International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

MULTIFACETED ANALYSIS OF CUSTOMER FEEDBACK ON AIRLINE SERVICES.

[1]Shirin Niyaz Al Ajmi , [2]Ferddie Quiroz Canlas, [3]Sarachandran Nair

[1]*Department Of Computing Muscat College* Muscat, Oman
Shirinajmimc98@Gmail.Com

[2]*Department Of Computing Muscat College* Muscat, Oman
Ferddie@Muscatcollege.Edu.Om

[3]*Department Of Computing Muscat College* Muscat, Oman
Saran@Muscatcollege.Edu.Om

Abstract

The Proposed Multifaceted Analysis Of Customer Feedback On Airline Service Is A Machine Learning Model To Predict Whether The Customer Is Pleased With The Services Provided By The Airline Or Dissatisfied, As Well As Whether The Customer Is Loyal To The Airline Or Not, So The Airline Can Strategize What Service They Need To Adapt For Each Of The Unique Customers To Generate A Larger Number Of Satisfied Customers. This Aim Was Operated Based On The Classification Model With The Help Of Imputation Which Is The Estimation Of The Missing Values Through Assumptions Based On Statistical Methods. Nonetheless, Multiple Classification Models, Including K-Nearest Neighbour, Logistic Regression, Naïve Bayes, And Random Forest, Were Compared To See Which Had The Highest Accuracy Score And Precision, Allowing The Best Model To Be Built. In Addition, The User May Complete A Short Survey In Which The Model Predicts Whether The Consumer Is Content Or Dissatisfied, As Well As If The Customer Is Loyal Or Disloyal. The Model Was Very Efficient To Find Out Customer Satisfaction And Loyalty To The Airline

Keywords— *Classification Model, CRISP-DM Process, K-Nearest Neighbour, Logistic Regression, Naïve Bayes, Random Forest, Gridsearchcv, Machine Learning.*

I. INTRODUCTION

Covid-19 Has Caused An Indirect Decrease In Airline Operations Worldwide By Prohibiting Most Global Air Travel. Almost All Airlines Across The World Experienced Significant Losses Because Of Being Unable To Conduct Airline Services, Which Are Their Primary Source Of Revenue. Nonetheless, Once The Pandemic Has Settled, The Demand For Air Travel Is Expected To Increase As People Resume Their Vacation And Travel Habits. One Of The Issues Of Global Airlines Is Keeping Their Customers Satisfied With The Travel Experience They Provide. Customers That Are Dissatisfied Result In Fewer Passengers And Lead To Huge Losses For The Company, Which Will Therefore Considerably Decrease Their Profit. For That Reason, This Problem Requires A Solution, As It Is Crucial That Customers Have A Pleasant

The Authors Would Like To Acknowledge Muscat College For The Funding Of This Research.

Received: August 11,2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Experience As It Directly Correlates To The Airline's Financial Well-Being. Moreover, Analyzing Every Aspect Of Service Quality Is Unquestionably One Of The Most Important Features Required To Get Better Ratings And Please Customers. As A Result Of This Analysis, The Airline May Improve By Adapting The Experience And Services To Fit Customers' Needs.

Air Travel Has Enormous Social Advantages Which Can Be Seen Easily, As It Enhances People's Quality Of Life By Widening Their Entertainment, Leisure, And Cultural Experiences. It Provides A Worldwide And Relatively Inexpensive And Accessible Method Of Visiting Distant Friends And Relatives And Going On Holidays. Due To Its Large Effects On Economies And People In General, Countless Airline Companies Are Initiated Every Year, Showcasing An Increased Level Of Competition Between Companies. To Differentiate Itself From The Masses Of Other Airlines, An Airline Must Offer Good Customer Service Before, During, And After Landing To Be Well Appreciated By Their Customers And Deliver A Pleasant Experience To The Consumers Of Their Service. Therefore, It Is Obvious That Successful Airline Companies Consider Feedback, Customer Loyalty, And Customer Satisfaction As Crucial Components Of Their Airlines To Improve Customer Value.

As Previously Stated, Researching, Studying, And Investigating Every Component Is Critical For All Airline Companies To Provide Better Service And Increase Customer Happiness, Therefore Feedback Is Required. Even Though There Are Various Methods For Manually Acquiring Customer Feedback, There Are Many Negatives Which Might Discourage Airline Companies, Such As The Fact That It Takes A Lot Of Work, Money, And Time To Ask The Client For Their Feedback Or Send Them A Questionnaire. Moreover, Many Customers Might Not Even Fill Out Feedback Forms Which Could Prove That Feedback Collection Is A Waste Of Time For The Company Itself. Furthermore, The Information Collected From The Customer May Not Be Accurate Or Up To Date.

This Paper Attempts To Develop A Machine-Learning Model Which Can Predict Customer Satisfaction And Loyalty And Help Airlines By Increasing Customer Loyalty And The Company's Profit. Through The Use Of Machine Learning, The Data Set That Was Acquired Is Used Alongside An Algorithm To Predict Future Customer Satisfaction And Loyalty While Gradually Increasing The Level Of Accuracy. The Proposed Machine Learning Model Was Done Using Python (Jupyter Notebook), Microsoft Excel For Comparison Of The Tables And Finally Different Python Libraries (Mainly Sklearn)

II. Review Of Related Studies

A. Invistico Airlines - Understanding Customer Satisfaction.

This Study Was Done Recently By Swathi Mudhelli, Naby Syeda, Alejandra Miramontes, And Meixi Sun. The Main Aim Is To Identify The Factor Needed To Influence The Satisfaction Variable Target. The Methodology Used Was Undertaking Exploratory Data Analysis To Recommend Different Ways To Address Satisfaction And Dissatisfaction. A Brief Introduction To The Dataset That Was Taken From An Open-Source Website 'Kaggle' That Had The Previous Customers' Information. They Used Logistic Regression And Random Forest For The

Received: August 11,2025

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Variable Importance. Furthermore, They Used T-Tests And Correlation Tests To Check The Relationship And The Importance Of The Features. Additionally, They Checked Their AIC Score And Used The Net Promoter Score To Measure Their Overall Satisfaction. Lastly, They Split The Data Using 80% Training And 20% Testing [1].

The Advantages Include; Categorizing The Dataset Into Two Groups (Personal Factors And Flight Factors) Which Made It Easier And More Simple For The Researchers To Analyse And The Airlines To Improve Upon. Moreover, They Used Different Visualisation Techniques And Organized A Detailed Interpretation Of The Five Most Important Variables. Additionally, After They Listed The Variables They Gave And Detailed Explanation Of How They Searched The Variable And How It Impacts The Satisfaction Level Of The Customer And Recommendations On What The Airline Can Do To Enhance The Customer's Experience. On The Other Hand, They Did Not Build A Model Or Do Any Validation Testing, Since They Only Focused On The Analysis. To Add To That, They Only Had One Target Variable, But They Mentioned That They Want To Understand The Relationship Between Customer Satisfaction And Customer Loyalty And What Is Their Effect On Their Future Work.

Both Studies Used The Same Splitting Data (80% Training And 20% Testing), Used Logistic Regression, Random Forest And A Correlation Method. Although Understanding Customer Satisfaction Did Not Specify Which Technique They Used, The Multifaceted Analysis Of Customer Feedback On Airline Services Used The Spearman Rho Correlation. Furthermore, This Study Used The Chi-Square Test Instead Of The T-Test, Where Also Validation Testing Has Been Taken Into Place And Has Two Target Variables Instead Of One.

B. Predicting Satisfaction Of Airline Passengers With Classification

This Thesis Was Done In 2020 By Tan Pengshi Alvin To Predict The Satisfaction Of Airline Passengers. Their Methodology Was To Use Five Different Models To Predict The Satisfaction Of The Airline Passengers, Which Are K-Nearest Neighbours, Logistic Regression, Gaussian Naïve Bayes, Decision Tree And Finally Random Forest. Furthermore, To Find The Best Model They Used AUC, Precision And Recall. In Addition, They Used Correlation, Heat Map, KDE (Kernel Density Estimation) And LASSO Regression To Do The Feature Selection And Check What Feature Needs To Be Dropped. For Their Model Selection, They Split The Data Using 80% Training And 20% Testing, To Add To That Cross-Validation Was Done Using Five Folds To Find The Best Hyper-Parameter With The Help Of 'Gridsearchcv' [2].

One Of The Most Important Advantages Of This Study Is That They Created An Excellent Web Interface That Contains The Exact Percentage Of The Satisfaction Level Of That Specific Customer (As Can Be Seen In Figure 1) With The Help Of The Flask App. They Also Gave A Very Detailed Introduction To The Dataset And What It Contains, Which Is From The Open Source Website 'Kaggle', And Have Re-Named Some Of The Variables For Better Understanding. Moreover, They Have Identified The Most Vital Factor That Leads To High Customer Satisfaction. Besides, They Have Used Validation With The Help Of A Random Forest Classifier To Decide The Most Effective Probability Level. To Add To That, They Used

Received: August 11,2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Exploratory Data Analysis To Conclude What Variables May Influence The Satisfaction Target. Finally, They Have Recommended Ways To Improve The Airlines' Services, Such As ‘Inflight Wi-Fi Service’ And ‘Ease Of Online Booking’. They Also Mentioned That They Want To Work On Clustering, Natural Language Processing And Building A Recommendation System For Their Future Work

In Contrast, They Don’t Have User Input And They Only Have One Target Variable (Customer Satisfaction) While The Multifaceted Analysis Of Customer Feedback On Airline Services Study Used Two (Customer Satisfaction And Customer Type).

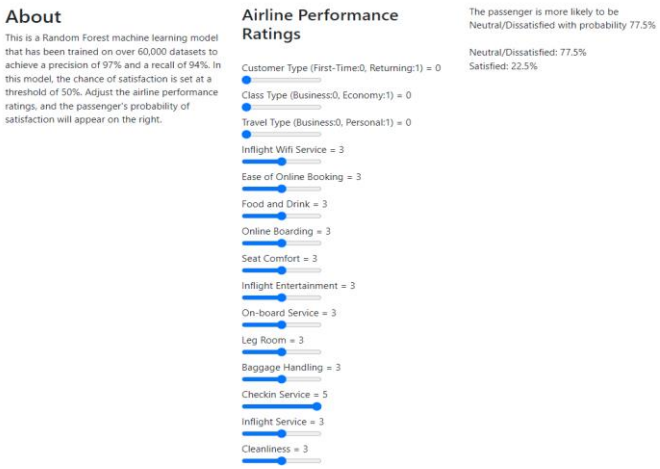

Fig. 1. Web Interface Of The Predicting Satisfaction Of Airline Passengers.

C. Investigating Airline Passenger Satisfaction: Data Mining Method

This Study Was Done By Tri Noviantoro And Jen-Peng Huang. The Main Goal Of This Research Is To Investigate Airline Passenger Satisfaction In Order To Recommend Four Vital Services To Improve The Airline’s Quality And Ability In Gaining High Customer Satisfaction (Which Include Online Boarding, Inflight Wi-Fi Service, Baggage Handling And Inflight Entertainment). The Methodology Used To Do This Is The Data Mining Method And Feature Selection Method To Select The Top Features With The Most Significance To The Dataset, Which Is The Wrapping Feature. This Approach Is Used To Improve Its Efficiency But It Took A Long Time To Calculate. Moreover, Validation Is Used To Determine The Accuracy Level [3].

An Important Merit Is That They Used Ten Different Classification Models Which Are Decision Tree, Random Forest, Gradient Boosting Tree, K-Nearest Neighbour, Naïve Bayes, Logistic Regression, Neural Network And Support Vector Machine And Mentioning The Source Of The Dataset That Was Taken From (The Open-Source Website ‘Kaggle’) And The Name Of The Company The Data Was Collected From. Furthermore, They Analysed The Data By Comparing The Different Variables And Seeing What The Resulting Relationship Between Them Is. This Was Done After Filtering Data That Doesn’t Contain Missing Values And Statistical Description Of The Variables, Finally, They Used Tenfold Cross Validation To Split

Received: August 11,2025

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The Data Into Training And Testing And Then Visualised The Predictive Model Using A ROC Graph.

Some Disadvantages Include That They Used Only One Target Variable (Customer Satisfaction) And They Didn't Mention The Ratio Of The Data Splitting.

D. Predicting Airline Customer Loyalty By Integrating Structural Equation Modeling And Bayesian Networks

This Study Is Written By Kattreeya Chanpariyavatevong, Waritwipulanusat, Thanapong Cham-Pahom, Sajjakaj Jomnonkwao, Dissakoon Chonsalasin And Vatanavongs Ratanavaraha. And It Aims To Find The Important Factors That Influence Airline Loyalty. The Methodology Used Includes Basiyan Network And Structural Equation Modelling (SEM) And They Researched The Relationship Between Them. They Also Used The Chi-Test To Measure The Significance Level And The Validated Structural Model To Determine The Factors That Impact Customer Loyalty [4].

The Data Collection Was Extremely Detailed As It Was Conducted Through Face-To-Face Surveys And 15-Minute Interviews At The Terminal Building Of Thailand Airport (From March To May 2019). The Interviews Conducted Were In All Four Parts Of Thailand (Southern, Northern, Central And Northeastern Thailand) And Had 400 Participants From Each Region. The Survey Was Also Very Thorough As Its Rating Was From 1-7 (One Strongly Disagreed And Seven Strongly Agreed). Furthermore, They Gave A Fully Detailed Analysis Of The Methodologies They Used And Recommended Strategies To The Airline On How They Can Increase Customer Loyalty And What Factors Influence It The Most.

However, They Only Explore The Factors Of Customer Loyalty And Didn't Create Any Classification Model While The 'Multifaceted Analysis Of Customer Feedback On Airline Services' Study Built A Classification Model Using Random Forest And Tested Four Different Models (Logistic Regression, K-Nearest Neighbor, Naïve Bayes And Random Forest). Moreover, It Has User Input For Validation Purposes.

E. Airline Recommendation Prediction Using Customer-Generated Feedback Data

This Research Was Done By Praphula Kumar Jain, Rajendra Pamula, Dilip Sharma, Sarfraj Ansari, And Lakshmibai Maddala. And It Aims To Predict The Reviews That Were Collected From Online Sites And Analyse Them. The Data Was Taken From The Online Site 'Airlinequality.Com'. This Dataset Was Collected From January 2011 To December 2015, And It Consists Of Four Different Airline Classes (Economy, Premium Economy, Business And First Class). Then They Performed Exploratory Data Analysis To Compare The Economy And Business Classes Since They Had The Most Information About These Two Classes. The Dataset Was Collected From 85 Different Countries And The Reviews Were From People Who Have Travelled With 186 Different Airlines. Moreover, They Used Machine Learning And Other Techniques Such As Tokenization, Lemmatization And The Help Of The Natural Language Toolkit. Finally, They Built Different Models (K-Nearest Neighbour, Support

Received: August 11,2025

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Vector Machine And Decision Tree) And Evaluated Their Implementation So That They Can Accurately Understand The Recommendations From The Users; Reviews [5].

The Strengths Include That They Mentioned That All The Reviews In The Dataset Were In English And There Were No Missing Values In It. They Also Evaluated Their Methodology By Mentioning That They Used Exoplanetary Data Analysis To Find Out What Priorities Different Classes Have. They Wrote Positive And Negative Recommendations For The Economy And Business Classes, And Concluded That The Support Vector Machine Had Better Results Overall With A Very High Precision Rate.

However, They Did Not Have Sufficient Data To Investigate The Premium Economy And First Class. Additionally, They Did Not Build An Interface. One Of The Major Differences Between This Study And 'Multifaceted Analysis Of Customer Feedback On Airline Services' Is That They Used Natural Language Processing And Machine Learning Techniques To Analyse The Reviews And The Customers' Sentiments On The Different Classes Of Their Airline.

F. A Deep Learning Approach To Analyze Airline Customernpropensities: The Case Of South Korea

This Investigation Was Held In South Korea And Was Written By So-Hyun Park, Mi-Yeon Kim, Yeon-Ji Kim And Young-Ho Park. The Aim Is To Find The Relationship Between Different Aspects That Effects And Stimulate Customer Churn Risk And Satisfaction By Studying And Analyzing The Air-Line Customer Data, And This Is Done By Applying Deep Learning Methods To The Survey Collected From The Users Who Have Used Mostly Korean Aeroplanes. Moreover, Analyze And Reviews The Social Services As Well As The Opinions And Points Of View Of The Cabin Crew And The Passengers Using The Aircraft, On Airline Customer Learning And Tendencies [6].

The Methodology They Utilized Was To Collect Data From Travellers Who Have Travelled At Least Once Between 2018 And 2022. Furthermore, They Divided The Data Into 80% Training And 20% Testing, As Well As Conducting Numerous Stages Such As Combining The Dataset, Cleaning Undesirable Data (Re-Moving Replies Who Did Not Complete Their Survey Correctly), And Ultimately Doing Feature Selection Using Pearson's Correlation. Therefore, They Used A Variety Of Machine Learning And Deep Learning Models To Determine The Best Model To Predict Airline Consumer Behaviors, Including K-Nearest Neighbors, Decision Trees, Random Forests, Gradient Boosting, Convolutional Neural Networks, And CNN Long Short-Term Memory Networks. Finally, Cross-Validation Was Carried Out To Identify The Ideal Hyperparameters Of The Machine Learning And Deep Learning Models.

The Benefits Include Providing A Detailed Explanation And Specifications About The Data And How It Is Collected, Where There Were 50 Questions In The Survey And The User Responded To All Of Them That Related To The Physical And Social Environment Of The Airlines, Brand Experience And Loyalty, And Customer Satisfaction. Moreover, A Detailed Description Of The Methodology Was Given. Similarly, Different Types Of Visualization

Received: August 11,2025

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Have Been Used To Explain Different Aspects Of The Studies, Such As Heatmaps, And Bar Charts. Lastly, They Analyzed And Evaluated The Confusion Matrix (By Calculating The Accuracy, Precision, Recall And F1 Score) And Found That The Deep Learning Model (CNN Long Short-Term Memory Networks) Is More Accurate At Predicting The Customer Churn Risk And Satisfaction, Likewise, It Was The Best Model.

One Of The Important Differences Between This Research And The 'Multifaceted Analysis Of Customer Feedback On Airline Services' Is The User Input. Moreover, There Wasn't Any Analysis Of The Services That Need To Be Improved.

G. Comparative Analysis Of Decision Tree Algorithms: Random Forest And C4.5 For Airlines Customer Satisfaction Classification

The Study Was Written By W. Baswardono. The Goal Is To Compare The Analysis Of The Decision Tree Algorithms Between Random Forest And C4.5 For The Airline Customer Satisfaction Dataset That Was Taken From The Open-Source Website Kaggle (U.S Airline Passenger Satisfaction). This Has Been Achieved By Evaluating The Confusion Matrix And Concluding The Accuracy, Precision, Recall And Area Under The Curve [7].

The Main Methodology They Used Was To Analyze The Dataset Using Different Ratios Of Splitting The Data, Such As 80% Training And 20% Testing, 90% Training And 10% Testing, And Finally 70% Training And 30% Testing, All With Ten-Fold Cross-Validation. The Performance Of Comparison Between Each Algorithm Has Taken Place To Find Which Algorithm Is Best To Use For The Dataset. Furthermore, The KDD Process Has Been Used To Find Out Meaningful, Practical, And Understandable Patterns In Large And Complex Data Sets (This Process Consists Of 6 Stages Which Are Data Selection, Data Processing, Data Cleaning, Data Transformation, Data Mining And Evaluation.).

The Positives Are That They Gave A Full Detailed Explanation Of The Data And From Which Source It Was Taken, Which Is (Kaggle), And The File Contains A Dataset For Customer Satisfaction Along With The Records. Furthermore, They Stated That The Data Will Be Tested To Determine The Level Of Accuracy Of Customer Satisfaction, So Classification Techniques Such As Data Collection, Data Processing, And Data Mining Science Will Be Used To Obtain Important Information, And Then Processing Techniques Will Be Utilized To Produce A Better And Enhanced Quality Of The Data, And Finally, Data Cleaning To Remove Noisy And Inconsistent Data. Similarly, They Calculated The Results Of Validation And Testing (The Level Of Accuracy, Precision, Recall, Error Classification, And The Area Under The Curve), Which They Compared With The Selected Classification Model, This Has Been Done To Confirm That The Results They Concluded Are Correct. In Addition, They Concluded That The Random Forest Classifier Has The Best Performance And Gives Better Results Than The C4.5 Algorithm. Furthermore, Many Visualizations Such As Tables Were Used To Compare And Display The Data And Mentioned That They Specifically Used The Two Classifiers Since It Is The Two Most Popular And Both Mentioned Algorithms Can Be Used With Large Datasets. Finally, They Suggested Several Future Works That Can Be Performed, Such As

Received: August 11,2025

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Clustering, And The Usage Of Different Classification Models With Different Hyperparameters.

On The Other Hand, A Limitation Is That They Did Not Mention What Target Variable They Used For Their Investigation. One Main Difference Is That The Current Study Is A Comparison Between 2 Different Algorithms Using Different Ratios In Splitting The Data With 1- Fold Cross-Validation And Comparing To Find Out The Best Model.

H. Predicting Airline Customer Satisfaction Using K-NN Ensemble Regression Models

The Paper, Written By V. García, R. Florencia-Juárez, J. P. Sánchez-Solís, G. Rivera-Zarate, R. Contreras-Masse Aims To Achieve An Experimental Study Using A Real-Data And The Sample That Was Taken Was Approximately 130000 Samples From The Airline Companies Located In The USA To Predict Customer Satisfaction. Whereas, The Different Methodology Has Been Used For Instance Analysis Of The Performance Of The K-Nearest Neighbour Model For Regression As A Base Classifier Of The BAGGING (Bootstrap Aggregat-ING). Moreover, They Used 5-Fold Cross-Validation To Avoid Inaccurate Performance And Use The Root Mean Square Error And Absolute Error To Measure And Evaluate The Regression Problem Performance [8].

The Main Strength Is That They Gave A Full Detailed Explanation Of The Data, For Instance, The Data Was Conducted With A Very Large Group Of The Dataset That Was Taken From Airline Companies In The USA Where It Was Collected In 2014 From Customer Satisfaction Questionnaires And One Target Variable. Finally, They Mentioned Future Work Where They Would Like To Expand Their Analysis And Study By Using Feature Selection And Removing Any Noisy, And Unimportant Data. On The Other Hand, They Didn’t Clean The Data Before Analyzing It, Besides, They Only Had One Target Variable.

I.Analysis Of Hotel Services By Their Symmetric And Asymmetric Effects On Overall Customer Satisfaction: A Comparison Of Market Segments

The Research Was Done By Özgür Davrasa And Meltem Caber. The Primary Goal Was To Examine And Investigate How The Performance Of Hotel Services, Both Symmetrically And Asymmetrically, Influences Total Customer Satisfaction. The Methodology That Was Used Is The Chi-Test To Find Out The Significant Relationship Between The Variables With The Help Of Eigenvalues. Finally, The Records With An Outlier More Than Or Less Than Three Standard Deviations From The Mean Base Were Removed [9].

They Mentioned A Full Detailed Explanation Regarding The Data Which Was Taken From Different 5-Star Hotels In Three Different Countries (Turkey, Germany, And Russia), The Customers Who Completed The Survey Were Informed About The Purpose Of The Study, And The Researchers Obtained The Customer's Permission. Similarly, They Stated That By Doing This, They Hoped To Better Understand The Services And Determine Which Services Should Be Prioritized To Improve Customer Happiness At The Hotel. They Also Employed A Variety Of Visuals To Describe The Data, Such As Line Graphs And Tables. Furthermore,

Received: August 11,2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

They Created A Statistical Description Of The Data And Examined It In Order To Assess The Hotel's Customer Service And How It Affects Customer Satisfaction. Lastly, They Identified Certain Services That May Be Improved.

J. Perceived Quality, Customer Satisfaction, And Customer Loyalty: The Case Of Lexus In Taiwan

Chwo-Ming Joseph Yu, Lei-Yu Wu, Yu-Ching Chiao, And Hsingshia Tai Conducted The Research Where Its Preliminary Purpose Is To Examine All Aspects Of Customer Satisfaction Connected To Toyota's Lexus Vehicles Using The Fornell Et Al., Customer Satisfaction Software. [10].

The Methodology Was That They Analyzed A Group Of 320 Taiwan Lexus Owners, And Those Who Used The Lexus Car For More Than 7 Months. They Utilized The Structural Equation Model, With The Help Of LISREL Software To Illustrate If The Quality Has A Direct Or Indirect Effect On Overall Customer Satisfaction. Furthermore, They Removed All The Not Answered Questions Where They Assumed Them To Be Missing Values, Therefore, The Final Analysis Of Data Was 319 Reviews On Lexus Car Owners In Taiwan Who Have Been Using Their Cars For More Than 7 Months. Furthermore, They Used CSI Model. Finally, They Used Standard Error And T-Value To Find The Relationship Between Customer Satisfaction To The Company.

Taiwanese Researchers Studied Over 1,000 Consumer Complaints And 879 Surveys Acquired By Word-Of-Mouth. The Strengths Are That They Did Not Publish The Owner's Business And Personnel Details In Their Data. Future Projects Have Located US-Based Research On Customer Complaints In Taiwan And Undertaken Exploratory Data Analysis For Improved Understatement.

III. METHODOLOGIES

A. CRISP-DM Methodology

Data Mining Projects Are About Learning Patterns From Data, So Predictions Can Be Made. A Data-Driven Project Is Very Different To A Standard Software Engineering Project And Many Other Projects Involving Design And Implementation. Standard Software Projects Begin With A Specification That Sets Out Precisely What The Software Will Do. The System Will Then Be Implemented And Tested Against That Specification Before It Is Signed Off And Put Into Use. One Difference Between A Data-Driven Project And A Full Software Product Is That The Individual Is Producing A Piece Of Functionality That Is Integrated Within A Larger System. Another Crucial Difference Is That The Results Are Partially Determined By The Data Itself. As A Result, It Might Be Difficult To Predict What An Effective Utilization Would Look Like Ahead Of Time. This Implies That Defining The Project's Goals From The Beginning Is Quite Challenging. Typically, This Means That At The Outset Of A Project, An Individual Can Only Describe In General Terms What They Want To Achieve. For Example, Being Able To Predict If The Customer Is Satisfied With The Current Services Of The Airline Or Not.

Received: August 11,2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The CRISP-DM (Cross Industry Standard Process For Data Mining) Standard Is Known As A Process Model That Offers A Framework To Follow When Running A Data Mining Project That Is Independent Of Both The Industry And The Technology Utilized. Other Standards Exist, But CRISP Is One Of The Most Well-Established And As Can Be Seen From Figure 2, A Survey That Is Taken From The 'KD Nuggets' Poll, CRISP-DM Is The Latest Top Process Used And Remains One Of The Most Popular. Figure 3 Demonstrates That The Framework Consists Of Six Stages, And It Is Not Linear, Whereas It Is A Circular Iterative Process That Attempts To Offer An Organized Method By Making The Aim Clear Before Starting The Project. Thus, It Is Possible To Step Back And Forth Between The Stages. The Most Important Movements Are Shown By Arrows In Figure3. The CRISP-DM Process Consists Of Six Steps, Which Consist Of Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation And Finally Deployment. The CRISP-DM Model Is Not About The Business Of Preparing The Data And Doing Analysis, It Is Very Much Driven By Research Or Business Objectives To Guide That Data And Analytical Processes [11].

One Of The Advantages Is That The Phases Can Be Executed In Any Sequence, However, It May Be Needed Sometimes To Return To Earlier Steps. This Technique Serves As A Project Guide Where It Describes The Different Phases Of A Project, The Activities Associated With Each Phase, And The Links Between Tasks. The Steps In This Methodology Are Used And Followed In Developing The System.

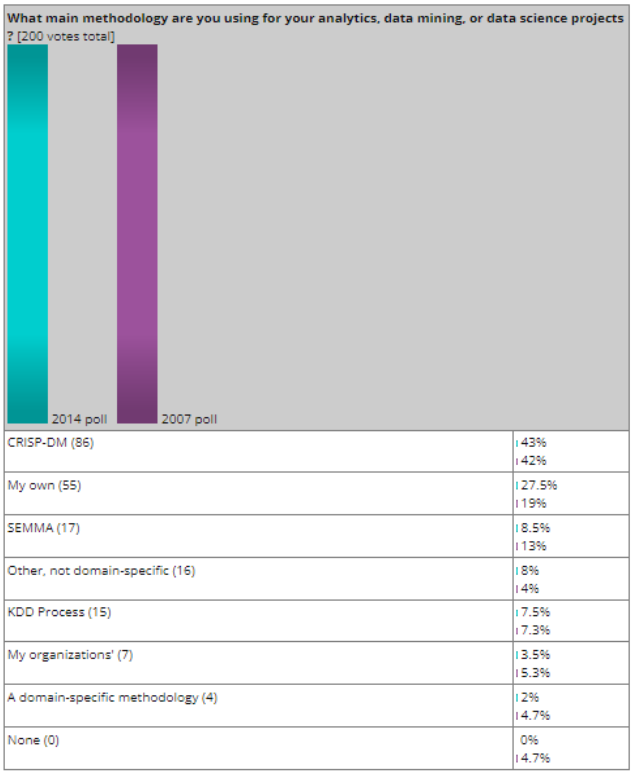

Fig. 2. Survey For Methodology Analytics *[30]*.

Received: August 11,2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

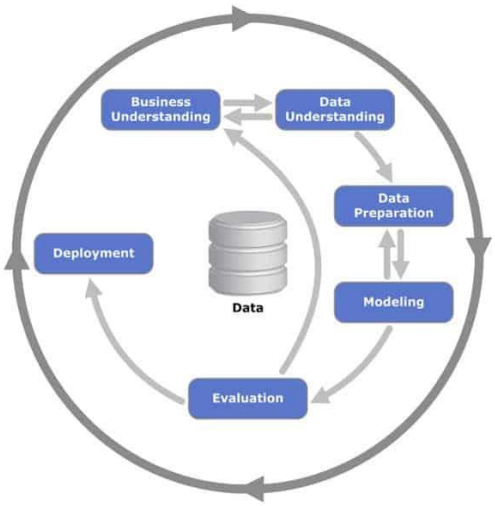

Fig. 3. : Cross Industry Standard Process For Data Mining *[31].*

B. Data Understanding

The Second Stage Is Data Understanding, In Which The Individual Should Grasp What And How Much Data Is Accessible, As Well As How Convenient It Would Be To Obtain More. What Variables Are Present, What Data Is Available, And How Does The Data Correlate To The Problem? What Is The Target Variable? The Individual Might Plot Some Basic Statistics To Grasp What The Dataset Might Look Like. At This Stage, The Individual Might Need To Backtrack To The CRISP-DM Process If It Doesn't Look Like The Data Is Suitable For The Problem. As Mentioned Before, The Dataset Was Acquired From The Open-Source Website 'Kaggle'. The Exact Name Of The Corporation Was Not Revealed For A Diverse Range Of Reasons, Including The Airline's Privacy. The Dataset Includes Customers Who Have Flown With This Airline And Their Opinions On Different Flight Questions.

The Dataset Includes More Than 3000 Records, And 23 Distinct Parameters With 2 Target Variables (Satisfaction And Customer Type) And 14 Of These Features Are Entries Of Passengers' Ratings (0-5) Of Various. All The Collected Data From The Survey Is About The Various Services Provided By The Airline Which Gives Vital Information That May Assist Airlines To Enhance Overall Customer Happiness.

1) Satisfaction (Target Variable1). It Is A Type Of String. This Variable Contains The Happiness Level Of The Customer For The Airline. It Consists Of Satisfaction And Dissatisfaction Options, And

2) Gender. This Variable Specifies The Gender Of The Passenger. It Consists Of Males And Females And The Field Type Is String.

3) Customer Type (Target Variable 2). This Variable Shows If The Customer Is Loyal Or Not. Thus, It Consists Of Loyal Customers And Disloyal Customers. And The Type Of Field Is String.

Received: August 11,2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

4) Age. This Field Shows The Actual Age Of The Passenger. It Consists Of A Number And The Data Type Is Integer.

5) Type Of Travel. This Field Shows The Purpose Of The Travel Flight Of The Passenger. It Consists Of Personal Travel And Business Travel. The Data Type Is String.

6) Class. This Field Defines The Travel Class In The Plane Of The Passenger, And Which Section They Are In. This Field Consists Of Eco: Economy Class, Eco Plus: Premium Economy Class, And Business Class. The Type Of The Data Is String.

7) Flight Distance. This Variable Contains The Distance Of The Current Journey. It Consists Of Numbers; Thus, The Data Type Is Integer.

8) Seat Comfort. This Is A Rating Variable And Shows How Comfortable Was The Passenger's Seat. It Consists Of 0-5 Where 0 Is Not Applicable And The General Rating Is 1-5. The Data Type Is Integer.

9) Departure/Arrival Time Convenient. This Variable Shows How Pleased Was The Passenger With The Time When A Plane Is Scheduled To Depart/Arrive. It Consists Of 0-5 Where 0 Is Not Applicable And The General Rating Is 1-5. The Data Type Is Integer.

10) Food And Drink. This Variable Contains The Passengers' Satisfaction Level With The Food And Drink Of The Airline. It Consists Of 0-5 Where 0 Is Not Applicable And The General Rating Is 1-5. The Data Type Is Integer.

11) Gate Location. This Variable Shows How Satisfied Are The Passenger With The Gate Distance Of The Airline (Near Or Far). It Consists Of 0-5 Where 0 Is Not Applicable And The General Rating Is 1-5. The Data Type Is Integer.

12) Inflight Wi-Fi Service. This Shows How Satisfied/Dissatisfied Are The Passengers With The Services During The Flight. It Consists Of 0-5 Where 0 Is Not Applicable And The General Rating Is 1-5. The Data Type Is Integer.

13) Inflight Entertainment. This Variable Describes How Satisfied Or Dissatisfied Are A Passenger With The Entertainment On The Flight It Consists Of 0-5 Where 0 Is Not Applicable And The General Rating Is 1-5. The Data Type Is Integer.

14) Online Support. This Variable Shows How Pleased Is The Passenger With The Airline, Providing Customer Service And Helping The Customer Resolve Their Problems. It Consists Of 0-5 Where 0 Is Not Applicable And The General Rating Is 1-5. The Data Type Is Integer.

15) Ease Of Online Booking. This Variable Displays How Satisfied/Dissatisfied Are The Passenger With The Online Booking, Such As Checking Their Booking And Ease Of Accessibility. It Consists Of 0-5 Where 0 Is Not Applicable And The General Rating Is 1-5. The Data Type Is Integer.

16) On-Board Service. This Variable Shows How Satisfied/Dissatisfied Are The Passenger With The Onboard Services On The Flight Such As Online Duty-Free, Shopping And Many

Received: August 11,2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

More. It Consists Of 0-5 Where 0 Is Not Applicable And The General Rating Is 1-5. The Data Type Is Integer.

17) Leg Room Service. This Variable Contains The Satisfaction Level Of The Passenger With The Leg Space Between The Seats. It Consists Of 0-5 Where 0 Is Not Applicable And The General Rating Is 1-5. The Data Type Is Integer.

18) Baggage Handline. This Variable Shows The Satisfaction Level Of The Passenger When Claiming Their Baggage. It Consists Of 0-5 Where 0 Is Not Applicable And The General Rating Is 1-5. The Data Type Is Integer.

19) Check-In Service. This Variable Contains The Satisfaction Level Of The Passenger When Going To Check-In (When Confirming That They Will Be On The Flight, Getting The Onboarding Pass, And Check-In Luggage). It Consists Of 0-5 Where 0 Is Not Applicable And The General Rating Is 1-5. The Data Type Is Integer

20) Cleanliness. This Variable Shows The Please Level Of The Passenger Regarding The Cleanliness Of The Airline. It Consists Of 0-5 Where 0 Is Not Applicable And The General Rating Is 1-5. The Data Type Is Integer

21) Online Boarding. This Variable Contains The Satisfaction Level Of The Passenger Regarding Online Boarding (The Easiness Of Accessibility). It Consists Of 0-5 Where 0 Is Not Applicable And The General Rating Is 1-5. The Data Type Is Integer

22) Departure Delay In Minutes. This Variable Shows How Satisfied/Dissatisfied Are The Passengers Regarding The Minutes Delayed During The Departure (The Lateness Of The Flight). It Consists Of 0-5 Where 0 Is Not Applicable And The General Rating Is 1-5. The Data Type Is Integer.

23) Arrival Delay In Minutes. This Variable Shows How Satisfied/Dissatisfied Are The Passengers Regarding The Minute's Arrival During The Departure (The Lateness Of The Flight Arriving At Its Destination). It Consists Of 0-5 Where 0 Is Not Applicable And The General Rating Is 1-5. The Data Type Is Integer.

C. Data Preparation

This Stage Consists Of Cleaning The Data And Converting It Into An Appropriate Format For Modelling. Where An Individual Will Deal With Missing Information, Correct Obvious Errors Such As A Person's Age, And Recode Evident Discrepancies. In Addition, When Dealing With Minority Values, Imbalanced Data, And Skewed Distributions, Identifying Which Variables Could Be Advantageous For Modelling By Examining Distributions And The Degree Of Noise In Each. At This Stage, We Also Divide The Data Into Training And Test Sets So That We May Assess The Model's Performance On New Data.

Data Preparation Is A Necessary Step As Many Indicators Proved That Without It, The Data Would Be Inaccurate. During The Data Preparation, Many Steps Were Taken To Ensure That The Data Is Error-Free. Firstly, The Chi-Test Square Was Used To Calculate The Significance

Received: August 11,2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Between Two Variables, To Check Which Features Would Be Best To Use, And It Has Been Done For The Nominal Value. The Second Step Consisted Of Using The Spearman Rho Test To Check The Correlation Between Variables To Analyze The Relationship Between Them And Has Been Utilized For Discrete Variables. Thirdly, Imputation Has Taken Place To Replace The Zero Variables Which Were Not Applicable For The Rating. Finally, To Finish Off The Data Preparation, Cleaning Was Used To Clear Out The Minority Or Anomaly Values To Guarantee Precise Results Using The Interquartile Range. Whereas The Splitting Was 80% For The Training And 20% For The Testing Finally Validation Has Been Done Through User Input

1) Chi-Square. The Chi-Square Test Has Been Utilized To Determine The Different Patterns And Satisfaction And Loyalty For The Airlines, If The Significance (P-Value) Of The Test Is Below 0.05, Then The Two Nominal Values Have A Significant Association. Moreover, To Check If The Variables Can Be Used For The Modelling Or Not And To Find A Pattern, The Calculation Of The Chi-Test Has Been Done. To Do This A Contingency Table Is Needed, Which Is A 2x2 Table. These Tables Can Help In Understanding The Discrete Distribution Of One Variable, Or The Relationship Between Two Variables. The Null Hypothesis Means There Is No Relationship Between The Variables And On The Other Hand, The Alternative Hypothesis Means There Is A Relationship Between The Variables. Finally, The Calculation Has Been Done Using The Formula In (1) To Find Out The Critical Value And The P-Value For The Level Of Significance.

$$\chi^2 = \sum \frac{(o-e)^2}{e} \qquad (1)$$

2) Spearman Rho Correlation Test. The Spearman Correlation Was Used On Discrete Values And The Test Evaluates And Measures The Strength And Direction Of A Relationship Between Two Ranking Variables. The Test Results Range From -1 To +1. If The Result Is -1, The Connection Is Inversely Proportionate (As One Variable Increases, The Other Decreases); If It Is 0, There Is No Relationship; And If It Is +1, The Relationship Is Indicated To Be A Perfect Positive Monotonic. Monotonic Implies That It Is Always Expanding Or Decreasing. A Positive Association Suggests That When One Variable Rises, The Other Rises As Well. To See If The Value Is Closer To 0 Or 1 Rea And Parker Suggest Interpretation Has Been Used Which Illustrates In Table 1 Which Is The Rule Of Thumb. Furthermore, A Heat Map Has Been Used To See The Correlation Between The Variables. In The End, All The Combinations With Negligible Values Whose P-Value Is More Than 5 And The Variables That Don't Have A Relationship, Have Been Removed. Whereas The Flight Distance And Food And Drink Variables Are Important And Impact The Result, And Even Though They Are Negligible, They Have A Relationship With The Target Variable And Their P-Value Is Less Than 5 Thus It Has Been Kept In The Dataset.

Received: August 11,2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Table 1: Rea And Parker Rule Of Thumb Interpretation Table.

$\|\rho\|$	Interpretation
0.00 < 0.10	Negligible
0.10<0.20	Weak
0.20<0.40	Moderate
0.40<0.60	Relatively Strong
0.60<0.80	Strong
0.80<=1.00	Very Strong

3) Cleaning. Cleaning Is A Vital Step Of Data Pre-Processing Which Is To Improve The Quality Of The Data, Hence It Increases The Efficiency, Eases The Mining Process, And Removes Noisy Data And Inconsistent Incomplete Data. Therefore Cleaning Is Done By Filling In The Missing Values, Smoothing The Noisy Data, Resolving The Inconsistency And Removing The Outliers. The Common Practice Is To Eliminate Flawed Values By Either Removing Them Manually Or Using Other Techniques. Figure 4 Depicts The Histogram Of "Age", The Left Indicates The Data Before Cleaning. As Can Be Seen From The Left-Hand Image, There Are Many Minority Values And Outliers, Thus Interquartile Has Been Used To Remove The 5th And 90th Percentiles Of The Data, Similarly, The Below Images Depicts The Histogram Of "Flight Distance". Elimination Of The 5th And 90th Percentiles Has Been Taken Into The Process To Clean The Data And Make It More Distributed. In Conclusion, The Removal Of The Data Through The Cleaning Process Is A Step That Improves The Results By Eliminating Any Anomalies.

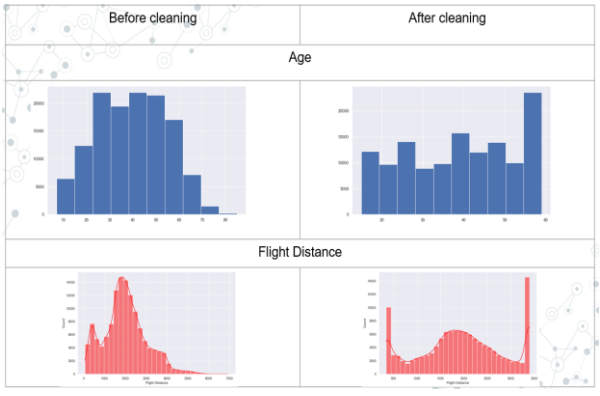

Fig. 4. : Histogram For Cleaning The ‘Age’ And The ‘Flight Distance Variables’.

4) Imputation. A Typical Issue In A Dataset Is Missing Values. There Are Several Approaches To Dealing With Lost Data. Complete Or Accessible Case Studies, Missing Measurement Approaches, And Population Mean Imputation Are Examples Of Simple And Widely Used Methods. These Methodologies, However, Result In Inefficiencies In The Analysis. Worse, They Provide Extremely Skewed Estimates Of The Connections Examined.

Received: August 11,2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

There Are More Advanced Strategies For Coping With Missing Data, Such As Multiple Embedding, That Can Produce Superior Outcomes. These Strategies Use Other Known Attributes Of Subjects To Translate Data For Missing Subjects To Predicted Values. Imputation Is Essentially Missing Values; These Missing Values Represent The Reality Of The Datasets, And When Computation Is Performed Alongside The Actual Value And The Missing Value, The Outcome Is "Nan", Which Means Not Accessible Or Not Relevant. As Mentioned Previously, There Are Several Techniques That Can Be Used To Solve The Problem. Removing Rows Containing Missing Values And Replacing The Missing Values With A Random Number May Result In Inaccurate Data Analysis [12]. Whereas Replacing The Missing Values With The Mean/Median (Average/Middle Value) Of The Dataset May Be A Better Solution Since It Will Not Decrease The Sample Size Of The Dataset, And The Median Of The Dataset Will Not Be Biased Since The Middle Value Of The Data Will Be Taken. In The Current Study, The Approach Of The Median Population Of The Dataset Has Been Taken, This Is Done Assuming That Zero Is Not Applicable, Or The Customer Did Not Rate The Feature. [13]. Thus, Many Steps Have Been Taken, Firstly, Converting The Zero To Nan Which Is Illustrated In Figure 5, And Then Imputing The Median Of The Column In The Dataset [14].

Replacing Zero's with NaN

	0	1	2	3	4
Gender	Female	Male	Female	Female	Female
Age	59	47	15	59	59
Type of Travel	Personal Travel	Personal Travel	Personal Travel	Personal Travel	Personal Travel
Class	Eco	Business	Eco	Eco	Eco
Flight Distance	341	2464	2138	623	354
Seat comfort	NaN	NaN	NaN	NaN	NaN
Food and drink	NaN	NaN	NaN	NaN	NaN
Inflight wifi service	2.0	NaN	2.0	3.0	4.0
Inflight entertainment	4.0	2.0	NaN	4.0	3.0
Online support	2.0	2.0	2.0	3.0	4.0
Ease of Online booking	3.0	3.0	2.0	1.0	2.0
On-board service	3.0	4.0	3.0	1.0	2.0
Leg room service	NaN	4.0	3.0	NaN	NaN
Baggage handling	3	4	4	1	2
Checkin service	5.0	2.0	4.0	4.0	4.0
Cleanliness	3.0	3.0	4.0	1.0	2.0
Online boarding	2.0	2.0	2.0	3.0	5.0

Replacing Nana with median values

	0	1	2	3	4
Age	59.0	47.0	15.0	59.0	59.0
Flight Distance	341.0	2464.0	2138.0	623.0	354.0
Seat comfort	3.0	3.0	3.0	3.0	3.0
Food and drink	3.0	3.0	3.0	3.0	3.0
Inflight wifi service	2.0	3.0	2.0	3.0	4.0
Inflight entertainment	4.0	2.0	4.0	4.0	3.0
Online support	2.0	2.0	2.0	3.0	4.0
Ease of Online booking	3.0	3.0	2.0	1.0	2.0
On-board service	3.0	4.0	3.0	1.0	2.0
Leg room service	4.0	4.0	3.0	4.0	4.0
Baggage handling	3.0	4.0	4.0	1.0	2.0
Checkin service	5.0	2.0	4.0	4.0	4.0
Cleanliness	3.0	3.0	4.0	1.0	2.0
Online boarding	2.0	2.0	2.0	3.0	5.0

Fig. 5. : Replacing The Zeros With 'NAN' And The Median Values With 'NANA'.

D. Modelling

Modelling Is The Fourth Phase Of The CRISP-DM, During This Phase, The Types Of Tests Which Are Best To Examine The Data Need To Be Determined. Thus, The First Step Is A Selection Of The Model Technique That Should Be Done, Secondly, A Test Needs To Be Designed, Next A Model Needs To Be Built, And Finally, The Model Needs To Be Assessed. During The Assessment Of The Model, Anomalies May Be Present. Therefore, It Is Not Possible To Blindly Accept The Results As They Need To Be Correct. Predominantly, The Model Should Be Trained On The Training Data And Tuned And Compared With The Testing Data Until The Best Model Has Been Found, This Model Should Perform Well On Unseen Data.

Received: August 11,2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

In The Subject Of Machine Learning, There Are Two Primary Categories Of Problems: Supervised Learning And Unsupervised Learning. When Dealing With Unsupervised Learning, Various Machine Learning And Statistical Approaches Are Applied To The Dataset To Construct A Structure Of Data And Gain A More Comprehensive Knowledge Of What Is Going On And What Features Are Controlling Other Features. Since There Are No Labels Accessible In Unsupervised Learning, No Predictions Can Be Made [15].

There Are Many Different Models That Can Be Used For Both Regression And Classification Problems, Conventional Statistical Methods Include Multiple Linear Regression For Regression Problems, Feeding Of The Various Individual Features Will Be Done And Attempting To Predict Some Real Number Output. Additionally Logistic Regression For Classification Problems, Where Similarly Feeding Several Individual Features And Getting A Binary Outcome Of 0 Or 1. Logistic Regression Is Typically Taking Multiple Linear Regression And Squashing It Through A Logistic Function. A Logistic Function Can Be Generalized To Classify Multiple Different Categories Were Also Any Number Of Machine Learning Can Be Done Instead [15].

Four Different Classification Models Have Been Used To Build The Prediction Of The Airline Customer Satisfaction Model, Which Are Logistic Regression, Naïve Bayes, K-Nearest Neighbours And Random Forest. Moreover, Different Hyperparameters Tuning Was Used, Additionally With Grid Search, Where The Four Highest Accuracy Score Models Were Taken After Comparing Each Of The Four Models With The Highest Precision For Picking The Best Model. Nevertheless, The Focus For Picking The Best Model Was Based On Precision Since We Want To Predict The Correct Prediction Ratio Of The Satisfied Customer, For Instance, If The Model Incorrectly Predicts The Customers Are Satisfied But Dissatisfied This Will Result In An Incorrect Analysis.

1) Logistic Regression. Logistic Regression Is A Classic Statistical Model And Supervised Learning For Classification Derived From Multiple Linear Regression, In Which The Dependent Variable Is A Function Of Many Independent Variables, Each With Its Own Unique Coefficients And An Intercept Term, All Of Which Are Crammed Within A Logistic Function. The Logistic Function Compresses The Linear Model's Output So That It Remains Between 0 And 1. The Result Of Logistic Regression Is A Probability Value Indicating Whether The Inputs Belong To Class 0 Or Class 1, Or Any Arbitrary Pair Of Classes The Person Selects To Feed Into The Model [16].

One Of The Main Points Of Convenience With Logistic Regression Is Interpretability, The Coefficients That An Individual Gets Out Of The Logistic Regression Directly Quantify The Influence Of Each Individual Independent Variable On The Output As With Multiple Linear Regression. In The Case Of Logistic Regression, That Parameter That Is Output Can Be Exponentiated To Get The Odds Ratio For That Particular Independent Variable. Essentially, If Holding Everything Else The Same The Odds Ratio Tells An Individual How Much Of An

Received: August 11,2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Increase In The Output Probability You Get With An Increase In That Independent Variable [17].

As Illustrated, In Figure 6, Model 2 (Which Is The Grid Search Hyper-Parameters) Is A Good Model Since It Has The Highest Accuracy Level Among The Rest Of The Models Which Is 83.6%. Moreover, In Figure 7 Model 3 Has The Highest Accuracy Level Which Is 86.8% Among The Logistic Regression Of The Customer Type Target Variable. The Accuracy Level Indicates The Portion Of The Right Prediction Model.

Model¤	Parameters¤	CA¤	F1¤	Precision¤	Recall¤
1¤	solver='liblinear',¶ random_state=42¤	83.5886973%¤	85.2363636%¤	84.4380403%¤	86.0499266%¤
2·(Grid· Search)¤	C=·1,¶ penalty='l1',¶ solver='liblinear',¶ random_state=42¤	83.6194949%¤	85.2589641%¤	84.4891513%¤	86.0429341%¤
3¤	C·=·10,¶ penalty=·'l1',·¶ solver·=·'saga',¶ random_state=42¤	73.1944872%¤	78.0353932%¤	71.0862069%¤	86.4904552%¤
4¤	C·=·10,·¶ penalty=·'l2',·¶ solver·=·'sag',·¶ random_state=42¤	76.5976286%¤	80.554045%¤	74.2393868%¤	88.0427942%¤

Fig. 6. : Logistic Regression With 'Satisfaction' Target Variable

Model¤	Parameters¤	CA¤	F1¤	Precision¤	Recall¤
1¤	solver='liblinear',¶ random_state=42¤	86.4721281%¤	55.7096042%¤	69.5624803%¤	46.4578516%¤
2·(Grid· Search)¤	C=·1,¶ penalty='l1',¶ solver='liblinear',¶ random_state=42)¤	86.8763474%¤	59.6137898%¤	68.2953312%¤	52.8904772%¤
3¤	C·=·10,·¶ penalty='l2',¶ solver·=·'newton-cg',¶ random_state=42¤	86.8917462%¤	59.6802842%¤	68.329718%¤	52.9745638%¤
4¤	C·=·10,·¶ penalty=·'l2',·¶ solver·=·'sag',·¶ random_state=42¤	81.6484447%¤	0.5009393%¤	35.2941176%¤	0.2522598%¤

Fig. 7. : Logistic Regression With 'Customer Type' Target Variable.

2) Naïve Bayes. A Naïve Bayes Classifier Works On The Principle Of Conditional Probability As Given By The Bayes Theorem. Bayes Theorem Gives The Conditional Probability Of An Event A Given Another Event B Has Occurred. $P(A \mid B)=(P(B \mid A)*P(A))/(P(B))$ Can Be Used To Find A Conditional Probability Where An Individual Can Find The Probability Of A Given The Event B Has Occurred By Knowing Some Parameters Which Are The Individual Probabilities Of A And B And Knowing The Probability Of B Given That A Has Occurred. This Theorem Is Very Powerful Since One May Know The Probability Of Certain Events But Doesn't Know The Probability Of Some Other Events And Using Certain Events Other Probabilities Can Be Found. Furthermore, It Is A

Received: August 11,2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Simple Assumption Which Reduces Calculations And Makes The Algorithm Simple Yet Effective. [18] [19] [20].

As It Can Be Seen From The Figure 8 , Model 1 Has The Highest Accuracy Level (81.4%) For The 'Satisfaction' Target Variable, However, In Figure 9, Model 2 (Which Is The Grid Search Hyper-Parameter) Has The Highest Accuracy Score (85.7%) Concerning The 'Customer Type' Target Variable.

Model	Parameters	CA	F1	Precision	Recall
1	nb_1 = GaussianNB()	81.4020634%	83.136105%	83.0057159%	83.2669044%
2 (Grid Search)	var_smoothing=3.5111917342151277e-08	81.3135202%	83.081213%	82.8271596%	83.3368296%
3	var_smoothing= 1e-03	60.0477364%	67.56875%	61.082547%	75.5961122%
4	var_smoothing= 1e-05	77.9065291%	81.1977853%	76.3900875%	86.6512831%

Fig. 8. Naive Bayes With 'Satisfaction ' Target Variable

Model	Parameters	CA	F1	Precision	Recall
1	nb_1_CT = GaussianNB()	84.7975054%	65.2774114%	56.1064087%	78.0323733%
2 (Grid Search)	var_smoothing=3.5111917342151277e-08	85.7406837%	63.5073892%	59.7626553%	67.7527854%
3	var_smoothing= 1e-10	84.6627656%	65.2052402%	55.77469%	78.473828%
4	var_smoothing= 1e-09	30.8862026%	14.17029211%	22.4002418%	10.3629117%

Fig. 9. : Naïve Bayes With 'Customer Type' Target Variable.

3) K-Nearest Neighbours. As Mentioned, Machine Learning Models Make Predictions By Learning From Past Data Available. The K-Nearest Neighbours Algorithm (KNN) Is A Classification Algorithm, Where Classification Determines What Group, Something Belongs In. Because KNN Is Based On Feature Similarity, Classification Can Be Done Using A KNN Classifier, Hence Having The Input Value Which Goes Into The Trained Model And Predicts The Output [28].

K-Nearest Neighbours Is One Of The Simplest Supervised Machine Learning Algorithms Mostly Used For Classification, Which Classifies A Data Point Based On How Its Neighbours Are Classified. KNN Stores All Available Classes And Classifies New Cases Based On Similarity Measures. 'K' In KNN Is A Parameter That Refers To The Number Of Nearest Neighbours To Include In Most Of The Voting Process [21].

Parameter Tuning Can Help In Choosing The Right Value Of 'K' Which Is Important For Better Accuracy. Fundamentally, 'K' Changes Depending On Where The Point Is, This

Received: August 11,2025

International Journal of Applied Mathematics
Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Drastically Changes The Answer. Choosing The 'K' And Getting It Right Is Vital So That There Won't Be Any Huge Bias. Another Option In Choosing 'K' Is To Take The Square Root Of 'N' Which Is The Total Number Of Values An Individual Has. KNN Can Be Used When The Data Is Labelled, The Data Is Noise-Free, And Since KNN Is A Lazy Learner, It Doesn't Learn Discriminative Functions From The Training Set [21].

As Is Illustrated In Figure 10, Model 3 Has The Highest Accuracy Level (79.5%), While In Figure 11 Which Is The 'Customer Type' Target Variable, Model 2 (Which Is The Grid Search Model) Has The Highest Accuracy Level Among The Rest Of The Hyper-Parameters Within The Group.

Model	Parameters	CA	F1	Precision	Recall
1	n_neighbors=15	79.223129%	81.2310902%	80.8011623%	81.6656178%
2 (Grid Search)	n_neighbors=13	79.4810594%	81.4311594%	81.1428175%	81.7215579%
3	leaf_size=13, n_neighbors=9	79.5773021%	81.4322215%	81.5206727%	81.343962%
4	leaf_size=24, n_neighbors=20	78.5686788%	80.1836756%	81.6632831%	78.7567303%

Fig. 10. : K-Nearest Naighbour With 'Satisfaction' Target Variable.

Model	Parameters	CA	F1	Precision	Recall
1	n_neighbors=15	84.3509393%	48.0245493%	61.2924282%	39.478663%
2 (Grid Search)	n_neighbors=13	84.4664306%	48.8917036%	61.5041428%	40.5717889%
3	n_neighbors=9	84.1507545%	48.9902119%	59.6560048%	41.5598066%
4	n_neighbors=20	84.4125346%	44.0359364%	64.2857143%	33.4874921%

Fig. 11. : -Nearest Naighbour With 'Customer Type' Target Variable.

4) Random Forest. A Random Forest Is A Sort Of Machine-Learning Model That Predicts Using An Ensemble Of Decision Trees. This Model Is Created By Collecting A Random Sample Of Data And Then Developing A Continuing Series Of Decision Trees On The Subsets, Resulting In A Vast Number Of Decision Trees, Each Of Which Gives Rise To A Bigger Model Or Group. The More Decision Trees Utilized With Diverse Criteria, The Better The Random Forest Performs Since It Is Essentially Boosting Prediction Accuracy, And If One Or Two Of These Smaller Decision Trees Are Not Important On A Particular Occasion, They Should Be Ignored [22].

One Of The Key Advantages Of Random Forest Is That It Can Assist Prevent Overfitting, Which Occurs When The Model Begins To Remember The Data Rather Than Attempting To

Received: August 11,2025

International Journal of Applied Mathematics
Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Generalize By Generating Predictions On Future Data. Essentially, It Aids In Overcoming The Constraints Of The Data, Which May Not Be Completely Representative Of All The Best Features In The Model. It Can Also Assist In Reducing Bias, Which Can Arise When A Certain Degree Of Mistake Is Included In The Model. When Cases Are Not Distributed Evenly Throughout Training, Bias Arises. As A Result Of How The Model Is Configured, Instead Of Seeing All The Data Points, One May Only View Half Of Them. [23].

Parameters Are Required To Build Up A Random Forest; These Parameters Include Node Size, Number Of Trees, And Number Of Features. This May Be Difficult At The Beginning Since Many Trees Are Required To Get The Maximum Predicted Accuracy, Yet Numerous Trees Consume A Lot Of Time To Train The Model And Take Up A Lot Of Memory Space [24] [25].

As Is Demonstrated In Figure 12, Model 4 Has The Highest Accuracy Level (94.6%) With The Target Variable Of 'Satisfaction', Similarly, Figure 13 Shows That Model 4 Has The Highest Accuracy Level (97.3%) Among The 'Customer Type' Target Variable Group.

Model	Parameters	CA	F1	Precision	Recall
1	n_estimators=70, oob_score=True, n_jobs=-1, random_state=42, max_features=None, min_samples_leaf=30	93.2784108%	93.8886944%	93.9939729%	93.7836515%
2 (Grid Search)	bootstrap=True, max_depth=4, max_features='sqrt', min_samples_leaf=2, min_samples_split=5, n_estimators=80, random_state=42	86.9571913%	88.2475371%	87.5610931%	88.944829%
3	bootstrap=True, max_depth=10, max_features='auto', min_samples_leaf=4, min_samples_split=10, n_estimators=600	92.2774869%	93.0260047%	92.5050128%	93.5528984%
4	bootstrap=False, max_depth=20, max_features='sqrt', min_samples_leaf=4, min_samples_split=2, n_estimators=400 ,random_state=42	94.6566061%	95.1044018%	95.9504662%	94.2731278%

Fig. 12. : -Random Forest With 'Satisfaction' Target Variable.

Received: August 11,2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Model	Parameters	CA	F1	Precision	Recall
1	n_estimators=70, oob_score=True, n_jobs=-1, random_state=42, max_features=None, min_samples_leaf=30	95.6806283%	87.8175896%	90.8151808%	85.0115619%
2 (Grid Search)	bootstrap=True, max_depth=4, max_features='sqrt', min_samples_leaf=2, min_samples_split=5, n_estimators=80, random_state=42	85.2556206%	33.1821354%	97.5384615%	19.9915913%
3	bootstrap=True, max_depth=10, max_features='auto', min_samples_leaf=4,min_samples_split=10,n_estimators=600	95.7460733%	87.7887059%	92.544268%	83.4980029%
4	bootstrap=False, max_depth=20, max_features='sqrt', min_samples_leaf=4, min_samples_split=2, n_estimators=400, random_state=42	97.355251%	92.62163%	94.6859903%	90.6453647%

Fig. 13. : -Random Forest With 'Customer Type' Target Variable.

6) User Input. This Is The Step Of Validation From The Existing Data. This Allows The Users To Have The Ability To Input The Data, Through Which The Answers Will Be Processed In The Model To Output The Predictions Of The Customer 'Satisfaction And Customer Type'. Figure 14 Illustrates A Simple User Input That Has Been Performed, In Which The Customer Answers The Questions, The Model Will Predict Whether The Customer Is Pleased Or Unsatisfied, As Well As Whether The Customer Is Loyal Or Disloyal, Based On The Question Answered.

```
chose the following for your gender: female=0, Male=1 1
What is your Age? 15
choose the folowing for your Travel Type: Business travel=0 ,Personal Travel=1 0
choose the following: With what class are you travelling?  Eco=1 or Business=0 or Eco Plus=2 1
Write just the number how log was the flight?123
'Rating:Very Satisfied =5 ,Satisfied =4 , Neutral=3 ,Dissatisfied=2 and Very Dissatisfied=1'
Rate how comftable was the seat 1-5 : 5
Rate how was the food and the drink 1-5 : 5
Rate how was Inflight wifi service the 1-5: 5
Rate how was Inflight entertainment service the 1-5: 5
Rate how was the onlinne support 1-5: 5
Rate how was the easiness of online booking 1-5: 4
Rate how was the on-board service  1-5: 3
Rate how was the leg room space in the plane 1-5: 5
Rate how was the Baggage handling  1-5: 5
Rate how was the checkin service  1-5: 5
Rate how was the Cleanliness of the plane 1-5: 5
Rate how was the online boarding  1-5: 5
'Thanks for your time :)'
Customer is:  ['satisfied'] but ['disloyal Customer']
```

Fig. 14. : User Interface With A Question And The Prediction.

Received: August 11,2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

IV. Results And Discussion

The Final Phase Of The CRISP-DM Is Evaluation When The Outcome Is Assessed. At This Point, The Chosen Model Is Applied To The Unseen Test Data To Quantify Its Performance If It Is Regarded As Adequate.

This Step Is A Comparison Of The Eight Models, With Each Model Displaying The 'Satisfaction' Target Variable And The 'Customer Type' Target Variable, As Well As The Confusion Matrix, Which Displays The Values Of The True Positive, True Negative, False Positive, And False Negative. The Confusion Matrix Compares The Actual And Expected Goal Values. Finally, The Area Under The Curve (AUC), Receiver Operating Characteristic Curve (ROC) Has Been Displayed. The Generated Model Was Used To Predict Outcomes On The Testing Dataset, And Metrics Of The Model's Effectiveness Were Calculated As A Result. A Confusion Matrix Exemplifies The Results Predicted By A Model For The Test Set Situations To Their Actual Outcomes. This Produced A 2x2 Table With Four Potential Outputs In The Instance Of The Binary Outcome For ‘Satisfaction’ And ‘Customer Type’ Here:

- True Positives (TP): Predicted Passengers Are Satisfied While The True Classes Are Satisfied.
- False Positives (FP): Predicted Passengers Are Satisfied While The True Class Is Dissatisfied.
- True Negatives (TN): Predicted Passengers Are Dissatisfied While The True Class Is Dissatisfied.
- False Negatives (FN): Predicted Passengers Are Dissatisfied While The True Class Is Satisfied.
- True Positives (TP): Correctly Predicted That The True Classes Are Disloyal
- False Positives (FP): Incorrectly Predicted That The True Classes Are Disloyal
- True Negatives (TN): Correctly Predicted That The Classes Are Loyal
- False Negatives (FN): Incorrectly Predicted That The Class Are Loyal

A Variety Of Success Measurements May Be Derived From These Four Numbers. The Accuracy Of The Model Is The Percentage Of Correctly Predicted Labels: (TN+TP)/(TP+FP+TN+FN). However, Another 2 Matrices That An Individual Would Encounter Are Precision And Recall. Precision Is The Quantity Of Predicted Positive That Is Positive TP/(TP+FP), Which Is The Division Of The True Positive By The Sum Of The True And False Positive. Another Metric That An Individual Should Consider Is Recall Which Is Also Known As Sensitivity. Recall Differs From Precision And Works To Replace The False Positives In This Part Of The Equation With The False Negatives TP/(TP+FN), When Using Precision, The Overall Aim Is On Ensuring That As Many Of The Positive Predictions That Were Made Were Indeed Positive. But With Recall, The Overall Goal Is On Capturing As Many Positive Cases As Possible In Some Circumstances, We Must Strike A Compromise Between Preventing False Positives And False Negatives; In These Cases, The F1 Score Can Be Utilized.

Received: August 11,2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The F1 Score Is The Harmonic Mean Of Accuracy And Recall, Which Is Calculated By Multiplying Precision By The Recall And Dividing The Result By The Total Precision And Recall. 2*((Precision X Recall))/ ((Precision+ Recall)) [26]. Lastly, The Area Under The Receiver Operator Characteristic Curve (AUC Score From The ROC Curve) Can Be Used. The ROC Curve Is A Probabilistic Graphical Depiction Of The Classifier's Ability To Differentiate Between Two Classes, With The X-Axis Representing The Classifier's False Positive Rate And The Y-Axis Representing The True Positive Rate Or Recall [27].

The Classification Threshold Can Be Varied. On One Side Of The Plot, The Threshold Can Be Increased, Classifying This As Positive, Meaning That There Is A Low False Positive Rate, But Also A Low True Positive Rate, In The Sense That There Are A Lot Of False Negatives. On The Other Side, The Classification Threshold Can Be Increased, Capturing More True Positives, But Also Increasing The Number Of False Positives. This Plot Allows The Individual To Measure How Good A Model Is At Separating The Positive And The Negative Classes, Which Is Done By Calculating The Area Under The Curve. By Calculating The Area Under The Curve, A Matric Can Indicate How Well The Model Can Distinguish Between Both Positive And False Classes [28].

After Comparing The Four Models Confusion Matrix Has Been Done And Compared Between The Four Models With The ‘Satisfaction’ Target Variable, And The Confusion Matrix With The ‘Customer Type’ Target Variable, Where The Highest Model Was Chosen Using The Highest Accuracy Level, And The Best Model Was Chosen Based On The Precision Since The Target Is To Check The Correct Positive Predictions That Were Made To Ensure The Satisfaction Of The Customers.

Finally, The Area Under The Curve Has Been Done Using The Four Models And As Can Be Seen From Figures 17 And 18 The Random Forest Classifier Has The Highest Curve.

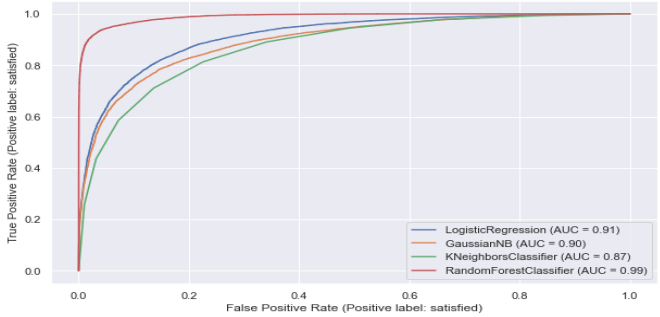

Fig. 15. : -Area Under The Curve Of The Four Models With ‘Satisfaction’ Target Variable.

Received: August 11,2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

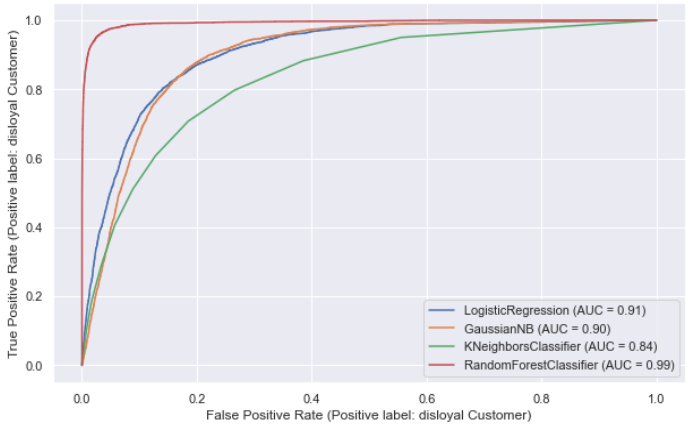

Fig. 16. : -Area Under The Curve With 'Customer Type' Target Variable.

V. CONCLUSION

The Goal Of Multifaceted Analysis Of Customer Feedback On Airline Services Is To Make The Feedback And Questionnaire Process More Efficient And Smoother, As Well As To Create And Achieve A Learning Model That Can Predict The Customer's Satisfaction Level And Loyalty To The Airline So That The Airline Can Strategize What Services It Needs To Adapt And Apply For Each Unique Customer In Order To Generate A Larger Number Of Satisfied Consumers. The Survey's Overall Scope Spans Numerous Areas Of Airline Services And Customer Information To Make It Simple And Easy For Customers To Review And Staff Members To Inspect And Research The Results, And This Is Done By Achieving The Following Two Goals:

- To Choose Which Model To Use For The Research. The Four Models Were Measured Using The Area Under The Curve To Make Sure Which Model Has Better Performance Between The Positive And Negative Classes. The Result Was That The Random Forest Is The Best Model At 99%.

- To Predict Customer Fulfilment And Loyalty To The Airline, Where The Model Was Fruitful In Predicting, The Results Needed Through The User Input Accurately. This Was Proven By Testing The Model With 5-10 Different People By Making Them Answer The Required Questions And Making Sure That The Results Matched The User's Intention.

Some Of The Limitations Of The Study Include The Possible Inaccuracy Since The Data Was From Open Sources, If The Data Was Collected From Viable Sources, The Result Could Be More Precise. Furthermore, There Are Many Other Models Could Have Been Used To Test The Data Such As Neural Network, Decision Tree And Many More. Thus, A Number Of Steps Could Be Taken To Improve The Analysis And The Model. Firstly, As Mentioned Before Different Classification Models Could Be Used To Test Different Classification Models In Future. Furthermore, The User Input Design Could Be Enhanced, Such As Creating A Web Interface With It So, Various Companies Can Use It With Real Data, Likewise, When Each

Received: August 11,2025

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Customer Finishes The Questionnaire, The Result Can Be Sent Back To The Staff Members So That They Can Analyze The Results Faster And More Effectively. Furthermore, Adding A Comment Section In The Web Interface So The Customer Can Write Their Overall Feedback On The Airline And Any Other Concerns That Have Not Been Addressed. In Addition, A Recommendation System Can Be Implemented As Well.

REFERENCES

[1] H. Vaza, "Medium," 16 July 2021. [Online]. Available: Https://Harshalvaza.Medium.Com/Invistico-Airlines-Understanding-Customer-Satisfaction-6108b500e592.

[2] T. P. Alvin, "Medium," 10 October 2020. [Online]. Available: Https://Towardsdatascience.Com/Predicting-Satisfaction-Of-Airline-Passengers-With-Classification-76f1516e1d16.

[3] J.-P. H. Tri Noviantoro, "Investigating Airline Passenger Satisfaction: Data Mining Method," *Research In Transportation Business & Management,* P. 100726, 2021.

[4] R. A. R. Bruce K. Behn, "Using Nonfinancial Information To Predict Financial Performance: The Case Of The U.S. Airline Industry," *Journal Of Accounting, Auditing & Finance,* Vol. 14, No. 1, Pp. 29-56, 1999.

[5] R. P. S. A. D. S. L. M. Praphula Kumar Jain, "IEEE Xplore," 1 November 2019. [Online]. Available: Https://Ieeexplore.Ieee.Org/Abstract/Document/9036251/Authors#Authors.

[6] M.-Y. K. Y.-J. K. Y.-H. P. So-Hyun Park, "A Deep Learning Approach To Analyze Airline Customer Propensities: The Case Of South Korea," *Applied Sciences,* Vol. 12, No. 4, 2022.

[7] D. K. A. M. A. D. M. A. W Baswardono, "IOP Science," *Comparative Analysis Of Decision Tree Algorithms: Random Forest And C4.5 For Airlines Customer Satisfaction Classification,* Vol. 1402, No. 6, P. 7, 2019.

[8] V. García, "Predicting Airline Customer Satisfaction Using K-Nn Ensemble Regression Models," 2019. [Online]. Available: Http://148.210.21.170/Handle/20.500.11961/8559.

[9] M. C. Özgür Davras, "Analysis Of Hotel Services By Their Symmetric And Asymmetric Effects On Overall Customer Satisfaction: A Comparison Of Market Segments," *International Journal Of Hospitality Management,* Vol. 81, Pp. 83-93, 2019.

Received: August 11,2025

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[10] L.-Y. W. Y.-C. C. H.-S. T. Chwo-Ming Joseph Yu, "Perceived Quality, Customer Satisfaction, And Customer Loyalty: The Case Of Lexus In Taiwan," *Total Quality Management & Business Excellence,* Vol. 16, No. 6, Pp. 707-719, 2005.

[11] "Simplilearn," 22 December 2020. [Online]. Available: Https://Www.Simplilearn.Com/What-Is-Data-Mining-Article.

[12] A. L. Roderick J., "UCLA Department Of Statistics Papers Title Regression With Missing X's: A Review," 2011.

[13] D. B. RUBIN, "Frontmatter," *Multiple Imputation For Nonresponse In Surveys,* 1987.

[14] J. Schork, "Mean Imputation For Missing Data (Example In R & SPSS)," 2022. [Online]. Available: Https://Statisticsglobe.Com/Mean-Imputation-For-Missing-Data/.

[15] Seldon, "Supervised Vs Unsupervised Learning Explained," 16 October 2021. [Online]. Available: Https://Www.Seldon.Io/Supervised-Vs-Unsupervised-Learning-Explained#:~:Text=The%20main%20difference%20between%20supervised.

[16] "Module 4 - Logistic Regression | The Programming Foundation," 2022. [Online]. Available: Https://Learn.Theprogrammingfoundation.Org/Getting_Started/Intro_Data_Science/Module4/?Gclid=Eaiaiqobchmi8paj9als-Qivsgolch0p_Gzceaayasaaeglrofd_Bwe.

[17] S. Jessica, "How Does Logistic Regression Work?," 15 July 2022. [Online]. Available: Https://Www.Kdnuggets.Com/2022/07/Logistic-Regression-Work.Html#:~:Text=Is%20Logistic%20Regression%3F-.

[18] C. Bishop, "Naïve Bayes Classifier We Will Start Off With A Visual Intuition, Before Looking At The Math…," 2006. [Online]. Available: Https://Www.Cs.Ucr.Edu/~Eamonn/CE/Bayesian%20Classification%20withinsect_Examples.Pdf.

[19] N. Kumar, "Naive Bayes Classifiers - Geeksforgeeks," 14 January 2019. [Online]. Available: Https://Www.Geeksforgeeks.Org/Naive-Bayes-Classifiers/.

[20] P. Vadapalli, "Naive Bayes Explained: Function, Advantages & Disadvantages, Applications In 2023 | Upgrad Blog," 5 October 2022. [Online]. Available: Https://Www.Upgrad.Com/Blog/Naive-Bayes-Explained/#:~:Text=Naive%20Bayes%20is%20suitable%20for.

[21] O. Harrison, "Machine Learning Basics With The K-Nearest Neighbors Algorithm," 14 July 2019. [Online]. Available: Https://Towardsdatascience.Com/Machine-

Received: August 11,2025

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Learning-Basics-With-The-K-Nearest-Neighbors-Algorithm-6a6e71d01761#:~:Text=Summary-.

[22] S. E. R, "Random Forest | Introduction To Random Forest Algorithm," 17 June 2021. [Online]. Available: Https://Www.Analyticsvidhya.Com/Blog/2021/06/Understanding-Random-Forest/#:~:Text=Random%20forest%20is%20a%20Supervised.

[23] I. N. Dmitry Devetyarov, "Prediction With Confidence Based On A Random Forest Classifier," P. 37–44, 2010.

[24] "What Is Random Forest? | IBM," 2020. [Online]. Available: Https://Www.Ibm.Com/Topics/Random-Forest.

[25] T. Buskirk, "Surveying The Forests And Sampling The Trees: An Overview Of Classification And Regression Trees And Random Forests With Applications In Survey Research," January 2018. [Online]. Available: Https://Www.Researchgate.Net/Publication/323155152_Surveying_The_Forests_And_Sampling_The_Trees_An_Overview_Of_Classification_And_Regression_Trees_And_Random_Forests_With_Applications_In_Survey_Research.

[26] S. Narkhede, "Understanding Confusion Matrix," Towards Data Science, 9 May 2018. [Online]. Available: Https://Towardsdatascience.Com/Understanding-Confusion-Matrix-A9ad42dcfd62.

[27] S. Narkhede, "Understanding AUC - ROC Curve," Towards Data Science, 26 June 2018. [Online]. Available: Https://Towardsdatascience.Com/Understanding-Auc-Roc-Curve-68b2303cc9c5.

[28] J. N. Mandrekar, "Receiver Operating Characteristic Curve In Diagnostic Test Assessment," *Journal Of Thoracic Oncology,* Vol. 5, No. 9, Pp. 1315-1316, 2019.

[29] T. P. Alvin, "Airline Flight Satisfaction," 10 October 2020. [Online]. Available: Https://Flight-Satisfaction-Prediction.Herokuapp.Com/.

[30] G. Piatetsky, "CRISP-DM, Still The Top Methodology For Analytics, Data Mining, Or Data Science Projects," October 2014. [Online]. Available: Https://Www.Kdnuggets.Com/2014/10/Crisp-Dm-Top-Methodology-Analytics-Data-Mining-Data-Science-Projects.Html.

[31] J. Saltz, "CRISP-DM For Data Science Teams: 5 Actions To Consider," 29 May 2020. [Online]. Available: Https://Www.Datascience-Pm.Com/Crisp-Dm-For-Data-Science-Teams-5-Actions-To-Consider/.

Received: August 11,2025

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[32] "Ordinal Vs Scale - Part 3: Test And Effect Size For Association (Spearman's Rho)," 2022. [Online]. Available: Https://Peterstatistics.Com/Crashcourse/3-Twovarunpair/Ordscale/Ordscale3.Html.

Received: August 11,2025