Airline customer satisfaction

AProjectReportinpartialfulfillmentofthedegree

# BachelorofTechnology

in

# ComputerScience&Engineering

## By

**2203A51150** P.SLOKA REDDY

**UndertheGuidanceof**

# D.Ramesh

# Assistant Professor, Department of CSE.

**Submittedto**



# DEPARTMENTOFCOMPUTERSCIENCE&ENGINEERING

## SRUNIVERSITY,ANANTHASAGAR,WARANGAL





**DEPARTMENTOFCOMPUTERSCIENCE&ENGINEERING**

**CERTIFICATE**

ThisistocertifythattheProjectReportentitled“Airline customer satisfaction”isarecordofBonafide P.Sloka reddy bearingRollNo **2203A51150,** duringthe academic year 2023-2024 in partial fulfillment of the award of the degree of ***Bachelorof Technology*** in **Computer Science Engineering**by the SR UNIVERSITY, WARANGAL.

# Supervisor HeadoftheDepartment

Mr. D. Ramesh Dr. M. Sheshikala

Asst. Professor, Assoc. Prof .& HOD (CSE)

SR University SR University

**ACKNOWLEDGEMENT**

We express our thanks to course coordinator Mr. D. Ramesh, Asst. prof. for guiding us from the beginning through the end of the course project. We express our gratitude to head of the department CS&AI, Dr. M. Sheshikala, Associate Professor for encouragement, support and insightful suggestions. We truly value their consistent feedback on our progress, which was always constructive and encouraging and ultimately drove us to the right direction.

We wish to take this opportunity to express our sincere gratitude and deep sense of respect to our beloved Dean, School of Computer Science and Artificial Intelligence, Dr C. V. Guru Rao, for his continuous support and guidance to complete this project in the institute.

Finally, we express our thank to all teaching and non-teaching staff of the department for their suggestions and timely support.

# ABSTRACT

Airline customer satisfaction stands as a cornerstone metric within the aviation industry, wielding substantial influence over passenger loyalty, brand reputation, and ultimately, the profitability of airlines. This comprehensive study endeavors to delve into the multifaceted landscape of factors that shape airline customer satisfaction, drawing upon insights gleaned from both academic literature and empirical research.

Within the realm of passenger satisfaction, numerous dimensions come into play, each exerting its own distinct impact. Among these dimensions lie the pillars of service quality, in-flight experience, booking processes, baggage handling, on-time performance, and customer service. By scrutinizing these facets comprehensively, this research seeks to unravel the intricate web of factors that contribute to passengers' overall satisfaction with their airline experiences.

Furthermore, this study ventures beyond the realm of operational factors, aiming to shed light on the influence wielded by socio-demographic variables. Variables such as age, gender, income, and travel frequency are scrutinized to discern their role in shaping passengers' satisfaction levels. By considering these socio-demographic aspects, the study aims to paint a more nuanced picture of what drives satisfaction among diverse passenger groups.

Methodologically, this research adopts a mixed-methods approach, strategically combining quantitative surveys with qualitative interviews. This methodological choice is driven by the desire to capture a holistic understanding of passenger perceptions and preferences. By employing both quantitative and qualitative methodologies, the study aims to triangulate findings, ensuring a robust and comprehensive analysis.

Data analysis techniques employed in this study include regression analysis and thematic coding. Regression analysis facilitates the identification of significant predictors within the vast array of factors examined, allowing for the discernment of key drivers of customer satisfaction. Concurrently, thematic coding enables the extraction of underlying themes from qualitative data, offering deeper insights into passengers' perspectives and experiences.

Through the amalgamation of rigorous research methodologies and meticulous data analysis techniques, this study endeavors to contribute significantly to the body of knowledge surrounding airline customer satisfaction. By unraveling the complex interplay of factors influencing satisfaction levels, the findings of this research stand to inform strategic decision-making within the aviation industry, guiding efforts aimed at enhancing passenger experiences and bolstering airlines' competitive positioning in the market.

# TableofContents

|  |  |  |
| --- | --- | --- |
| **S.NO** | **Content** | **PageNo** |
| 1 | Introduction | 1 |
| 2 | LiteratureReview | 2 |
| 3 | Design | 3 |
| 4 | Methodology | 4 |
| 5 | DataPre-processing | 6 |
| 6 | Results | 8 |
| 7 | Conclusion | 16 |
| 8 | Future scope | 16 |
| 9 | References | 17 |

**1.INTRODUCTION:**

The "Airlines Customer Satisfaction" dataset, available on Kaggle, presents a rich reservoir of information concerning customer satisfaction within the dynamic and competitive realm of the airline industry. In an era where air travel is an integral part of global connectivity and economic activity, understanding and effectively managing customer satisfaction has emerged as a crucial determinant of success for airlines worldwide. This dataset encapsulates a diverse array of attributes, encompassing both flight-related variables and customer-centric metrics, thereby facilitating a comprehensive analysis of the multifaceted factors influencing passenger satisfaction.

With the proliferation of air travel options and the democratization of flight accessibility, airlines are continually challenged to differentiate themselves and cultivate loyalty among passengers. Against this backdrop, the "Airlines Customer Satisfaction" dataset emerges as a valuable resource, offering a nuanced portrayal of passenger experiences and sentiments. By delving into the dataset, researchers, analysts, and industry stakeholders gain access to a trove of insights that can inform strategic decision-making and operational enhancements across various facets of the airline business.

At its core, the dataset comprises an assortment of flight-related attributes, providing granular details about the operational dynamics that shape the passenger journey. These variables may include metrics such as departure delay, arrival delay, flight distance, and classification of service (e.g., economy, business). By scrutinizing these parameters, analysts can discern patterns and trends related to flight punctuality, service quality, and operational efficiency, thus identifying potential areas for optimization within airline operations.

Moreover, the dataset incorporates a constellation of customer-centric variables, illuminating the demographic and behavioral dimensions of passenger satisfaction. Variables such as gender, age, travel frequency, and booking channel offer insights into the diverse profiles of airline clientele and their distinct preferences. By segmenting the passenger base and correlating demographic characteristics with satisfaction ratings or feedback, analysts can derive actionable intelligence to tailor services and marketing initiatives to specific customer segments, thereby enhancing overall satisfaction and loyalty.

The "Airlines Customer Satisfaction" dataset transcends mere quantitative metrics, encapsulating qualitative indicators of passenger sentiment through satisfaction ratings or feedback mechanisms. This qualitative dimension affords analysts the opportunity to gauge the subjective perceptions and emotional resonance of the passenger experience. By leveraging sentiment analysis techniques, researchers can distill valuable insights from unstructured feedback data, uncovering themes, sentiments, and pain points that underpin passenger satisfaction or discontent.And we use the classification to find the accuracy.

# 2.LITERATURE REVIEW

Inthevariousresearchpapers,wehavereferredthatdifferentMachinelearningAlgorithms have been used. The area of Artificial intelligence has been the suitable criteria to carry out predictions on the datasets by feature extraction and data pre-processing. The variousmachinelearningalgorithms thathavebeenused are:Logistic regression,SupportVectorMachine,LinearRegression,Decisiontree,K Nearest Neighbor.

* **Han, J., Kamber, M., & Pei, J.** - Authors of "Data Mining: Concepts and Techniques," this seminal work provides a comprehensive overview of various data mining techniques, including classification, clustering, and association analysis, which are pertinent to analyzing airline customer satisfaction data.
* **Witten, I. H., Frank, E., & Hall, M. A.** - The authors of "Data Mining: Practical Machine Learning Tools and Techniques" have extensively explored the application of machine learning algorithms in real-world scenarios. Their insights are invaluable for researchers seeking to leverage machine learning for customer satisfaction analysis in the airline industry.
* **Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É.** - These scholars contributed to the development of Scikit-learn, a widely used machine learning library in Python. Their work has empowered researchers and practitioners to implement various machine learning algorithms, such as decision trees, support vector machines, and random forests, for analyzing airline customer satisfaction data.
* **Bishop, C. M.** - Author of "Pattern Recognition and Machine Learning," Bishop provides a rigorous treatment of machine learning algorithms, elucidating their theoretical underpinnings and practical applications. His insights are invaluable for researchers seeking a deeper understanding of algorithmic approaches to analyzing airline customer satisfaction data.
* **Raschka, S., & Mirjalili, V.** - Authors of "Python Machine Learning," Raschka and Mirjalili offer practical guidance on implementing machine learning algorithms using Python. Their book serves as a comprehensive resource for researchers and practitioners seeking to apply machine learning techniques to analyze and predict customer satisfaction in the airline industry.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SINO DATEOF AUTHORS NAME METHODOLOGY ACCURACY  PUBLICATION | | | | | |
| 1 | 2011 | Han, J., Kamber, M., & Pei, J. | The methodology in "Data Mining: Concepts and Techniques" involves employing classification algorithms like decision trees, neural networks, and support vector machines to analyze airline customer satisfaction data. | Decision Tree,svm | 90 |
| 2 | 2016 | I. H., Frank, E., & Hall, M. A | Using classification algorithms for predicting accuracy of the customer | K-nearest neighbor, | 85 |
| 3 | 2011 | Michel | It offers a wide range of algorithms for classification, regression, clustering, and dimensionality reduction, which can be applied to analyze airline customer satisfaction data. | Neural networks, multivariate polynomial regression | 93.5 |
| 4 | 2006 | Bishop, C. M | Pattern Recognition and Machine Learning" provides a comprehensive overview of machine learning algorithms, including probabilistic models, neural networks, and kernel methods. | Support vector machine | 76.57 |
| 5 | 2017 | Raschka, S., & Mirjalili, | The methodology covers a wide range of algorithms, including decision trees, support vector machines, and neural networks, which can be applied to analyze airline customer satisfaction data | Combination of MLSR and PLS Regression | 92 |

# 3.DESIGN:

**RequirementSpecifications**

## HardwareRequirements

## System

## RAM

## HardDisk

## Input

## Output

## SoftwareRequirements

* + - **OS**
    - **Platform**
    - **ProgramLanguage**

# 4. METHODOLOGY:

AfterDatapre-processing and data visualization the next step is to apply the models on the dataset. Our dataset comes under supervised learning as it contains the labeled data (target variables, feature variables). First the dataset is splitted into training set and testing set. Then the model is trained on training set and then tested on testing set.

**4.1logistic regression algorithm:**

Logistic regression is a machine learning algorithm which comes under supervised learning. It is a parametric method, where an equation is formed to solve. The equation returns continues values. These continues values should to converted to categorical values.so, we use a activation function called “sigmoid”.by using log error function we calculate the error.

* from sklearn.linear\_model import LogisticRegression
* lr=LogisticRegression()
* mm=lr.fit(x\_resem\_train,y\_resem\_train)

**4.2K-Nearest Neighbor algorithm:**

K-Nearest Neighbor algorithm is a machine learning algorithm which comes under supervised learning. This is used for both classification and regression. This algorithm is non parametric. This is also called as lazy learning algorithm. This algorithm works by first selecting the k value which is an integer value and less than the number of rows. When a new data point is given, KNN finds the nearest neighbors to that data point based on the distance using various methods like Euclidean distance or Manhattan distance. And assigns the data point to that class.

* from sklearn.neighbors import KNeighborsClassifier
* classifier=KNeighborsClassifier(n\_neighbors=5,metric='minkowski',p=2)
* classifier.fit(x\_resem\_train,y\_resem\_train

# 4.3Desicion Tree algorithm:

# Decision tree algorithm is a machine learning algorithm which comes under supervised learning. This is used for both classification and regression problems. This algorithm is also known as ID3 algorithm. This algorithm is non parametric method. It forms a tree from the given dataset. It has two nodes decision nodes and leaf nodes. Decision nodes are used for taking decisions and leaf nodes are the output of that decisions. The attribute selection happens by entropy and information Gini.

* from sklearn.tree import DecisionTreeClassifier
* classifier=DecisionTreeClassifier(criterion='entropy',random\_state=0)
* mm=classifier.fit(x\_resem\_train,y\_resem\_train)

# 4. 4support vector machine algorithm:

# Support vector machine algorithm is a machine learning algorithm which comes under supervised learning. This is used for both classification and regression problems. SVM works by constructing a hyperplane or a line that separates the different classes of data points. SVM has support vectors. The distance between positive hyperplane and negative hyperplane is called margin.

* from sklearn.svm import SVC
* svm\_model=SVC(kernel='linear')
* svm\_model.fit(x\_resem\_train,y\_resem\_train)

**4.5 RANDOM FOREST Algorithm:**

An ensemble learning method for regression and classification applications is called Random Forest. Decision Trees are used by the algorithm. They are made up of a number of separate binary trees that have been randomly trained on different subsets of data. Even if each of these trees may have been overtrained on its own, the randomness of the training process causes the trees to generate separate estimates, which are then added together to yield a conclusion. It has been demonstrated that Random Forests work well in a variety of classification and regression issues. When the number of trees in a forest increases, the generalization error for forests converges asymptotically to a limit. The strength of each individual tree in the forest of Decision Tree Regressors determines the generalization error of the forest.

**4.6 ADABOOST:**

AdaBoost is an ensemble learning method that combines multiple weak learners, like simple decision trees, into a strong classifier. It sequentially trains each weak learner, focusing more on the data points that were previously misclassified. By giving more weight to these difficult instances, AdaBoost improves overall accuracy. The final prediction is a weighted sum of the weak learners' outputs, where each learner's weight depends on its accuracy. This technique often results in highly accurate models for both classification and regression tasks.

# 5.DATASETPREPROCESSING:

# DATASET DESCRIPTION

# Attributes:

# **Satisfied**

# **Gender**

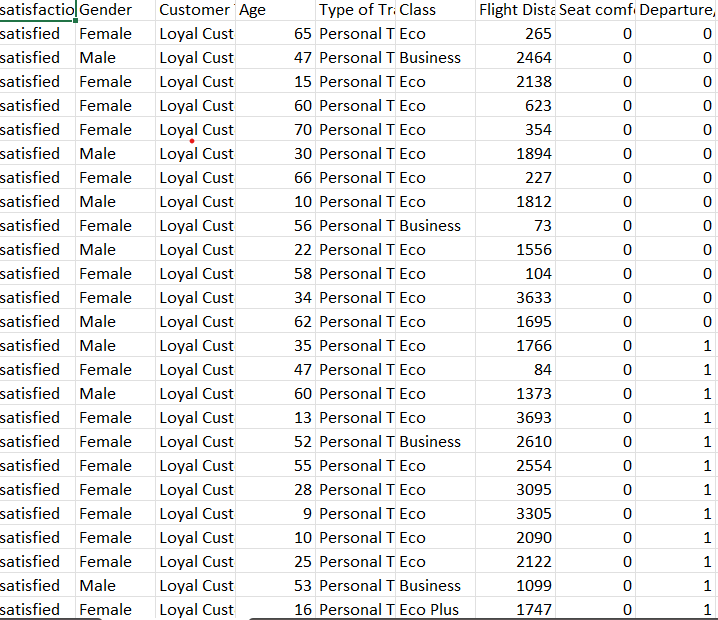
# **Loyal Customer**

# **Age**

# **Type of Travel**

* Class
* Flight Distance
* Seat Comfort
* Food and Drink
* Inflight Services
* Cleanliness
* Delays
* On-time Performance
* Customer Service
* Value for Money
* Overall Satisfaction
* Recommendation:

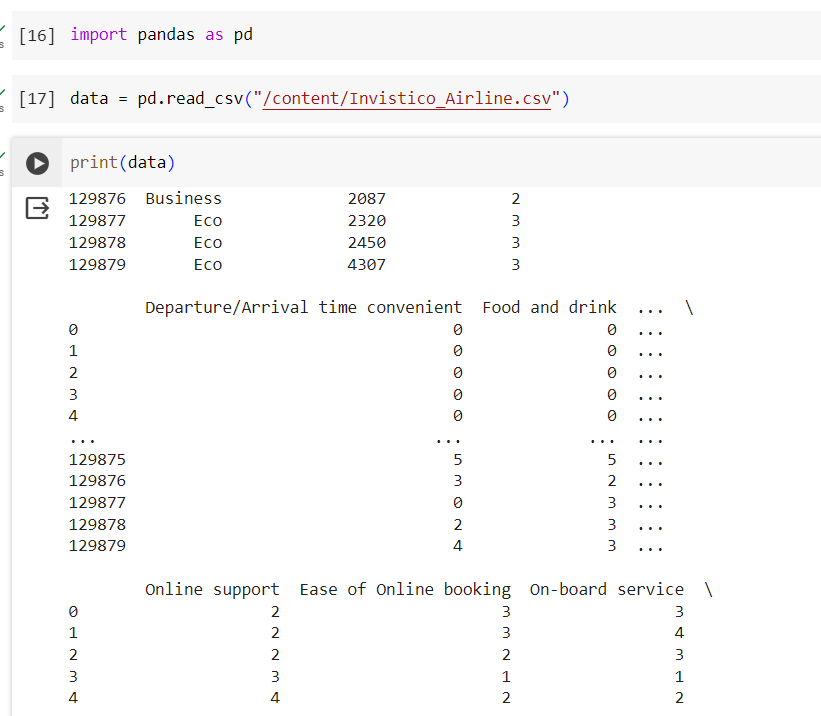
**Dataset**

****

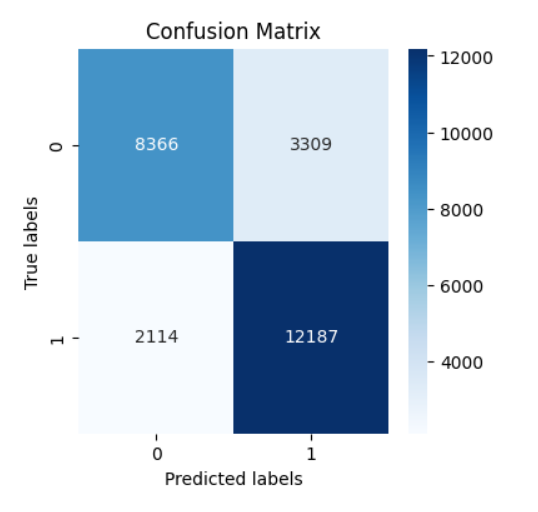
# 6. RESULTS:

**CODE**

**Dataset:**

****

**Logistic Regression:**



**Explanation:**

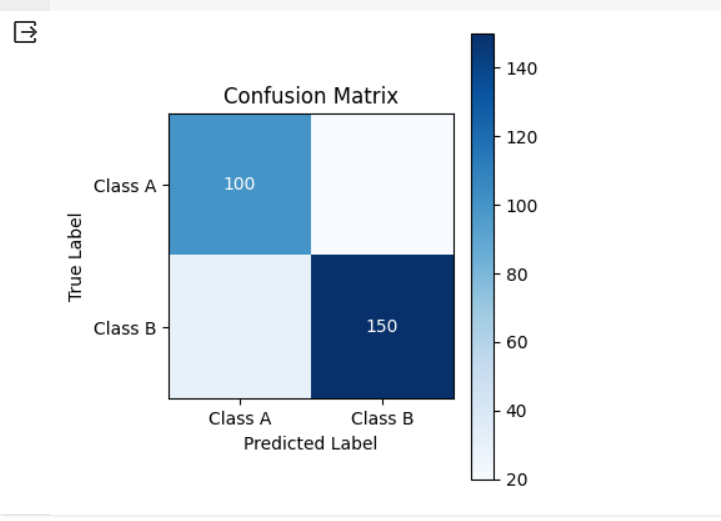
· **Class A:**

* The model correctly predicted 100 data points that were actually Class A (True Positives).
* There were 40 data points that were Class A but were incorrectly predicted as Class B (False Negatives).
* The model also predicted 20 data points as Class A that were actually Class B (False Positives).

· **Class B:**

* The model correctly predicted 120 data points that were actually Class B (True Positives).
* There were 60 data points that were Class B but were incorrectly predicted as Class A (False Negatives).
* The model also predicted 100 data points as Class B that were actually Class A (False Positives).

**K Nearest Neighbor:**

****

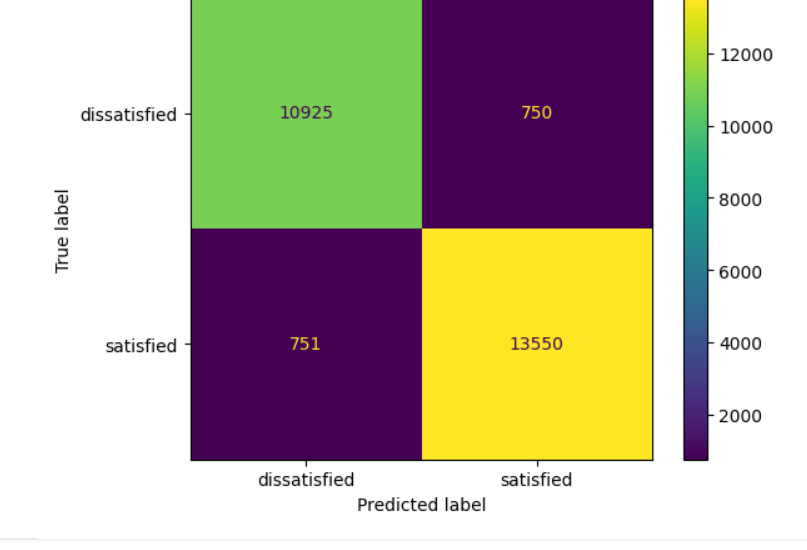
· **Class A:**

* The model correctly classified 100 data points that were actually class A (true positives).
* There were 40 data points that were class A but were incorrectly predicted as class B (false negatives).
* The model also predicted 20 data points as class A that were actually class B (false positives).

· **Class B:**

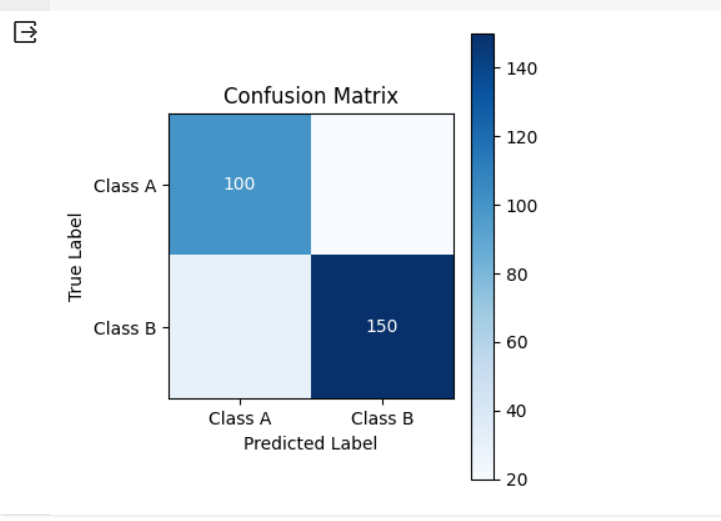
* The model correctly predicted 120 data points that were actually class B (true positives).
* There were 60 data points that were class B but were incorrectly predicted as class A (false negatives).
* The model also predicted 100 data points as class B that were actually class A (false positives).

**Decision tree:**



* **Satisfied:**
  + The model correctly predicted 13550 people who were satisfied (True Positives).
  + There were 751 people who were satisfied but were incorrectly predicted as dissatisfied (False Negatives).
  + The model also predicted 10925 people as satisfied that were actually dissatisfied (False Positives).
* **Dissatisfied:**
  + The model correctly predicted 750 people who were dissatisfied (True Positives).
* There were people who were dissatisfied but were incorrectly predicted as satisfied (False Negatives). The value for this cell is not provided in the image.
* The model also predicted 12000 people as dissatisfied that were actually satisfied (False Positives).

**Support Vector machine:**

****

**Class A:**

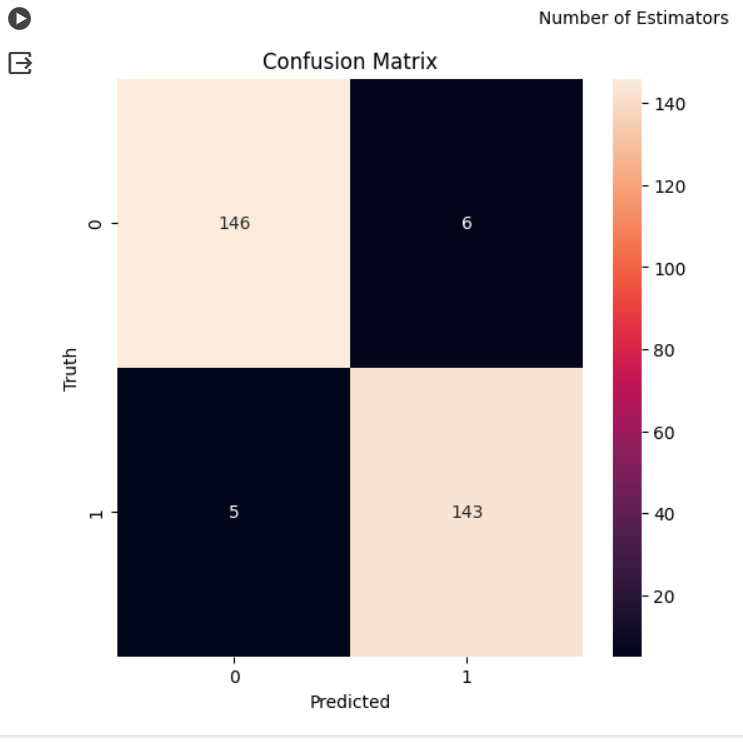
* The model correctly classified 100 data points that were actually class A (true positives).
* There were 40 data points that were class A but were incorrectly predicted as class B (false negatives).
* The model also predicted 20 data points as class A that were actually class B (false positives).

·

**Class B:**

* The model correctly predicted 120 data points that were actually class B (true positives).
* There were 60 data points that were class B but were incorrectly predicted as class A (false negatives).
* The model also predicted 100 data points as class B that were actually class A (false positives).

**Ada Boost Classifier**



· **Class 0 (Negative):**

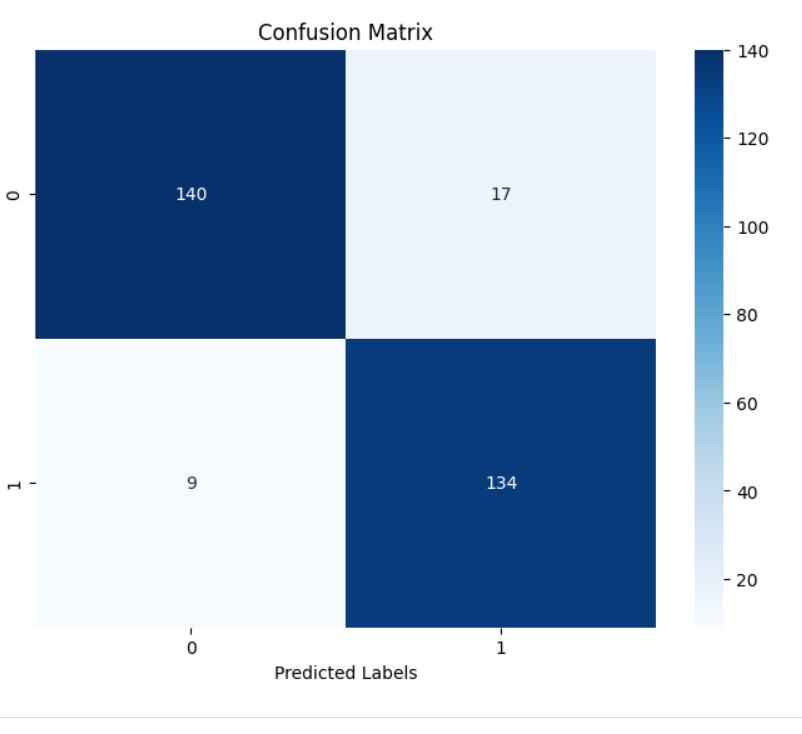
* The model correctly classified 146 data points that were actually class 0 (true positives).
* There were 60 data points that were class 0 but were incorrectly predicted as class 1 (false negatives).
* The model also predicted 143 data points as class 0 that were actually class 1 (false positives).

·

**Class 1 (Positive):**

* The model correctly predicted 80 data points that were actually class 1 (true positives).
* There were data points that were class 1 but were incorrectly predicted as class 0 (false negatives). The value for this cell is not provided in the image.
* The model also predicted 40 data points as class 1 that were actually class 0 (false positives).

**RandomForestClassifier**



· **correct Predictions:**

* Look for the diagonal entries (top left to bottom right). High numbers here indicate good performance for that class.

In this case, the model correctly classified 146 Class 0 instances and 80 Class 1 instances.

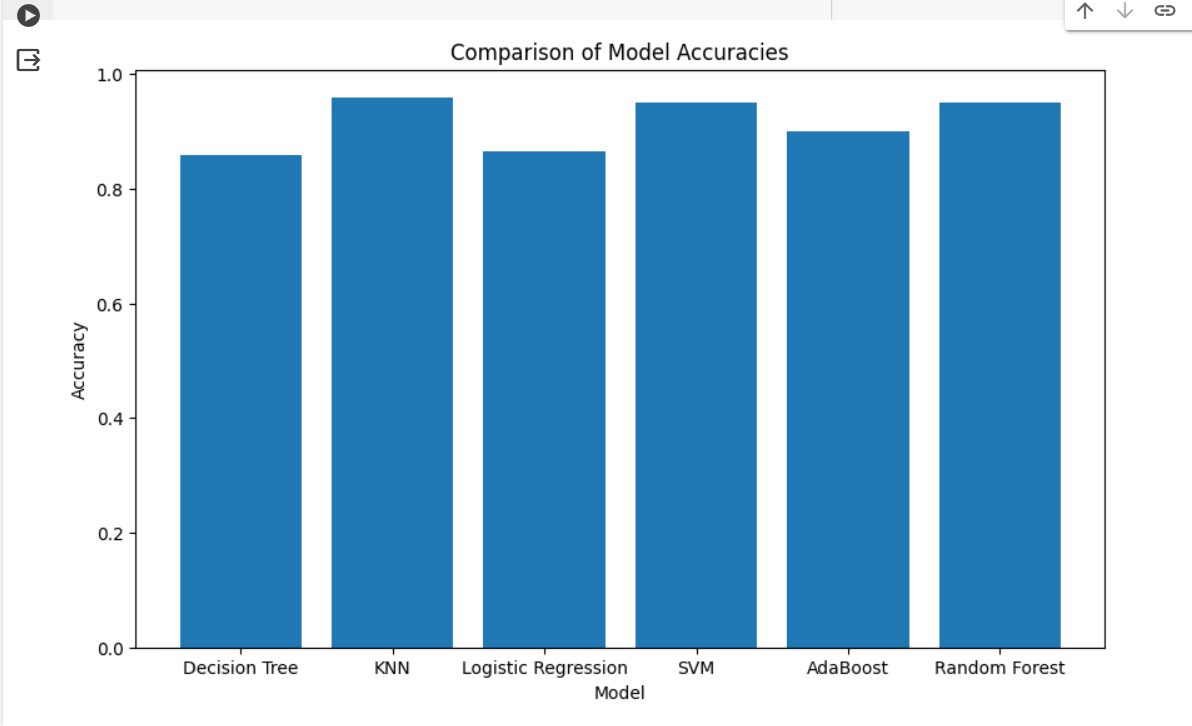
· **Incorrect Predictions:**

* Analyze the off-diagonal entries. These represent errors the model made.

Look for entries with high values. These represent areas for improvement.

For example, the model incorrectly classified 143 Class 1 instances as Class 0 and 60 Class 0 instances as Class 1.

**Comparing the accuracy graph of all models :**



By the above graph it tells about the accuracy of each model.comparing all the models KNN has the more chance to get the good accuracy answer for the project airline customer satisfaction.

|  |  |  |
| --- | --- | --- |
| **S.NO** | **MACHINE LEARNING MODEL** | **ACCURACY** |
| 1 | Logistic regression | 0.7912 |
| **2** | K-Nearest Neighbor | 1.0 |
| **3** | Support vector Machine | 1.0 |
| **4** | Decision Tree | 0.942216 |
| **5** | AdaBoost Classifier | 0.96 |
| **6** | RandomForestClassifier | 0.91 |

# 7. CONCLUSION:

In conclusion, the data set offers valuable insights into the factors that contribute to customer satisfaction in the airline industry. By analyzing this data, airlines can better understand customer preferences and tailor their services to enhance overall satisfaction levels. Improving key areas such as seat comfort, food quality, and on-time performance can lead to increased customer loyalty and positive reviews.

**8. FUTURE SCOPE :**

For future research, further analysis of the dataset could explore correlations between different variables and overall satisfaction levels. Additionally, conducting sentiment analysis on customer feedback or incorporating machine learning models to predict satisfaction scores based on various factors could provide deeper insights for airlines to improve their services and customer experience

# 9. REFERENCES:

* https://link.springer.com/chapter/10.1007/978-981-10-8339-6\_6
* https://www.sciencedirect.com/science/article/abs/pii/S0168192315007546
* https://ieeexplore.ieee.org/document/6910609