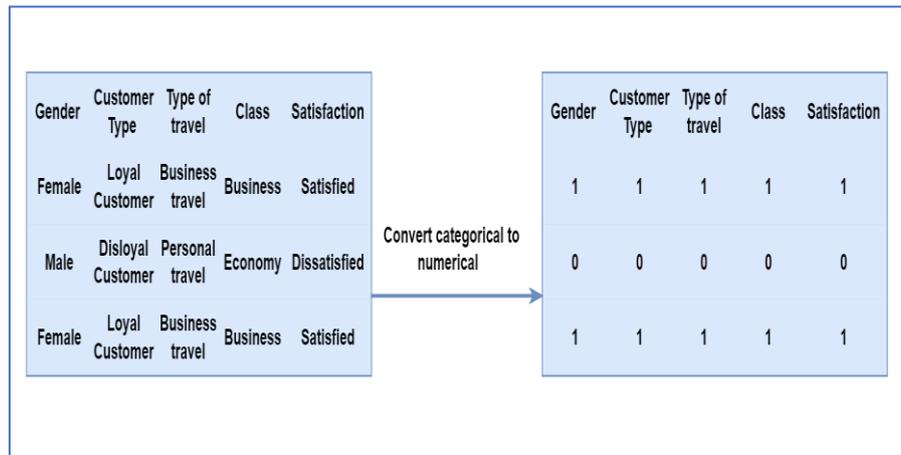
- A. Data Cleaning: In this step, features that were not useful in the prediction process such as ID number is removed. Algorithm 3.1 describe the data cleaning.

<b>Algorithm 3.1: Data Cleaning</b>
<b>Input:</b> Two-dimensional array $D1[n*m]$ where $m$ is the number of features and $n$ is the number of instances
<b>Output:</b> Two-dimensional array $D1[n*m]$ after cleaning the data.
<b>Begin:</b>
For $i = 0$ to $n$
Remove ( $D1[i,0]$ )
end for $i$
<b>End</b>

- B. Handling missing values: In the used dataset, there are some missing values of 393 rows that are observed in the column of arrival delay in minutes. These missing values are replaced with the mode value. Algorithm 3.2 describe the handling of missing values.

<b>Algorithm 3.2: Handling Missing Values</b>
<b>Input:</b> Two-dimensional array $D1[n*m]$ where $m$ is the number of features and $n$ is the number of instances
<b>Output:</b> Two-dimensional array $D1[n*m]$ after handling missing values.
<b>Begin:</b>
For $i = 1$ to $m$
For $j = 1$ to $n$
if $D1 [j][i]$ in feature $i$ is missing then
$D1 [j][i] = M =$ The most repeated value in feature $i$
end if
end for $j$
end for $i$
<b>End</b>

C- Converting categorical feature into a numerical feature: After handling missing values the next step transforming categorical features into numerical features. This step is applied to make the proposed system capable to deal with all features in the dataset. Figure 3.3 describes the idea of the transformation of the dataset.



**Figure 3.3: Transformation of the dataset from categorical to numerical**

**D - Unifying age feature:** The age feature was coded by sorting age into eight groups which are (7-16 years: coded as 1), (17-26 years: coded as 2), (27-36 years: coded as 3), (37-46 years: coded as 4), (47-56 years: coded as 5), (57-66 years: coded as 6), (67-76 years: coded as 7), and finally (77 years and more: coded as 8).

**E- Handling Outliers:** The preprocessing includes identifying outliers. It is found that five features had outliers which are customer type, flight distance, check-in service, departure delay in minutes, and arrival delay in minutes. The generic steps of handling outliers are described in Algorithm 3.3.

**Algorithm 3.3: Handling Outliers**

**Input :** Two- dimensional array  $D1[n*m]$  where  $m$  is the number of features and  $n$  is the number of instances

**Output :** Two-dimensional array  $D1[n*m]$  after handling outliers

**Begin:**

- 1- Define the threshold of outliers detection (e.g= 1.5)
- 2- Calculate the 25<sup>th</sup> and 75<sup>th</sup> percentile for each column of the dataset
- 3- Calculate the interquartile range (IQR) for each column
- 4- Define the lower and upper bound for each column based on IQR and threshold
- 5- Replace the value of the fall outside of the boulder with the corresponding bound value
- 6- The result should be free from outliers

**End**

F- Normalization: The purpose of this step is to prevent features with high values from heavily influencing the calculation results. Therefore, all feature values are transformed through min-max normalization, with a scale of 0 to 1. Algorithm 3.4 describe the normalization of dataset.

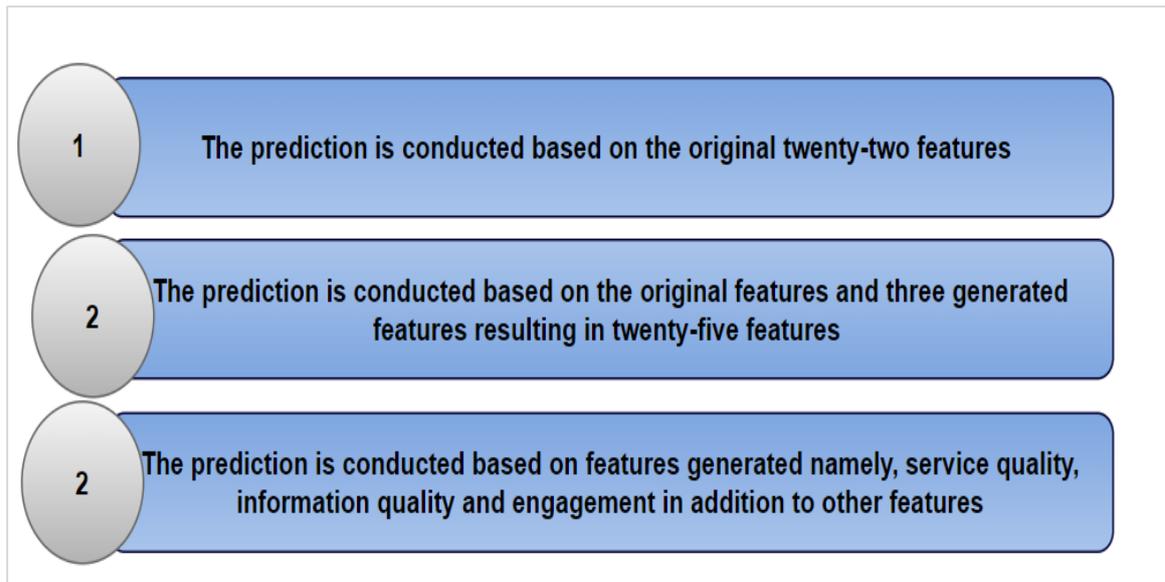
<b>Algorithm 3.4: Normalization</b>
<b>Input:</b> Two-dimensional array D1[n*m] where n is the number of instances, m is the number of features and v represents the first value of each feature in D1
<b>Output:</b> Two-dimensional array D1[n*m] after applying normalization
<b>Begin:</b>
Let Min[j] and Max [j] two arrays holding (minimum and maximum value)
For j= 1 to m
Set min and max in the first value of feature j
For i= 1 to n
If v <sub>ji</sub> < min then
min= v <sub>ji</sub>
else if v <sub>ji</sub> > max then
max= v <sub>ji</sub>
End if
End if
modify each value v in feature j according to the equation:
$v^{\sim} = \frac{v - \min}{\max - \min}$
End for i
End for j
<b>End</b>

### 3.3.2 Feature Generation

Based on the original features, three new features are created namely as; service quality, information quality, and engagement. New features are generated in this study to enhance the accuracy of the prediction process. The generated features are demonstrated as follows:

- **Service quality:** This feature is created based on its definition and the D&M-ISS model. Service quality is generated using nine features in the original dataset. This includes the food & drink attribute, legroom service, set comfort, cleanliness, online boarding, inflight service, on-board service, check-in service, and bagging handling. The mean of score measures in the original dataset is used to represent service quality.
- **Information quality:** This feature is generated based on departure delay in minutes and arrival delay in minutes. This generation relies on the key concepts of this feature as identified in the D&M-ISS model. The mean score measure of these two features is used to represent information quality.
- **Engagement:** This attribute reflects the level of entertainment offered to customers. As such it is calculated by using mean from inflight entertainment and Wi-Fi service.

In this thesis, the prediction of the airline customer satisfaction process involves three distinct phases. The initial phase utilizes the original features exclusively without incorporating any new ones. In the second phase, the original features are combined with additional features that have been generated. Finally, the third phase involves the integration of the newly created features with the features that were not used. Figure 3.4 describes the phases of the prediction process. Algorithm 3.5 describes the generic steps of feature generation.



**Figure 3.4: The phases of prediction**

**Algorithm 3.5: feature generation**

**Input:** two-dimensional array  $D1[n*m]$  where  $n$  is the number of instances and  $m$  is the number of features

**Output:** two-dimensional array  $D1[n*m]$  after generation new features

**Begin:**

1. Initialize a variable sum to 0
2. Repeat the following steps for each feature:
  - a. Input the value of the feature
  - b. Add the feature value to the sum
3. Calculate the mean by dividing the sum by the number of features

**End**

### 3.3.3 Feature Selection

The feature selection method is applied on the Airline customer satisfaction dataset to reduce the dimensionality of feature space, select the important features for the prediction process, and enhance the prediction accuracy. The output of this step is to select a subset of the most important features. The traditional feature selection methods are applied. Wrapper methods choose features' subsets based on the quality performance of the

prediction model. Algorithm 3.6 describes the wrapper feature selection technique.

<b>Algorithm 3.6: Wrapper Feature Selection</b>	
<b>Input:</b> Two dimensional array $D1(n,m)$ where $n$ : number of instance , $m$ : number of features	
<b>Output:</b> Significant features	
<b>Begin:</b> Set subset as empty { }	
Repeat For $d= 1$ to $m$ Subset = {subset, features} Evaluate the model using (3 cross-validation) End for until evaluating all subsets	
Identifying the most significant features depend on the performance	
<b>End</b>	

### 3.3.4 Optimizer Technique

During this stage, genetic optimization is used to determine appropriate hyperparameters of the classifications model. In this study, genetic optimizer techniques are used to improve machine learning algorithms. It is observed that applying the genetic approach to the KNN (k-nearest neighbor) algorithm improved the prediction accuracy. However, the random forest algorithm does not show significant benefits from genetic adaptation. This is because the Random Forest Algorithm is already considered a robust algorithm and therefore it is not affected by the use of genetic optimization. Algorithm 3.7 describes the generic steps of genetic optimizer algorithm.

#### Algorithm 3.7: Genetic optimizer

**Input:**

- training dataset that contain the significant features
- parameter of algorithms (KNN,RF)
- population size: number of individuals in the population
- mutation rate: probability of mutation for each gene
- cross over: probability of crossover between individuals
- generation: number of generations
- fitness function: function to evaluate the fitness of an individual

**Output:**

The best hyperparameters of each algorithm with best accuracy

**Begin:**

## 1. Initialization:

Initialize population size individuals randomly.

## 2. Evaluation:

For each individual in the population:

Evaluate the fitness of the individual using the objective function(accuracy).

3. Selection:

Select individuals from the population to form the next generation.

4. Crossover:

- Randomly select two individuals from the new generation.

5. Mutation:

For each individual in the new generation ( apply mutation to the individual's genes ).

6. Replacement:

Replace the current population with the new generation of offspring.

7. Termination:

Repeat steps 2 to 6 for a certain number of generations (generations number) or until a stopping criterion is met (e.g., maximum number of generations).

8. Result:

Return the best individual in the final population as the optimized solution or best set of parameters.

**End**

### 3.3.5 The Classification Methods

The primary objective of prediction models is to identify patterns related to a particular issue. These models are utilized to anticipate the target class for a given instance. Classification techniques are useful when the values in the target class are discrete. This thesis focuses on predicting customer satisfaction with airline companies. Consequently, classification techniques are suitable for addressing this problem. While various machine learning algorithms can be employed to predict the target class using a set of features, this thesis primarily discusses the fundamental principles of the data mining algorithms utilized. The total number of instances related to customer satisfaction is 129,880 collected from two files. The first file includes 25,975

instances of customer satisfaction while the second file includes 103,903 instances. The total number of instances is used to construct our models. Four classifiers are implemented to identify the highest accuracy that would be obtained. The applied classifiers are:

1. K-Nearest Neighbor (KNN).
2. Random Forest (RF).
3. Multi-Layer Perceptron (MLP).
4. Long Short-Term Memory (LSTM).

After applying the four classifiers, it is shown that Random Forest (RF) outperformed others. As a result, the proposed system is based on Random Forest for predicting customers' satisfaction. Algorithm 3.8 shows the Random Forest technique.

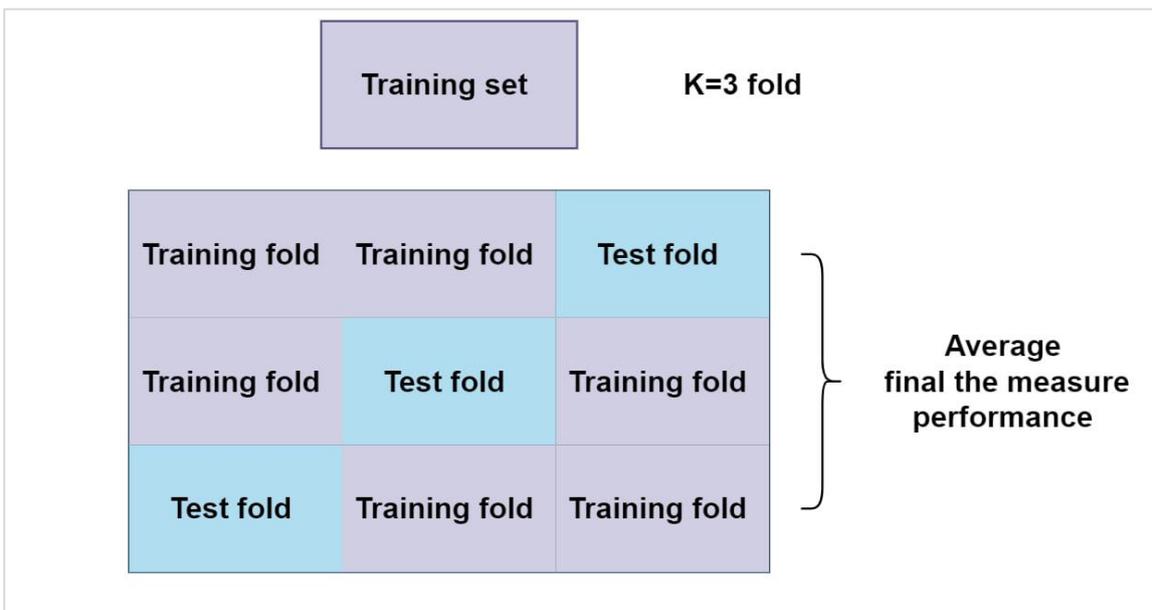
Algorithm 3.8: Random Forest
<p><b>Input:</b></p> <p>-Training dataset, number of trees, max depth , number of features at each split.</p> <p><b>Output:</b></p> <p>-Random Forest model</p> <p><b>Begin:</b></p> <ol style="list-style-type: none"> <li>1. Data Preparation: Prepare the training dataset by dividing it into features (X) and corresponding labels (Y).</li> </ol>
<ol style="list-style-type: none"> <li>2. Random Subset Selection: Randomly select a subset of features from the total set of features.</li> </ol>
<ol style="list-style-type: none"> <li>3. Tree Building: For each tree, build a decision tree using the selected subset of features and a subset of the training dataset.</li> </ol>

4. Prediction: The final prediction can be determined through majority voting (for classification) or averaging (for regression).

**End**

### 3.3.6 Evaluating the Performance of the Proposed Model

Four distinct metrics were employed to assess the effectiveness of a specific classification algorithm. These metrics encompass accuracy, f1-score, precision, and recall. These metrics are determined by analyzing a confusion matrix, a tabular representation that outlines the count of correctly and incorrectly predicted examples by a classification model. In comparing the findings of this research with earlier studies, only the accuracy metric is taken into consideration. The performance of the customers' satisfaction is evaluated using 3-fold Cross-Validation. This means that each instance uses the same number of times for training and exactly once for testing, as shown in Figure 3.5.



**Figure 3.5: Three-fold cross validation for the airline customers satisfaction dataset**

# ***Chapter Four***

## ***Experimental Results and Discussion***

## **Chapter Four**

### **Experimental Results and Discussion**

#### **4.1 Overview**

The proposed methodology illustrated in Chapter three is implemented to cover the research aims illustrated in Chapter one. The experimental results are described and discussed in this Chapter.

#### **4.2 Software and Hardware Environment**

The proposed system is implemented using the following hardware and software requirements.

**Hardware:** Processor Intel core i5, RAM 8GB, Storage 1 TB.

**Operating System:** Windows10 (64) bit.

**Programming Language:** python language.

**Integrated Development Environment IDE:** the system is implemented by Python 3.8.12, Jupyter Notebook, and Google colab.

#### **4.3 Results of Data Preprocessing**

Six important steps are performed on the Airline customer satisfaction namely, data cleaning, handling missing values, converting categorical features to numerical features, unifying the range of age feature, handling outliers and normalization. The data preprocessing steps are achieved on a small sample of airline customer satisfaction. In the data cleaning step, ID customers feature has been removed from the airline customers satisfaction dataset. The reason is that this feature is not useful for the prediction process. Figure 4.1 clarifies the outcomes of removing ID customers.

ID	gender	Age	Customer type	Type of travel	Class	Flight distance	Check in service	Departure delay in minutes	Arrival delay in minutes	Satisfaction
1	Male	53	Loyal customer	Personal Travel	Business	109	3	25	18	Dissatisfied
2	Male	39	Loyal customer	Business travel	Economy	1012	2	11	9	Dissatisfied
3	Male	38	Disloyal customer	Business travel	Business	759	5	0	NaN	Satisfied
4	Female	52	Loyal customer	Personal Travel	Economy	3659	3	0	NaN	Satisfied

**Data cleaning**

gender	Age	Customer type	Type of travel	Class	Flight distance	Check in service	Departure delay in minutes	Arrival delay in minutes	Satisfaction
Male	53	Loyal customer	Personal Travel	Business	109	3	25	18	Dissatisfied
Male	39	Loyal customer	Business travel	Economy	1012	2	11	9	Dissatisfied
Male	38	Disloyal customer	Business travel	Business	759	5	0	NaN	Satisfied
Female	52	Loyal customer	Personal Travel	Economy	3659	3	0	NaN	Satisfied

Figure 4.1: Removing ID customer feature from a sample dataset

After completing the data cleaning step, the missing values are processed in each feature. Missing values only appeared in arrival delay in minutes. Therefore, the mode method on the departure delay in minutes feature is applied to compute the missing values. Figure 4.2 shows a sample of dataset after handling missing values.

gender	Age	Customer type	Type of travel	Class	Flight distance	Check in service	Departure delay in minutes	Arrival delay in minutes	Satisfaction
Male	53	Loyal customer	Personal Travel	Business	109	3	25	18	Dissatisfied
Male	39	Loyal customer	Business travel	Economy	1012	2	11	9	Dissatisfied
Male	38	Disloyal customer	Business travel	Business	759	5	0	NaN	Satisfied
Female	52	Loyal customer	Personal Travel	Economy	3659	3	0	NaN	Satisfied

**Missing values has been processed**

gender	Age	Customer type	Type of travel	Class	Flight distance	Check in service	Departure delay in minutes	Arrival delay in minutes	Satisfaction
Male	53	Loyal customer	Personal Travel	Business	109	3	25	18	Dissatisfied
Male	39	Loyal customer	Business travel	Economy	1012	2	11	9	Dissatisfied
Male	38	Disloyal customer	Business travel	Business	759	5	0	0	Satisfied
Female	52	Loyal customer	Personal Travel	Economy	3659	3	0	0	Satisfied

Figure 4.2: A sample of the dataset after handling the missing values

The transformation of converting each categorical feature to the numeric features is performed, to ensure that the proposed system becomes capable of handling all the features present in the dataset. Figure 4.3 shows a sample of data after the transformation process. After the transformation, the values of age feature were set within a specific range between 1 and 8. Figure 4.4 shows a sample of dataset after the unifying process.

gender	Age	Customer type	Type of travel	Class	Flight distance	Check in service	Departure delay in minutes	Arrival delay in minutes	Satisfaction
Male	53	Loyal customer	Personal Travel	Business	109	3	25	18	Dissatisfied
Male	39	Loyal customer	Business travel	Economy	1012	2	11	9	Dissatisfied
Male	38	Disloyal customer	Business travel	Business	759	5	0	0	Satisfied
Female	52	Loyal customer	Personal Travel	Economy	3659	3	0	0	Satisfied

↓

**After converting data from categorical to the numerical**

gender	Age	Customer type	Type of travel	Class	Flight distance	Check in service	Departure delay in minutes	Arrival delay in minutes	Satisfaction
0	53	1	0	1	109	3	25	18	0
0	39	1	1	0	1012	2	11	9	0
0	38	0	1	1	759	5	0	0	1
1	52	1	0	0	3659	3	0	0	1

**Figure 4.3: A sample of the dataset after transformation**

gender	Age	Customer type	Type of travel	Class	Flight distance	Check in service	Departure delay in minutes	Arrival delay in minutes	Satisfaction
0	53	1	0	1	109	3	25	18	0
0	39	1	1	0	1012	2	11	9	0
0	38	0	1	1	759	5	0	0	1
1	52	1	0	0	3659	3	0	0	1

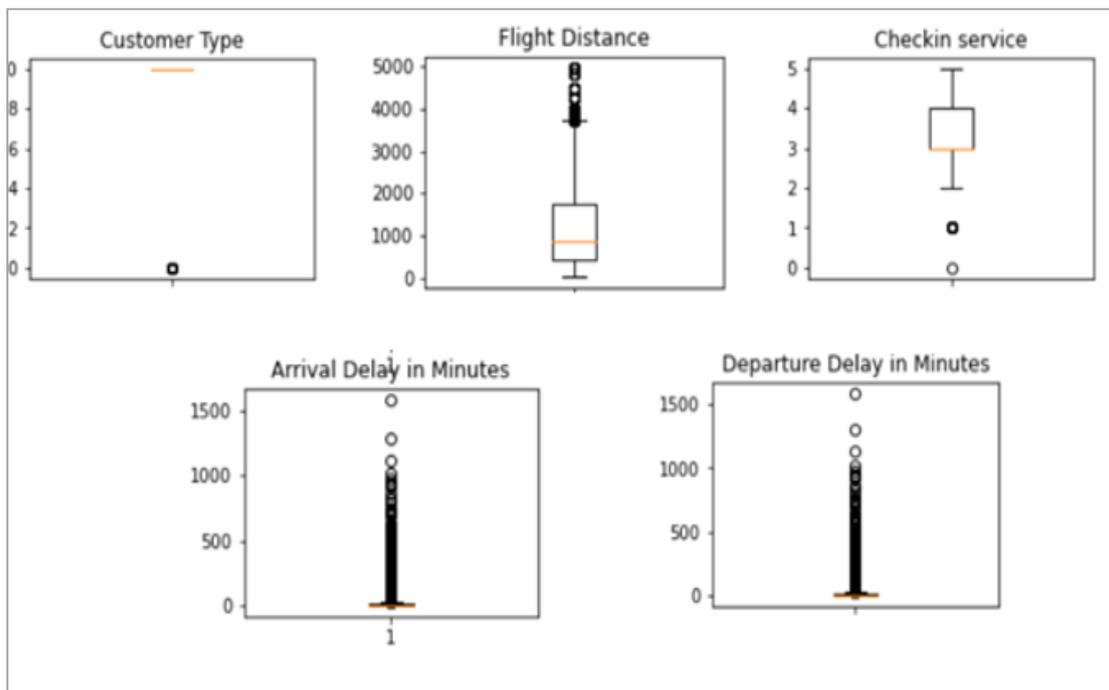
↓

**After unifying Age feature in to range between 1,8**

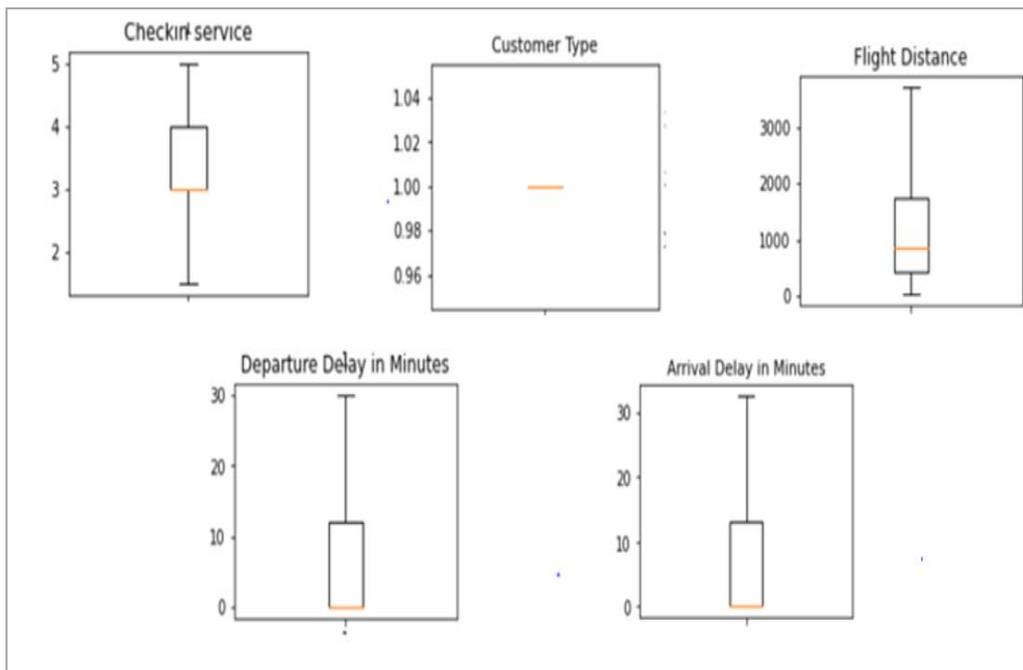
gender	Age	Customer type	Type of travel	Class	Flight distance	Check in service	Departure delay in minutes	Arrival delay in minutes	Satisfaction
0	5	1	0	1	109	3	25	18	0
0	4	1	1	0	1012	2	11	9	0
0	4	0	1	1	759	5	0	0	1
1	5	1	0	0	3659	3	0	0	1

**Figure 4.4: A sample of the dataset after unifying age feature**

In this thesis, the outliers have been taken into consideration and addressed appropriately. To ensure that the model is not sensitive to extreme values and can deliver consistent and reliable results. Outliers have appeared in customer type, flight distance, check-in service, departure delay in minutes, and arrival delay in minutes. The capping technique is applied to handle the outlier in these features. Figure 4.5 shows the outliers identified in these features, whereas Figure 4.6 shows the features after manipulating the capping technique through handling outliers.



**Figure 4.5: Outliers before using the capping technique**



**Figure 4.6: Handling outliers by using the capping technique**

To avoid the difference in large values that dominate the results, normalization is performed. The min-max normalization method has been used to range the feature values between zero and one. Figure 4.7 depicts a sample of the dataset after normalizing.

gender	Age	Customer type	Type of travel	Class	Flight distance	Check in service	Departure delay in minutes	Arrival delay in minutes	Satisfaction
0	5	1	0	1	109	3	25	18	0
0	4	1	1	0	1012	2	11	9	0
0	4	0	1	1	759	5	0	0	1
1	5	1	0	0	3659	3	0	0	1

↓

**Data normalization**

Gender	Age	Customer type	Type of travel	Class	Flight distance	Check in service	Departure delay in minutes	Arrival delay in minutes	Satisfaction
0	0.28	1	0	1	0.24	0.14	0.83	0.55	0
0	0.28	1	1	0	0.75	1	0.36	0.27	0
0	0.71	0	1	1	0.24	0.42	0	0	1
1	0.28	1	0	0	0.26	0.71	0	0	1

**Figure 4.7: A sample of the dataset after normalization**

## 4.4 Results of Feature Generation

Three new features are constructed from the original dataset which are: Service quality, information quality, engagement. Figure 4.8 shows the original dataset before the generation of features, whereas Figure 4.9 shows the dataset after the generation of features.

Gender	Customer type	Age	Type of travel	Class	.....	Leg room service	Baggage handling	Check in service	Inflight service	Cleanliness
0	0	0.14	0	0		0.6	0.75	0.71	1	1
0	0	0.71	1	1		1	0.50	0.42	0.8	0.8
1	0	0.28	1	1		0.8	0.50	0.14	0.6	0.4
1	0	0.57	0	1		0.4	0.70	1	0.8	0.6

**Figure 4.8: A sample of the original dataset before the generation of features**

Gender	Customer type	Age	Type of travel	Class	.....	Leg room service	Baggage handling	Check in service	Inflight service	Cleanliness	Engagement	Information Quality	Service Quality
0	0	0.14	0	0		0.6	0.75	0.71	1	1	0.8	0.69	0.82
0	0	0.71	1	1		1	0.50	0.42	0.8	0.8	0.4	0.10	0.41
1	0	0.28	1	1		0.8	0.50	0.14	0.6	0.4	0.7	0.32	0.85
1	0	0.57	0	1		0.4	0.70	1	0.8	0.6	0.8	0	0.47

**Figure 4.9: A sample of the dataset after the generation of features**

## 4.5 The selected features

This study aimed at predicting customers' satisfaction with airline companies, highlighting the most features that may influence customers' satisfaction. The data of airline companies contains a few numbers of features. However, it should be noted that not all of these features are appropriate for the prediction task. Some of them lead to decreasing the accuracy of the prediction model and increasing its time complexity. Consequently, this study employed a feature selection process to address this issue. Tables 4.1 and 4.2 highlight the subsets of features that appeared after applying the SFS method in the original features and features generation (twenty-five) features using KNN and RF classifiers. Tables 4.3 and 4.4 represent the subsets of features after applying the same method with features generation and features that were not used in the generation process (twelve) features using KNN and RF classifiers. Figure 4.10 shows the contributions of each attribute for the decision making.

**Table 4.1: Subsets of twenty-five features by using the KNN classifiers**

Subsets	Name of features	Score
Subset1	Online boarding	0.787
Subset2	Type of Travel, Online boarding	0.836
Subset3	Type of Travel, Online boarding, and Engagement	0.872

Subsets	Name of features	Score
Subset4	Type of Travel, Online boarding, Inflight entertainment, and Engagement	0.907
Subset5	Type of Travel, Online boarding, Inflight entertainment, Engagement, and Service Quality	0.924
Subset6	Type of Travel, Online boarding, Inflight entertainment, Cleanliness, Engagement, and Service Quality	0.931
Subset7	Type of Travel, Class, Online boarding, Inflight entertainment, Cleanliness, Engagement, and Service Quality	0.934
Subset8	Type of Travel, Class, Inflight Wi-Fi service, Online boarding, Inflight entertainment, Cleanliness, Engagement, and Service Quality	0.937
Subset9	Type of Travel, Class, Inflight Wi-Fi service, Gate location, Online boarding, Inflight entertainment, Cleanliness, Engagement, and Service Quality	0.942
Subset10	Type of Travel, Class, Inflight Wi-Fi service, Gate location, Online boarding, Inflight entertainment, Inflight service, Cleanliness, Engagement, and Service Quality	0.945
Subset11	Type of Travel, Class, Inflight Wi-Fi service, Ease of Online booking, Gate location, Online boarding, Inflight entertainment, Inflight service, Cleanliness, Engagement, and Service Quality	0.946
Subset12	Type of Travel, Class, Inflight Wi-Fi service, Ease of Online booking, Gate location, Online boarding, Inflight entertainment, Baggage handling, Inflight service, Cleanliness, Engagement, and Service Quality	0.946
Subset13	Type of Travel, Class, Inflight Wi-Fi service, Ease of Online booking, Gate location, Online boarding, Seat comfort, Inflight entertainment, Baggage handling, Inflight service, Cleanliness, Engagement, and Service Quality	0.947
Subset14	Customer Type, Type of Travel, Class, Inflight Wi-Fi service, Ease of Online booking, Gate location, Online boarding, Seat comfort, Inflight entertainment, Baggage handling, Inflight service, Cleanliness, Engagement, and Service Quality	0.947
Subset15	Customer Type, Age, Type of Travel, Class, Inflight Wi-Fi service, Ease of Online booking, Gate location, Online boarding, Seat comfort, Inflight entertainment, Baggage handling, Inflight service, Cleanliness, Engagement, and Service Quality	0.94.6
Subset16	Customer Type, Age, Type of Travel, Class, Inflight Wi-Fi service, Departure/Arrival time convenient, Ease of Online booking, Gate location, Online boarding, Seat comfort, Inflight entertainment, Baggage handling, Inflight service, Cleanliness, Engagement, and Service Quality	0.945
Subset17	Customer Type, Age, Type of Travel, Class, Inflight Wi-Fi service, Departure/Arrival time convenient, Ease of Online booking, Gate location, Food and drink, Online boarding, Seat comfort, Inflight entertainment, Baggage handling, Inflight service, Cleanliness, Engagement, and Service Quality	0.944

Subsets	Name of features	Score
Subset18	Customer Type, Age, Type of Travel, Class, Flight Distance, Inflight Wi-Fi service, Departure/Arrival time convenient, Ease of Online booking, Gate location, Food and drink, Online boarding, Seat comfort, Inflight entertainment, Baggage handling, Inflight service, Cleanliness, Engagement, and Service Quality	0.942
Subset19	Customer Type, Age, Type of Travel, Class, Flight Distance, Inflight Wi-Fi service, Departure/Arrival time convenient, Ease of Online booking, Gate location, Food and drink, Online boarding, Seat comfort, Inflight entertainment, Leg room service, Baggage handling, Inflight service, Cleanliness, Engagement, and Service Quality	0.940
subset20	Customer Type, Age, Type of Travel, Class, Flight Distance, Inflight Wi-Fi service, Departure/Arrival time convenient, Ease of Online booking, Gate location, Food and drink, Online boarding, Seat comfort, Inflight entertainment, On-board service, Leg room service, Baggage handling, Inflight service, Cleanliness, Engagement, and Service Quality	0.938
subset21	Customer Type, Age, Type of Travel, Class, Flight Distance, Inflight Wi-Fi service, Departure/Arrival time convenient, Ease of Online booking, Gate location, Food and drink, Online boarding, Seat comfort, Inflight entertainment, On-board service, Leg room service, Baggage handling, Check-in service, Inflight service, Cleanliness, Engagement, and Service Quality	0.936
Subset22	Customer Type, Age, Type of Travel, Class, Flight Distance, Inflight Wi-Fi service, Departure/Arrival time convenient, Ease of Online booking, Gate location, Food and drink, Online boarding, Seat comfort, Inflight entertainment, On-board service, Leg room service, Baggage handling, check-in service, Inflight service, Cleanliness, Engagement, Information Quality, and Service Quality	0.933
Subset23	Customer Type, Age, Type of Travel, Class, Flight Distance, Inflight Wi-Fi service, Departure/Arrival time convenient, Ease of Online booking, Gate location, Food and drink, Online boarding, Seat comfort, Inflight entertainment, On-board service, Leg room service, Baggage handling, Check-in service, Inflight service, Cleanliness, Arrival Delay in Minutes, Engagement, Information Quality, and Service Quality	0.932
Subset24	Customer Type, Age, Type of Travel, Class, Flight Distance, Inflight Wi-Fi service, Departure/Arrival time convenient, Ease of Online booking, Gate location, Food and drink, Online boarding, Seat comfort, Inflight entertainment, On-board service, Leg room service, Baggage handling, Check-in service, Inflight service, Cleanliness, Departure Delay in Minutes, Arrival Delay in Minutes, Engagement, Information Quality, and Service Quality	0.930
Subset25	Gender, Customer Type, Age, Type of Travel, Class, Flight Distance, Inflight Wi-Fi service, Departure/Arrival time convenient, Ease of Online booking, Gate location, Food and drink, Online boarding, Seat comfort, Inflight entertainment, On-board service, Leg room service, Baggage handling, Check-in service, Inflight service, Cleanliness, Departure Delay in Minutes, Arrival Delay in Minutes, Engagement, Information Quality, and Service Quality	0.925

**Table 4.2: Subsets of twenty-five features by using the RF classifier**

Subsets	Name of features	Score
Subset1	Online boarding	0.790
Subset2	Type of Travel and Online boarding	0.849
Subset3	Type of Travel, Inflight Wi-Fi service, and Online boarding	0.890
Subset4	Type of Travel, Inflight Wi-Fi service, Online boarding, and service quality	0.925
Subset5	Type of Travel, Flight distance, Inflight Wi-Fi service, Online boarding, and service quality	0.947
Subset6	Type of Travel, Flight distance, Inflight Wi-Fi service, Gate location, Online boarding, and service quality	0.956
Subset7	Type of Travel, Class, Flight distance, Inflight Wi-Fi service, Gate location, Online boarding, and service quality	0.960
Subset8	Type of Travel, Class, Flight distance, Inflight Wi-Fi service, Gate location, Online boarding, Engagement, and service quality	0.963
Subset9	Type of Travel, Class, Flight distance, Inflight Wi-Fi service, Gate location, Online boarding, Inflight service, Engagement, and service quality	0.967
Subset10	Type of Travel, Class, Flight distance, Inflight Wi-Fi service, Gate location, Online boarding, Baggage handling, Inflight service, Engagement, and service quality	0.969
Subset11	Type of Travel, Class, Flight distance, Inflight Wi-Fi service, Gate location, Online boarding, Seat comfort, Baggage handling, Inflight service, Engagement, and service quality	0.970
Subset12	Age, Type of Travel, Class, Flight distance, Inflight Wi-Fi service, Gate location, Online boarding, Seat comfort, Baggage handling, Inflight service, Engagement, and service quality	0.971
Subset13	Age, Type of Travel, Class, Flight distance, Inflight Wi-Fi service, Gate location, Online boarding, Seat comfort, Baggage handling, Inflight service, Cleanliness, Engagement, and service quality	0.972
Subset14	Age, Type of Travel, Class, Flight distance, Inflight Wi-Fi service, Departure/Arrival time convenient, Gate location, Online boarding, Seat comfort, Baggage handling, Inflight service, Cleanliness, Engagement, and service quality	0.972
Subset15	Age, Type of Travel, Class, Flight distance, Inflight Wi-Fi service, Departure/Arrival time convenient, Gate location, Online boarding, Seat comfort, Baggage handling, check in service, Inflight service, Cleanliness, Engagement, and service quality	0.973
Subset16	Age, Type of Travel, Class, Flight distance, Inflight Wi-Fi service, Departure/Arrival time convenient, Gate location, Online boarding, Seat comfort, On-board service Baggage handling, check in service, Inflight service, Cleanliness, Engagement, and service quality	0.973

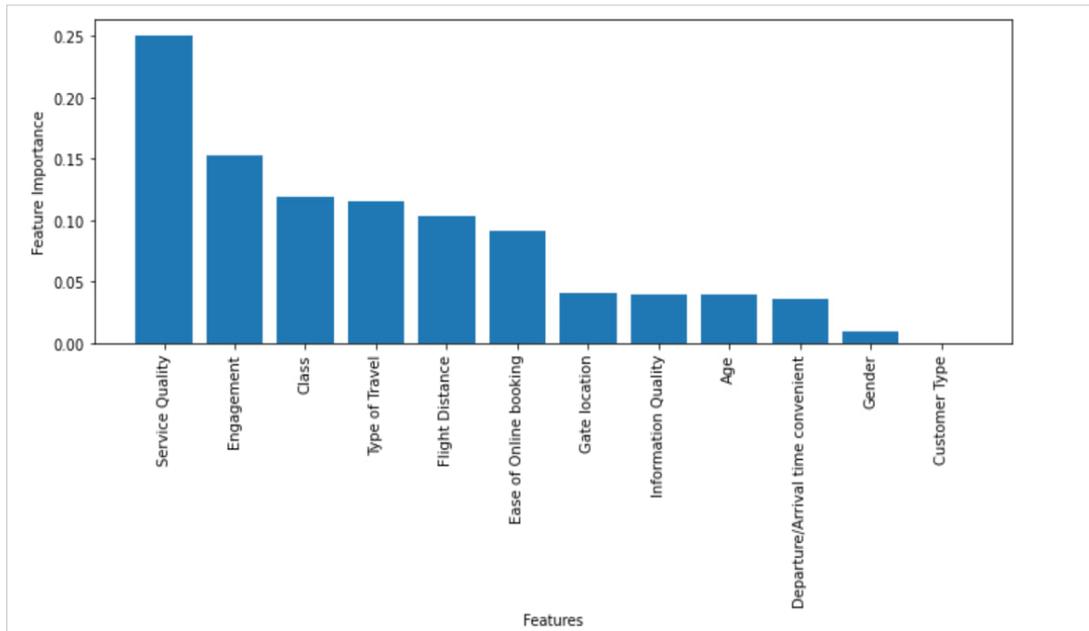
Subsets	Name of features	Score
Subset17	Age, Type of Travel, Class, Flight distance, Inflight Wi-Fi service, Departure/Arrival time convenient, Ease of online booking, Gate location, Online boarding, Seat comfort, On-board service, Baggage handling, check in service, Inflight service, Cleanliness, Engagement, and service quality	0.9741
Subset18	Age, Type of Travel, Class, Flight distance, Inflight Wi-Fi service, Departure/Arrival time convenient, Ease of online booking, Gate location, Online boarding, Seat comfort, Inflight entertainment, On-board service, Baggage handling, check in service, Inflight service, Cleanliness, Engagement, and service quality	0.9743
Subset19	Gender, Age, Type of Travel, Class, Flight distance, Inflight Wi-Fi service, Departure/Arrival time convenient, Ease of online booking, Gate location, Online boarding, Seat comfort, Inflight entertainment, On-board service, Baggage handling, check in service, Inflight service, Cleanliness, Engagement, and service quality	0.9747
Subset20	Gender, Customer type, Age, Type of Travel, Class, Flight distance, Inflight Wi-Fi service, Departure/Arrival time convenient, Ease of online booking, Gate location, Online boarding, Seat comfort, Inflight entertainment, On-board service, Baggage handling, check in service, Inflight service, Cleanliness, Engagement, and service quality	0.9744
Subset21	Gender, Customer type, Age, Type of Travel, Class, Flight distance, Inflight Wi-Fi service, Departure/Arrival time convenient, Ease of online booking, Gate location, Online boarding, Seat comfort, Inflight entertainment, On-board service, Leg room service, Baggage handling, check in service, Inflight service, Cleanliness, Engagement, and service Quality	0.9743
Subset22	Gender, Customer type, Age, Type of Travel, Class, Flight distance, Inflight Wi-Fi service, Departure/Arrival time convenient, Ease of online booking, Gate location, Online boarding, Seat comfort, Inflight entertainment, On-board service, Leg room service, Baggage handling, check in service, Inflight service, Cleanliness, Engagement, Information quality, and service quality	0.9741
Subset23	Gender, Customer type, Age, Type of Travel, Class, Flight distance, Inflight Wi-Fi service, Departure/Arrival time convenient, Ease of online booking, Gate location, Food and drink, Online boarding, Seat comfort, Inflight entertainment, On-board service, Leg room service, Baggage handling, check in service, Inflight service, Cleanliness, Engagement, Information quality, and service quality	0.973
Subset24	Gender, Customer type, Age, Type of Travel, Class, Flight distance, Inflight Wi-Fi service, Departure/Arrival time convenient, Ease of online booking, Gate location, Food and drink, Online boarding, Seat comfort, Inflight entertainment, On-board service, Leg room service, Baggage handling, check in service, Inflight service, Cleanliness, Departure delay in minutes, Engagement, Information quality, and service quality	0.973
Subset25	Gender, Customer type, Age, Type of Travel, Class, Flight distance, Inflight Wi-Fi service, Departure/Arrival time convenient, Ease of online booking, Gate location, Food and drink, Online boarding, Seat comfort, Inflight entertainment, On-board service, Leg room service, Baggage handling, check in service, Inflight service, Cleanliness, Departure delay in minutes, Arrival delay in minutes, Engagement, Information quality, and service quality	0.973

**Table 4.3: Subsets of twelve features by using the KNN classifier**

Subsets	Name of features	Score
Subset1	Service quality	0.740
Subset2	Type of travel and service quality	0.828
Subset3	Type of travel, Engagement, and service quality	0.867
Subset4	Type of travel, Ease of online booking, Engagement, and service quality	0.888
Subset5	Type of travel, Ease of online booking, Gate location, Engagement, and service quality	0.902
Subset6	Type of travel, Class, Ease of online booking, Gate location, Engagement, and service quality	0.912
Subset7	Type of travel, Class, Departure/Arrival time convenient, Ease of online booking, Gate location, Engagement, and service quality	0.918
Subset8	Customer type, Type of travel, Class, Departure/Arrival time convenient, Ease of online booking, Gate location, Engagement, and service quality	0.918
Subset9	Customer type, Age, Type of travel, Class, Departure/Arrival time convenient, Ease of online booking, Gate location, Engagement, and service quality	0.916
Subset10	Customer type, Age, Type of travel, Class, Flight distance, Departure/Arrival time convenient, Ease of online booking, Gate location, Engagement, and service quality	0.915
Subset11	Gender, Customer type, Age, Type of travel, Class, Flight distance, Departure/Arrival time convenient, Ease of online booking, Gate location, Engagement, and service quality	0.911
Subset12	Gender, Customer type, Age, Type of travel, Class, Flight distance, Departure/Arrival time convenient, Ease of online booking, Gate location, Engagement, information quality, and service quality	0.905

**Table 4.4: Subsets of twelve features by using the RF classifier**

Sublets	Name of features	Score
Subset1	Service quality	0.755
Subset2	Type of travel and service quality	0.845
Subset3	Type of travel, flight distance, and service quality	0.874
Subset4	Type of travel, flight distance, ease of online booking , and service quality	0.905
Subset5	Type of travel, flight distance, ease of online booking, engagement, and service quality	0.925
Subset6	Type of travel, flight distance, ease of online booking, gate location, engagement, and service quality	0.937
Subset7	Type of travel, class, flight distance, ease of online booking, gate location, engagement, and service quality	0.943
Subset8	Type of travel, class, flight distance, Departure/Arrival time convenient, ease of online booking, gate location, engagement, and service quality	0.948
Subset9	Age, type of travel, class, flight distance, Departure/Arrival time convenient, ease of online booking, gate location, engagement, and service quality	0.9510
Subset10	Gender, age, type of travel, class, flight distance, Departure/Arrival time convenient, ease of online booking, gate location, engagement, and service quality	0.9516
Subset11	Gender, customer type, age, type of travel, class, flight distance, Departure/Arrival time convenient, ease of online booking, gate location, engagement, and service quality	0.9515
Subset12	Gender, customer type, age, type of travel, class, flight distance, Departure/Arrival time convenient, ease of online booking, gate location, engagement, information quality, and service quality	0.9514



**Figure 4.10: The contribution of features to decision making**

## 4.6 Results of the Classification Methods

The total number of classified customers is 129,880 splits by using cross validation (3-fold). The performance of the four classifiers' predictions is assessed by evaluating their accuracy. The accuracy measure is particularly influenced by the distribution of the target variable.

### 4.6.1 Results of applying machine learning techniques

After the preprocessing steps, feature generation, feature selection, optimization, classification, and performance evaluation stages for two classifiers are performed. The accuracy of the obtained machine learning algorithms for each phase in the prediction process is evaluated. In the first phase, the prediction accuracy is obtained based on the original twenty-two features of the dataset as a baseline. The only preprocessing step that is implemented with the baseline is handling the missing values. The findings are summarized in Table 4.5. It can be noticed that the accuracy for the K-Nearest Neighbor (KNN) classifier is lower than that for the Random Forest classifier (RF). The second phase prediction the accuracy is obtained base of

the original features and feature generation twenty-five features. Table 4.6 explains the accuracy after completing the preprocessing and feature generation.

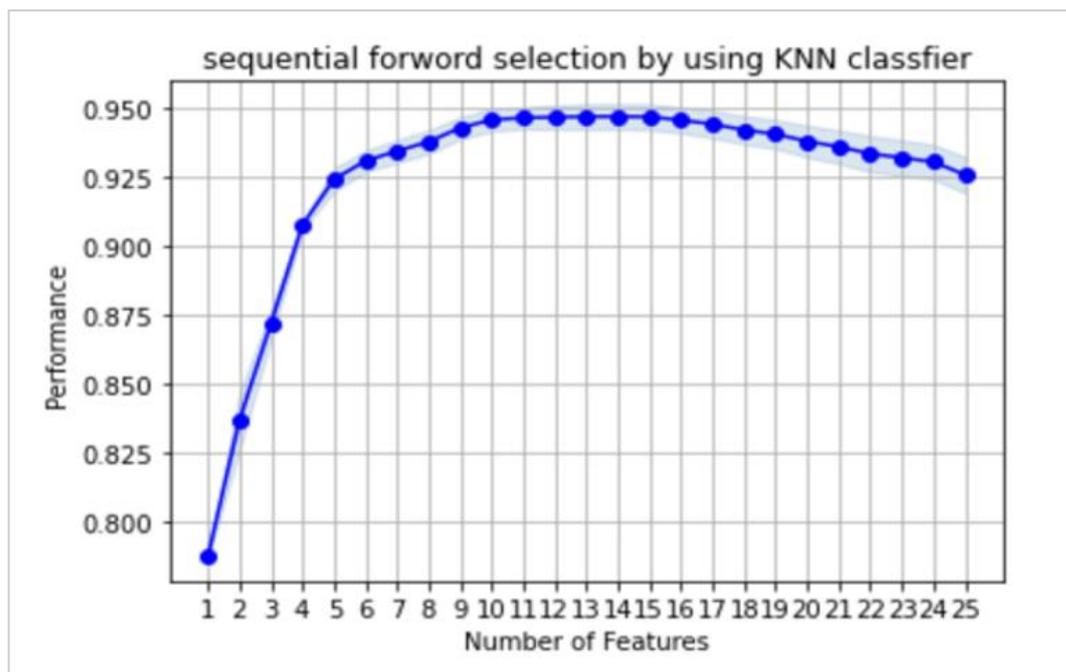
**Table 4.5: A summary of the accuracy with the twenty-two features after handling missing values**

Classifier	Accuracy	Precision	Recall	F1-score	Handling missing values	Handling outliers	Normalization
KNN	0.78	0.78	0.77	0.77	✓	X	X
RF	0.96	0.96	0.96	0.96	✓	X	X

**Table 4.6: A summary of the accuracy and other metrics with the base attributes and the generated features**

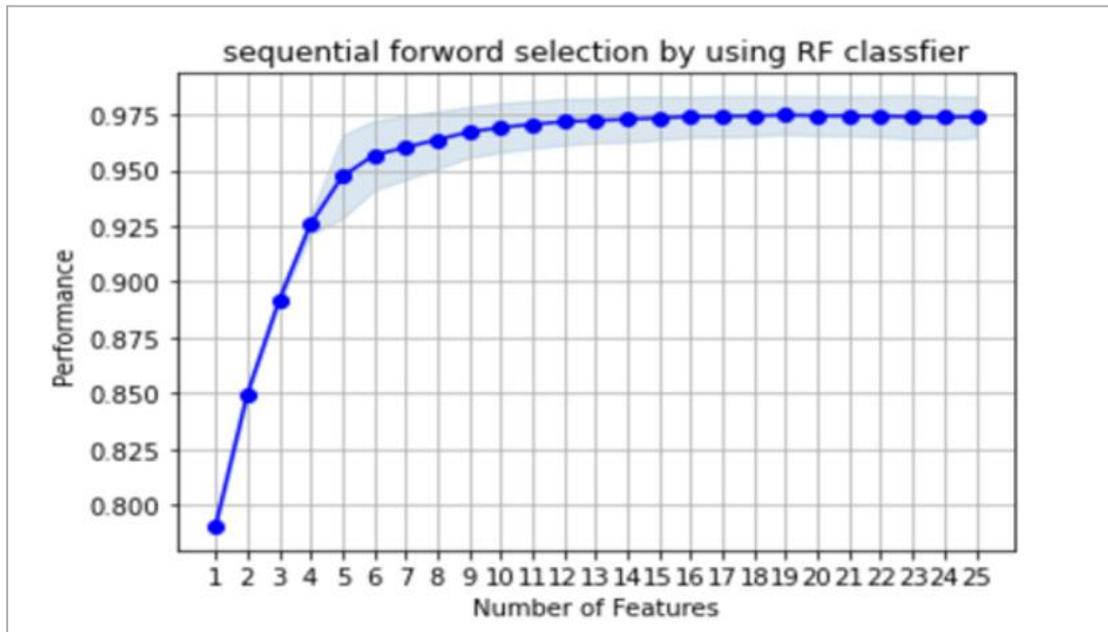
Classifier	Accuracy	Precision	Recall	F1-score	Handling missing values	Handling outliers	Normalization
KNN	0.93	0.925	0.92	0.925	✓	✓	✓
RF	0.97	0.96	0.96	0.96	✓	✓	✓

After generating the new features, the sequential forward feature selection algorithm is applied. When the KNN algorithm is used. It can be seen in Figure 4.11 that the subset of 12 features had the highest accuracy consisting of; Type of Travel, Class, Inflight Wi-Fi service, Ease of Online booking, Gate location, Online boarding, Seat comfort, Inflight entertainment, Baggage handling, Inflight service, Cleanliness, Engagement and Service Quality.



**Figure 4.11:** The performance accuracy of 25 subset of features by using KNN

After using the Random Forest algorithm, it can be noticed from Figure 4.12 that the subset of 19 features had the highest accuracy, and the name of these features are; Gender, Age, Type of Travel, Class, Flight Distance, Inflight Wi-Fi service, Departure/Arrival time convenient, Ease of Online booking, Gate location, Online boarding, Seat comfort, Inflight entertainment, On-board service, Baggage handling, Check-in service, Inflight service, Cleanliness, Engagement and Service Quality.



**Figure 4.12: The performance accuracy of 25 subset of features by using RF**

Genetic optimizer is applied after completing the feature selection to enhance the classifier accuracy. Table 4.7 explains a comparison of the classifiers' accuracy by using the wrapper feature selection and genetic optimizer based on the twenty-five features. According to the results in Table 4.7, it can be concluded that the performance accuracy of the Random Forest was not enhanced after implementing both the wrapper feature selection and the genetic optimizer techniques. This may be attributed to the fact that the Random Forest, which is composed of multiple trees, has the capacity of identifying significant features [51]. On the other hand, the accuracy of the K-Nearest Neighbor (KNN) classifier increases from 95% to 96% after applying the genetic optimizer. Table 4.8 describe the best hyperparameters of KNN and RF after applied the genetic optimizer with twenty-five features.

**Table 4.7: A comparison of classifiers' accuracy in the case of genetic optimizer and wrapper feature selection techniques based on twenty-five features**

Classifiers	Wrapper feature selection				Genetic optimizer
	Accuracy	Precision	Recall	F1-score	Accuracy %
<b>KNN</b>	<b>0.95</b>	<b>0.945</b>	<b>0.945</b>	<b>0.945</b>	<b>0.96</b>
<b>RF</b>	<b>0.97</b>	<b>0.975</b>	<b>0.97</b>	<b>0.975</b>	<b>0.97</b>

**Table 4.8: A summary of the best hyperparameters after applying the genetic optimizer with twenty -five features**

No of features	Algorithm	Input parameters	Output best hyperparameters
25	<b>KNN</b>	<b>N-neighbor: range (1,20)</b>	<b>Neighbor: 1</b>
	<b>RF</b>	<b>N-estimators: range (1,200)</b>	<b>N-estimators: 151</b>
		<b>Max_features: (sqrt, log)</b>	<b>Max_features: sqrt</b>
		<b>Max_depth: range (5,75)</b>	<b>Max_depth: 42</b>
		<b>Min_samples_split: range (2,11)</b>	<b>Min_samples_split: 3</b>
		<b>Min_samples_leaf: range (1,11)</b>	<b>Min_samples_leaf: 1</b>

In the final phase prediction, the accuracy is obtained by using features generated namely; service quality, information quality and engagement in addition to other features. This process aims to minimize the number of features from twenty-two features to twelve features to decrease the model's complexity. Figure 4.13 describes the relationship and strength between features. The intensity of the color shows that there is a connection between the features. Table 4.9 summarize the accuracy of the classifiers after minimizing the number of features.

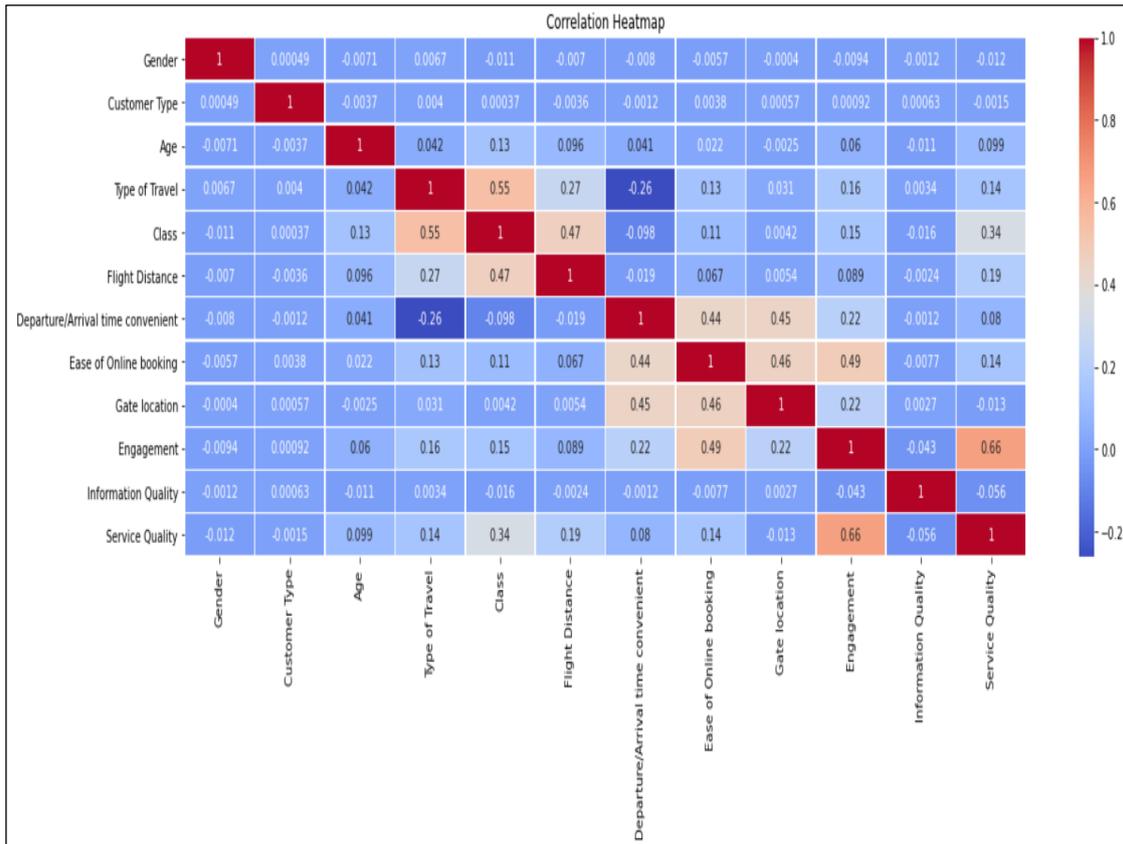


Figure 4.13: The relationship between features

Table 4.9: A summary of the accuracy with twelve features

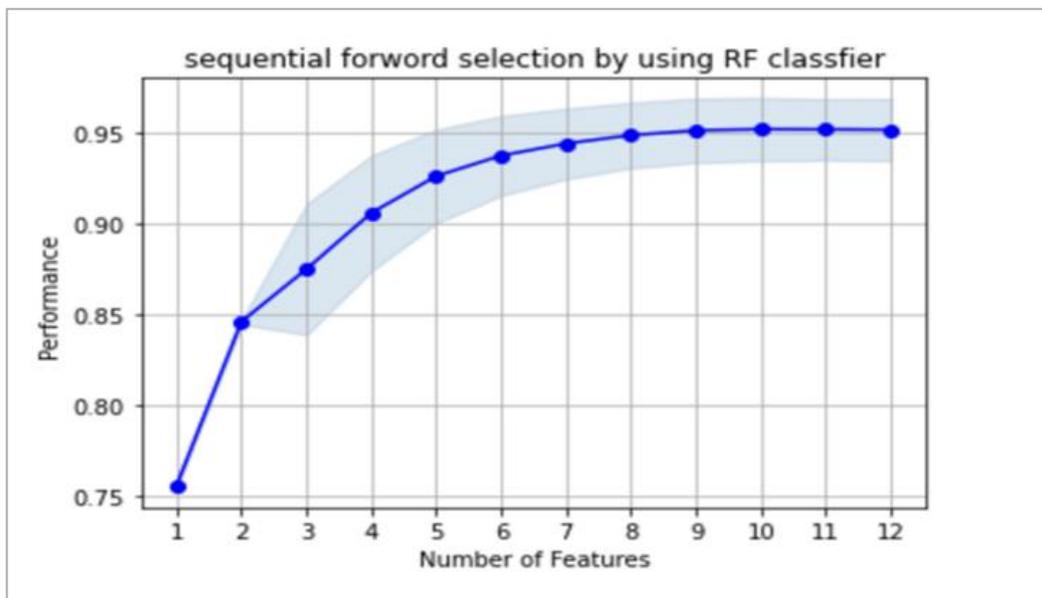
Classifier	Accuracy	Precision	Recall	F1-score	Handling missing values	Handling outliers	Normalization
KNN	0.91	0.905	0.905	0.905	✓	✓	✓
RF	0.95	0.95	0.95	0.95	✓	✓	✓

After evaluating accuracy with twelve features, the forward sequential feature selection algorithm is chosen. The results reveal that the service quality is the best feature. The second step is a comparison of service quality with each feature. This is conducted until obtaining a sequence of twelve subsets of features that had different degrees of accuracy.

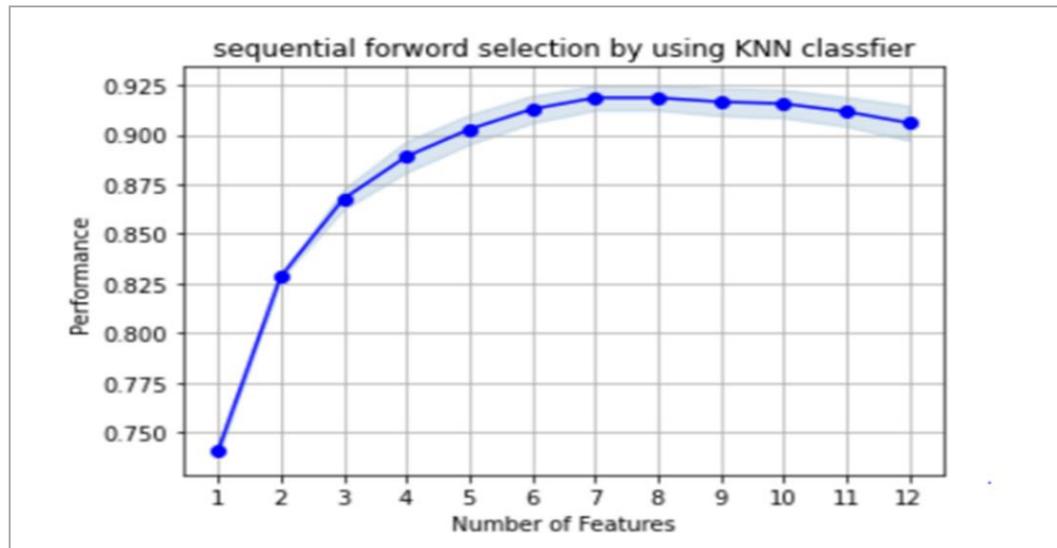
Figure 4.14 depicts that the subset of ten features with the Random Forest classifier (RF) had the highest accuracy. These features are gender, age, type of travel, class, flight distance, departure/arrival time convenience, ease of online booking, gate location, engagement, and service quality. For The K-Nearest Neighbor (KNN) classifier, service quality is also the best-selected feature. By comparing service quality with each feature until obtaining a sequence of twelve subsets of features that had different degrees of accuracy, a subset of seven and eight features had the highest accuracy as shown in Figure 4.15. These features are the type of travel, class, departure/arrival time convenience, ease of online booking, gate location, engagement and service quality for the subset of seven features. For the other subset, these features are customer type, type of travel, class, departure/arrival time convenience, ease of online booking, gate location, engagement and service quality. The implementation of the feature selection step with the Random Forest classifiers (RF) and K-Nearest Neighbor (KNN) is shown in Figure 4.14 and Figure 4.15. It is clear that the new features generated in this research namely, service quality and engagement are significant predictors of customers' satisfaction. This result is matching with previous literature about user satisfaction in other fields [30],[52].

The results also suggest that information quality was not a significant predictor of airline customer satisfaction. This finding is in agreement with previous literature [53]. According to the results, airline companies may be aware of service quality and engagement features to ensure customers' satisfaction. Table 4.10 explains a comparison of the classifiers' accuracy by using the wrapper feature selection and genetic optimizer based on the twelve features. According to the results in Table 4.10, it can be concluded that the performance accuracy of the Random Forest was not enhanced after implementing both the wrapper feature selection and the genetic optimizer

techniques. This may be attributed to the fact that the Random Forest, which is composed of multiple trees, has the capacity of identifying significant features [51]. However, such findings are in line with other studies [1], [11]. The outcomes of this present research are better than another research study [10]. However, twenty-two features are used in [10], whereas this thesis deals with only twelve. On the other hand, the accuracy of the K-Nearest Neighbor (KNN) classifier increases from 92% to 93% after applying the genetic optimizer. Table 4.11 summarized the best hyperparameters after applied the genetic optimizer.



**Figure 4.14:** The performance accuracy of twelve features based on the RF



**Figure 4.15:** The performance accuracy of twelve features based on KNN

**Table 4.10:** A comparison of the classifiers' accuracy in the case of wrapper feature selection and the genetic optimizer techniques based on twelve features

Classifiers	Wrapper feature selection				Genetic optimizer
	Accuracy	Precision	Recall	F1-score	Accuracy %
<b>KNN</b>	<b>0.92</b>	<b>0.915</b>	<b>0.92</b>	<b>0.92</b>	<b>0.93</b>
<b>RF</b>	<b>0.95</b>	<b>0.95</b>	<b>0.95</b>	<b>0.95</b>	<b>0.95</b>

**Table 4.11: A summary of the best hyperparameters after applying the genetic optimizer with twelve features**

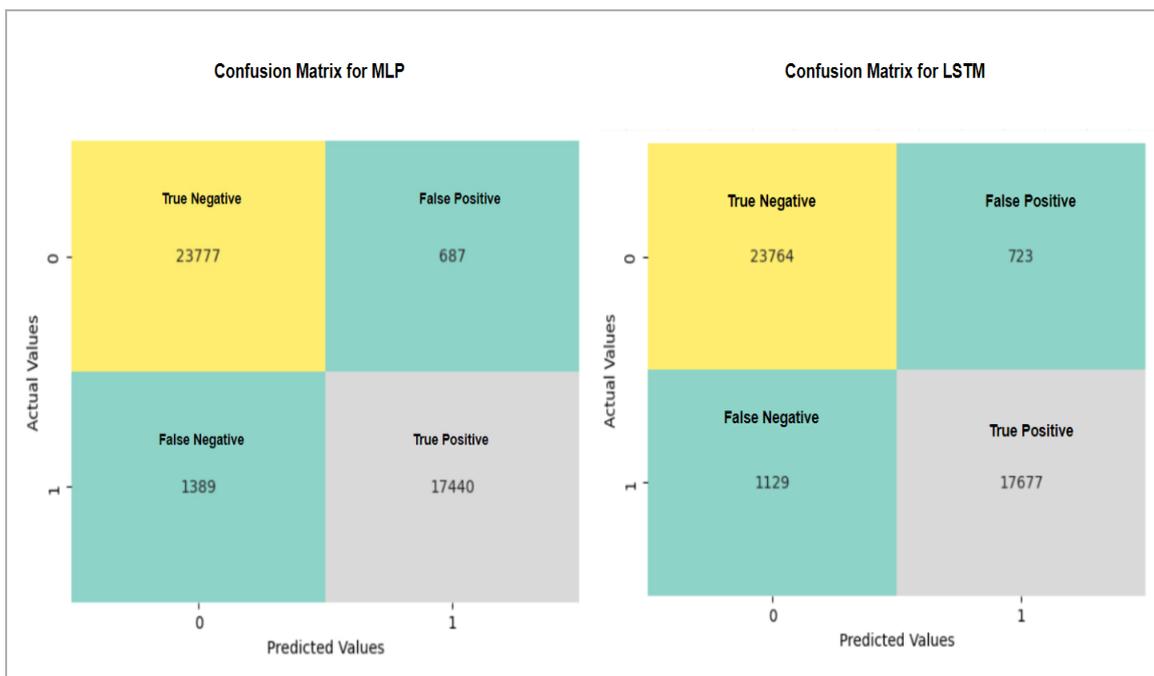
No of features	Algorithm	Input parameters	Output best hyperparameters
12	KNN	N-neighbor: range (1,20)	Neighbor: 1
	RF	N-estimators: range (1,200)	N-estimators: 81
		Max_features: (sqrt, log)	Max_features: sqrt
		Max_depth: range (5,75)	Max_depth: 46
		Min_samples_split: range (2,11)	Min_samples_split: 2
		Min_samples_leaf: range (1,11)	Min_samples_leaf: 1

#### 4.6.2 Results of applying deep learning techniques

After the preprocessing steps, feature generation, classification, and performance evaluation stages for the classifier are performed. The accuracy obtained in the first phase by using the original twenty- two features with deep learning algorithms is summarized in Table 4.12. Figure 4.16 shows the confusion matrix for MLP and LSTM algorithm with twenty -two features.

**Table 4.12: A summary of the accuracy of deep learning techniques based on twenty-two features**

Classifier	Accuracy	Precision	Recall	F1-score	Handling missing values	Handling outliers	Normalization
MLP	95.2	95	95	95	✓	✓	✓
LSTM	95.72	95.5	95.5	95.5	✓	✓	✓

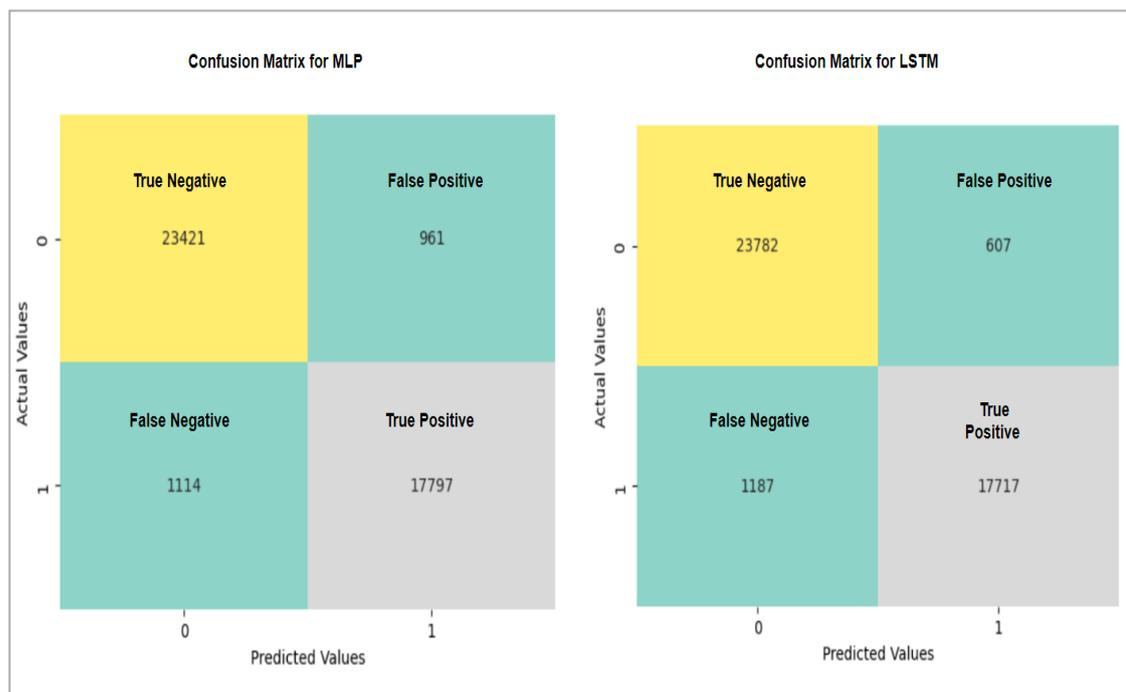


**Figure 4.16: The confusion matrix of MLP and LSTM with original features**

In the second phase, prediction accuracy is obtained using the original features and the three generated features resulting in twenty-five features, by using the deep learning techniques is summarized in Table 4.13. The confusion matrix by using MLP and LSTM algorithm is shown in Figure 4.17.

**Table 4.13: A summary of the accuracy of deep learning with the twenty-five features**

Classifier	Accuracy	Precision	Recall	F1-score	Handling missing values	Handling outliers	Normalization
MLP	95.2	95	95	95	✓	✓	✓
LSTM	96	95.5	95.5	95.5	✓	✓	✓

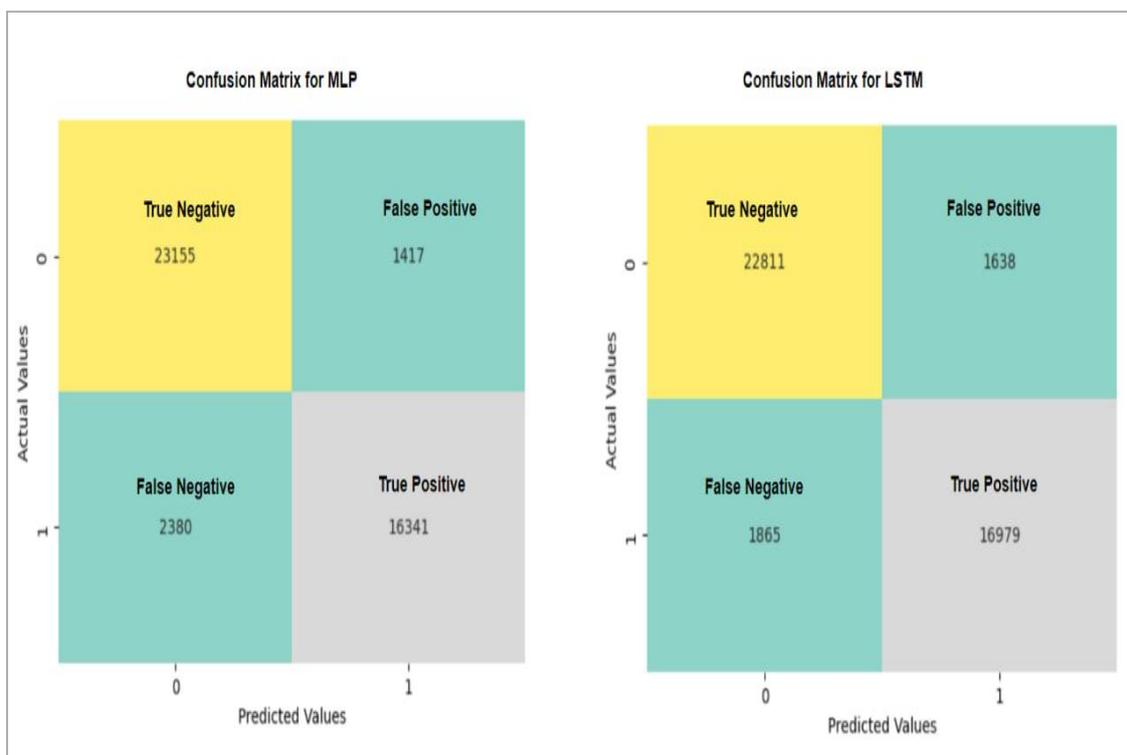


**Figure 4.17: The confusion matrix of MLP and LSTM by using twenty-five features**

In the third phase prediction the accuracy is obtained by using deep learning techniques. In this phase, the accuracy is obtained by using features generated namely; service quality, information quality and engagement in addition to other features. Table 4.14 demonstrates that the accuracy of LSTM is higher than MLP. Figure 4.18 shows the confusion matrix of MLP and LSTM with twelve features. Table 4.15 A comparison between the findings of previous literature and this study. Table 4.16 shows the comparison of time executions based on twenty -five and twelve features.

**Table 4.14: A summary of the accuracy of deep learning techniques based on twelve features**

Classifier	Accuracy	Precision	Recall	F1-score	Handling missing values	Handling outliers	Normalization
MLP	91.23	91.5	90.5	91	✓	✓	✓
LSTM	91.91	91.5	91.5	92	✓	✓	✓



**Figure 4.18: The confusion matrix of MLP and LSTM with twelve features**

**Table 4.15: A summary of the comparison between findings of previous literature and this thesis**

Reference		Algorithm used	Accuracy %
Amalia et.al [1]		RF	95.8
		SPAARC	95
Baswardono et.al [10]		RF	93
		C4.5	92
Hoon et.al [11]		RF	95.7
		DNN	86
The proposed system	By using the original features	RF	96
		KNN	78
		MLP	95.2
		LSTM	95.7
	By using the generated and original features	RF and Genetic optimizer	97
		KNN and Genetic optimizer	96
		MLP	95.2
		LSTM	96
	By using the generated and other unused features in the generation process	RF and Genetic optimizer	95
		KNN and Genetic optimizer	93
		MLP	91.2
		LSTM	91.9

Table 4.16: The time execution of all cases

No. features	Machine learning techniques			
Twenty- five features	Time execution			
	Classifiers	Preprocessing	Feature selection	Genetic optimizer
	KNN	2.534 Second	43800 Second	1457.15 Second
	RF	4.374 Second	55600 Second	11504.04 Second
Twelve features	Classifiers	Preprocessing	Feature selection	Genetic optimizer
	KNN	1.342 Second	32891 Second	286.101 Second
	RF	4.113 Second	14400 Second	4634.13 Second
Deep learning techniques				
Twenty- five features	Time execution			
	Classifiers	Preprocessing and implementing		
	MLP	7200 Second		
	LSTM	10800 Second		
Twelve features	Classifiers	Preprocessing and implementing		
	MLP	1908 Second		
	LSTM	5400 Second		

## 4.7 Summary

- Many preprocessing techniques are applied before passing the research data into the proposed classification model. The prediction accuracy of the models is obtained for customer airline satisfaction by using the original features and the features generated which are service quality, information quality, and engagement in addition to other features. This process aims to minimize the number of features from twenty-two (original) features to twelve features to decrease the model's complexity
- Feature selection is used to avoid unnecessary features and select the most important ones. During each prediction period, the Sequential Forward Selection (SFS) technique is utilized to identify significant features.
- The machine learning models employed in this thesis demonstrate superior predictive performance in comparison to deep learning approaches. These findings align with previous studies that have consistently shown that machine learning algorithms outperform deep learning in terms of performance accuracy. The results reveal that service quality and engagement are significant predictors of customers' satisfaction, whereas information quality is not a predictor feature.

# ***Chapter Five***

## ***Conclusions and Future Works***

## **Chapter Five**

### **Conclusions and Future Works**

#### **5.1 Conclusions**

- The predicting is conducted based on different number of features. According to the findings, it can be noticed that RF outperformed all other classifiers.
- It can be concluded that dealing with a number of twelve features can minimize the model's complexity and result in reasonable prediction accuracy.
- The research outcomes can help airline companies improve their services and respond to customers' needs.
- Any dataset should be cleaned well from missing values and outliers as these could affect the model accuracy.

#### **5.2 Theoretical and practical implications**

This research has many implications that would be beneficial for airline companies to improve the quality of services provided to customers. This can include:

- 1) The findings indicate that service quality is the most influential factor in determining customer satisfaction.
- 2) The overall outcomes outperform previous literature with the integration of optimization techniques.
- 3) An updated customer satisfaction framework is proposed based on earlier theories and the original features of the dataset.

- 4) The model accuracy is compared based on the original features of the dataset and the generated features.
- 5) The most effective features on customers' satisfaction are identified.
- 6) It is crucial for airline companies to prioritize and maintain high levels of service quality to meet customers' needs effectively.
- 7) The study emphasizes the concept of engagement, which relates to customers' enjoyment and positive experiences with a specific airline or technology.
- 8) The research suggests that incorporating engagement activities during flights can enhance customers' satisfaction. This should motivate airline companies to offer more engagement activities, ultimately leading to the cultivation of a loyal customer base.

### **5.3 Future Work**

Many future directions can be inferred based on this present thesis finding such as:

1. The proposed model can be implemented on other datasets such as collecting a new dataset of Iraqi airline companies.
2. Other methods of feature selection can be implemented which may show different or better results in comparison to the used method in this thesis.

# *References*

## References

- [1] S. Amalia, I. Deborah, and I. N. Yulita, “Comparative analysis of classification algorithm: Random Forest, SPAARC, and MLP for airlines customer satisfaction,” *Sinergi*, vol. 26, no. 2, pp. 213–222, 2022, doi: 10.22441/sinergi.2022.2.010.
- [2] M. Deveci, M. E. Çiftçi, İ. Z. Akyurt, and E. D. R. S. Gonzalez, “Impact of COVID-19 pandemic on the Turkish civil aviation industry,” *Sustain. Oper. Comput.*, vol. 3, no. October 2021, pp. 93–102, 2022, doi: 10.1016/j.susoc.2021.11.002.
- [3] J. Ohliati and B. S. Abbas, “Measuring Students Satisfaction in Using Learning Management System,” pp. 180–189.
- [4] E. Saglik, A. Caglar Gulluce, U. Kaya, and K. C. Ozhan, “Service Quality and Customer Satisfaction Relationship: A Research in Erzurum Ataturk University,” *Am. Int. J. Contemp. Res.*, vol. 4, no. 1, pp. 100–117, 2014.
- [5] D. van Lierop and A. El-Geneidy, “Enjoying loyalty: The relationship between service quality, customer satisfaction, and behavioral intentions in public transit,” *Res. Transp. Econ.*, vol. 59, pp. 50–59, 2016, doi: 10.1016/j.retrec.2016.04.001.
- [6] Y. H. S. Al-Mamary, “Measuring information systems success in yemen: Potential of Delone and Mclean’s model,” *Int. J. Sci. Technol. Res.*, vol. 8, no. 7, pp. 793–799, 2019.
- [7] W. H. DeLone and E. R. McLean, “The DeLone and McLean model of information systems success: A ten-year update,” *J. Manag. Inf. Syst.*, vol. 19, no. 4, pp. 9–30, 2003, doi: 10.1080/07421222.2003.11045748.
- [8] F. D. Davis, “Perceived usefulness, perceived ease of use, and user

- acceptance of information technology,” *MIS Q. Manag. Inf. Syst.*, vol. 13, no. 3, pp. 319–339, 1989, doi: 10.2307/249008.
- [9] R. Nunkoo, T. D. Juwaheer, and T. Rambhunjun, “Applying the Extended Technology Acceptance Model to Understand Online Purchase Behavior of Travelers,” *Appl. Ext. Technol. Accept. Model to understand online Purch. Behav. Travel. Proc. 21st Int. Bus. Res. Conf. 10 - 11 June, 2013, Ryerson Univ.*, pp. 1–10, 2013.
- [10] W. Baswardono, D. Kurniadi, A. Mulyani, and D. M. Arifin, “Comparative analysis of decision tree algorithms: Random forest and C4.5 for airlines customer satisfaction classification,” *J. Phys. Conf. Ser.*, vol. 1402, no. 6, pp. 0–6, 2019, doi: 10.1088/1742-6596/1402/6/066055.
- [11] S. Hoon Hong, B. Kim, and Y. Gyu Jung, “Correlation Analysis of Airline Customer Satisfaction using Random Forest with Deep Neural Network and Support Vector Machine Model,” *Int. J. Internet, Broadcast. Commun.*, vol. 12, no. 4, pp. 26–32, 2020.
- [12] P. Surendran, “Technology Acceptance Model: A Survey of Literature,” *Int. J. Bus. Soc. Res.*, vol. 2, no. 4, pp. 175–178, 2012, doi: 10.1016/j.biortech.2015.06.132.
- [13] A. G. R. Amalia and A. N. L. I. Fahrudi, “The Relationship Between Perceived Ease of Use, Perceived Usefulness and Perceived Loss of Control with User Satisfaction in Mandatory Setting,” *Proc. 3rd Annu. Int. Conf. Public Bus. Adm. (AICoBPA 2020)*, vol. 191, pp. 171–173, 2021, doi: 10.2991/aebmr.k.210928.034.
- [14] I. Pancasila, S. Haryono, and B. A. Sulistyono, “Effects of work motivation and leadership toward work satisfaction and employee performance: Evidence from Indonesia,” *J. Asian Financ. Econ. Bus.*, vol. 7, no. 6, pp. 387–397, 2020, doi: 10.13106/jafeb.2020.vol7.no6.387.

- [15] D. Chakraborty, “Customer Satisfaction Towards Food Service Apps in Indian Metro Cities,” pp. 1–11, 2019, doi: 10.1177/2319714519844651.
- [16] D. Hidayat, A. Bismo, and A. R. Basri, “The Effect of Food Quality and Service Quality Towards Customer Satisfaction and Repurchase Intention (Case Study of Hot Plate Restaurants),” *Manaj. Bisnis*, vol. 10, no. 1, pp. 1–5, 2020, doi: 10.22219/jmb.v10i1.11913.
- [17] A. I. Ojo, “Validation of the delone and mclean information systems success model,” *Healthc. Inform. Res.*, vol. 23, no. 1, pp. 60–66, 2017, doi: 10.4258/hir.2017.23.1.60.
- [18] F. Li, H. Lu, M. Hou, K. Cui, and M. Darbandi, “Customer satisfaction with bank services: The role of cloud services, security, e-learning and service quality,” *Technol. Soc.*, vol. 64, no. July 2020, p. 101487, 2021, doi: 10.1016/j.techsoc.2020.101487.
- [19] S. Mahato and J. Goet, “Service Quality, Customer Satisfaction and Customer Loyalty in Nepalese Restaurant Industry,” *Int. J. Innov. Sci. Res. Technol.*, vol. 5, no. 12, pp. 1255–1261, 2020, [Online]. Available: [www.ijisrt.com](http://www.ijisrt.com)
- [20] B. B. and K.-C. Chang<sup>1</sup>, “The Influence Factors of Online Purchase on Customer Satisfaction in Mongolian Airlines,” *Int. Proc. Econ. Dev. Res.*, vol. 57, no. 1, pp. 80–85, 2012, doi: 10.7763/IPEDR.
- [21] E. Yeoh, J. Kim, and L. Chan, “Malaysian Low Cost Airlines: Key Influencing Factors on Customers’ Repeat Purchase Intention,” *World Appl. Sci. Journal Special Issue Tour. Hosp.*, vol. 12, no. January 2011, pp. 35–43, 2011.
- [22] I. P. Julianto and I. N. P. Yasa, “The Analysis of Technology Acceptance Model (TAM) on The Use of Accounting Information System,” vol. 69, no. Teams 2018, pp. 276–280, 2019, doi: 10.2991/teams-18.2019.48.

- [23] G. A. Putri, A. K. Widagdo, and D. Setiawan, “Analysis of financial technology acceptance of peer to peer lending (P2P lending) using extended technology acceptance model (TAM),” *J. Open Innov. Technol. Mark. Complex.*, vol. 9, no. 1, p. 100027, 2023, doi: 10.1016/j.joitmc.2023.100027.
- [24] Delone and McLean, “The quest for the dependent variable. Information Systems Research,” *Inf. Syst. Res.*, vol. 3, no. 1, pp. 60–95, 1992, doi: . <http://dx.doi.org/10.1287/isre.3.1.60>.
- [25] M. I. A. Sabah, U. K. Rashid, and J. Nasuredin, “The effect of information systems success model on job performance in Iraqi banks,” *Rev. Int. ...*, vol. 11, pp. 11–12, 2021, doi: 10.48047/rigeo.11.10.XXXX.
- [26] T. Mudzana and M. Maharaj, “Measuring the success of business-intelligence systems in South Africa: An empirical investigation applying the DeLone and McLean Model,” *SA J. Inf. Manag.*, vol. 17, no. 1, pp. 1–7, 2015, doi: 10.4102/sajim.v17i1.646.
- [27] K. Çelik and A. Ayaz, “Validation of the Delone and McLean information systems success model: a study on student information system,” *Educ. Inf. Technol.*, vol. 27, no. 4, pp. 4709–4727, 2022, doi: 10.1007/s10639-021-10798-4.
- [28] E. Daryanto, “The Influence of Information System Quality, Information Quality and Perceived Usefulness on User Satisfaction of Personnel Information Systems (Study at The Indonesian Army Crypto and Cyber Centre),” *J. Posit. Sch. Psychol.*, vol. 2022, no. 3, pp. 9814–9830, 2022, [Online]. Available: <http://journalppw.com>
- [29] R. A. Kumala, “The Impact of Information Quality and System Quality On User Satisfaction In Web-Based Information System.,” vol. 2, no. 2, pp. 1–12, 2014, [Online]. Available: <http://repository.ub.ac.id/117526/>
- [30] R. U. Mensah, Ishmael and Dei Mensah, “Effects of Service Quality and

- Customer Satisfaction on Repurchase Intention in Restaurants on University of Cape Coast Campus,” *J. Tour. Herit. Serv. Mark.*, vol. 4, no. 1, pp. 27–36, 2018, doi: <http://doi.org/10.5281/zenodo.1247542>.
- [31] J. Munusamy, S. Chelliah, and H. Mun, “Service quality delivery and its impact on customer satisfaction in the banking sector in Malaysia,” *Int. J. Innov. Manag. Technol.*, vol. 1, no. 4, pp. 398–404, 2010.
- [32] D. Ferreira1\*, “The Influence of Service Quality Dimensions on Customer Satisfaction at the Port Elizabeth International Airport,” 2020, [Online]. Available: doi: 10.9734/bpi/mono/978-93-90516-46-9%0A
- [33] A. Bere, “Exploring determinants for mobile learning user acceptance and use: An application of UTAUT,” *ITNG 2014 - Proc. 11th Int. Conf. Inf. Technol. New Gener.*, pp. 84–90, 2014, doi: 10.1109/ITNG.2014.114.
- [34] M. Steinbach, P. Tan, and V. Kumar, “*Introduction to data mining.*” Pearson Education, Inc, 2006.
- [35] K. B. and N. B. A. Jović, “A review of feature selection methods with applications,” *Comput. Biol. Med.*, vol. 112, no. May, pp. 25–29, 2015, doi: 10.1016/j.compbio.2019.103375.
- [36] L. Al Shalabi and Z. Shaaban, “Normalization as a Preprocessing Engine for Data Mining and the Approach of Preference Matrix,” *Proc. Int. Conf. Dependability Comput. Syst. DepCoS-RELCOMEX 2006*, pp. 207–214, 2006, doi: 10.1109/DEPCOS-RELCOMEX.2006.38.
- [37] G. Chandrashekar and F. Sahin, “A survey on feature selection methods,” *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, 2014, doi: 10.1016/j.compeleceng.2013.11.024.
- [38] A. Zia, M. Aziz, I. Popa, S. A. Khan, A. F. Hamedani, and A. R. Asif, “Artificial Intelligence-Based Medical Data Mining,” *J. Pers. Med.*, vol.

- 12, no. 9, 2022, doi: 10.3390/jpm12091359.
- [39] M. Compere, “A Case Study on Visual-Inertial Odometry using Supervised , Semi-supervised and Unsupervised learning methods,” *2019 IEEE Int. Conf. Artif. Intell. Virtual Real.*, pp. 203–2034, doi: 10.1109/AIVR46125.2019.00043.
- [40] P. R. Togatorop, M. Sianturi, D. Simamora, and D. Silaen, “Optimizing Random Forest using Genetic Algorithm for Heart Disease Classification,” *Lontar Komput. J. Ilm. Teknol. Inf.*, vol. 13, no. 1, p. 60, 2022, doi: 10.24843/lkjiti.2022.v13.i01.p06.
- [41] Z. H. Kilimci, “Evaluation of Customer Satisfaction about Telecom Operators in Turkey by Analyzing Sentiments of Customers through Twitter,” *Data Sci. Appl.*, vol. 3, no. 2, pp. 15–20, 2020.
- [42] Z. Zhang, “Introduction to machine learning: K-nearest neighbors,” *Ann. Transl. Med.*, vol. 4, no. 11, pp. 1–7, 2016, doi: 10.21037/atm.2016.03.37.
- [43] D. S. Nimaje and D. P. Tripathy, “Assessment of Fire Risk of Indian Coals Using Artificial Neural Network Techniques,” *Am. J. Min. Metall.*, vol. 3, no. 2, pp. 43–53, 2015, doi: 10.12691/ajmm-3-2-2.
- [44] A. Graves, “Framewise phoneme classification with bidirectional LSTM and other neural network architectures,” vol. 18, pp. 602–610, 2005, doi: 10.1016/j.neunet.2005.06.042.
- [45] S. Oh, J. Yoon, Y. Choi, Y. A. Jung, and J. Kim, “Genetic Algorithm for the Optimization of a Building Power Consumption Prediction Model,” *Electron.*, vol. 11, no. 21, pp. 1–12, 2022, doi: 10.3390/electronics11213591.
- [46] H. M and S. M.N, “A Review on Evaluation Metrics for Data Classification Evaluations,” *Int. J. Data Min. Knowl. Manag. Process*,

- vol. 5, no. 2, pp. 1–11, 2015, doi: 10.5121/ijdkp.2015.5201.
- [47] X. J. Lim, E. C. X. Aw, and K. G. C. Teoh, “Factors influencing repurchase intention in online shopping context: The mediating role of satisfaction,” *J. Appl. Struct. Equ. Model.*, vol. 2, no. 1, pp. 29–43, 2018, doi: 10.47263/jasem.2(1)04.
- [48] A. Amin and T. C. T. Chandra, “the Effect of Information Quality and Innovation on Customer Loyalty Mediating By Customer Satisfaction,” *Int. Conf. Bus. ...*, pp. 1213–1231, 2022, [Online]. Available: <http://61.8.77.171/index.php/icobuss1st/article/view/302>
- [49] R. Kumar, “Impact of Demographic Factors on Consumer Behaviour - A Consumer Behaviour Survey in Himachal Pradesh,” *Glob. J. Enterp. Inf. Syst.*, vol. 6, no. 2, pp. 35–47, 2014, doi: 10.15595/gjeis/2014/v6i2/51844.
- [50] S. H. Wang and Y. L. Pang, “Demographics and customer satisfaction,” *EPRA Int. J. Environ. Econ. Commer. Educ. Manag. J.*, vol. 8, no. 9, pp. 8–13, 2021, doi: 10.36713/epra0414.
- [51] P. Jiang, H. Wu, W. Wang, W. Ma, X. Sun, and Z. Lu, “MiPred: Classification of real and pseudo microRNA precursors using random forest prediction model with combined features,” *Nucleic Acids Res.*, vol. 35, pp. 339–344, 2007, doi: 10.1093/nar/gkm368.
- [52] F. D. Davis, “Perceived Usefulness, perceived ease of use and user acceptance of information technology,” vol. 13, no. 3, pp. 319–340, 1989.
- [53] A. F. Shodiq, S. Hidayatullah, and Y. T. Ardianto, “Influence of Design, Information Quality and Customer Services Website on Customer Satisfaction,” *Artic. Int. J. Sci. Eng. Res.*, vol. 9, no. 12, pp. 746–750, 2018.

## الخلاصة

تتنافس الشركات مع بعضها البعض لتقديم أفضل الخدمات التي يمكن أن ترضي عملائها. يمثل رضا المستخدمين جانباً رئيسياً تهدف الشركات إلى تحقيقه لأنه يرسم مستقبل الشركة نحو النجاح أو الفشل. ومع ذلك، من الضروري فهم ما يمكن أن يؤدي إلى رضا العملاء عن شركة أو تقنية أو منتج معين. يهدف هذا البحث إلى (1) التنبؤ برضا العملاء عن شركات الطيران، (2) تحديد الميزات الأكثر فعالية في رضا العملاء، (3) تقديم إطار معدل لرضا العملاء، و(4) تحسين دقة التنبؤ باستخدام عدة تقنيات مختلفة.

تم استخدام مجموعة البيانات الخاصة بـ 129880 عميلاً في هذا البحث. ويشمل الميزات الديموغرافية وتصورات عملاء شركات الطيران. تم تطبيق أربع تقنيات لاستخراج البيانات للتنبؤ برضا العملاء، وتلك التقنيات هي الغابة العشوائية (RF) وأقرب جار (KNN)، والإدراك متعدد الطبقات (MLP)، والذاكرة الطويلة قصيرة المدى (LSTM). تم اتباع العديد من الخطوات لتعزيز دقة التنبؤ للتقنيات المنفذة مثل التعامل مع القيم المفقودة والقيم المتطرفة، وتطبيع قيم الميزات، وإنشاء ميزات جديدة. علاوة على ذلك، تم تطبيق تقنيات اختيار الميزات والتحسين الجيني. لقياس أداء المصنفات، تم تطبيق ثلاث مراحل للتنبؤ تعتمد على: (1) اثنان وعشرون ميزة أصلية لمجموعة البيانات، (2) خمسة وعشرون ميزة تتضمن الاثنان والعشرين الأصلية والميزات المولدة وهي جودة الخدمة وجودة المعلومات والمشاركة، و(3) اثنتي عشرة ميزة الميزات تتضمن ثلاث ميزات مولدة فضلاً عن تسع ميزات أصلية. تم تنفيذ المرحلة الثالثة للتنبؤ لغرض تقليل تعقيد النموذج.

تم إجراء التنبؤ بناءً على عدد الميزات المضمنة. أولاً، يستخدم النموذج المقترح الميزات الأصلية الاثنان والعشرين حيث كان أداء RF، KNN، MLP، LSTM هو 96، 95.2، 95.7، 95.7 على التوالي. ثانياً، مع الميزات الأصلية والميزات الثلاثة الجديدة، كانت دقة التقنيات الأربعة RF، KNN، MLP، LSTM هو 97، 96، 95.2، و96 على التوالي. أخيراً، استناداً إلى اثنتي عشرة ميزة فقط تتضمن الميزات التي تم إنشاؤها وميزات أخرى لم يتم استخدامها في عملية التوليد، كان أداء النموذج لـ RF و KNN و MLP و LSTM هو 95، 93، 91.2، 91.9 على التوالي. ووفقاً لهذه النتائج، يمكن ملاحظة أن مصنف RF تفوق على جميع المصنفات الأخرى.

ويمكن الاستنتاج أن التعامل مع اثنتي عشرة ميزة يمكن أن يقلل من تعقيد النموذج ويؤدي إلى دقة تنبؤ معقولة. يمكن أن تساعد نتائج البحث شركات الطيران على تحسين خدماتها والاستجابة لاحتياجات العملاء.



جمهورية العراق

وزارة التعليم العالي والبحث العلمي

جامعة بابل

كلية تكنولوجيا المعلومات - قسم البرمجيات

# تنبؤ رضا العملاء عن شركات الطيران: تطبيق تقنيات التعلم الآلي والعميق

رسالة مقدمة إلى

مجلس كلية تكنولوجيا المعلومات - جامعة بابل كجزء من متطلبات

نيل درجة الماجستير في تكنولوجيا المعلومات / البرمجيات

من قبل

شهد حسين علي

بإشراف

أ.م.د أحمد حبيب العزاوي

٢٠٢٣ م

١٤٤٥ هـ