MACHINE LEARNING

MINI PROJECT
ON

AIRLINE FLIGHT DELAY PREDICTION

INTRODUCTION

As global air travel continues to expand, the frequency of flight delays has also risen, causing significant inconvenience for passengers and substantial financial strain on airlines. This growth exacerbates the crowded situation at airports and causes financial difficulties within the airline industry. Air transportation delay indicates the lack of efficiency of the aviation system. It is a high cost to both airline companies and their passengers. In 2007 alone, flight delays in the United States resulted in an estimated economic cost of $32.9 billion, which included a $4 billion reduction in GDP, according to the Total Delay Impact Study. These delays highlight inefficiencies in the aviation system that need to be addressed to improve airline operations and enhance passenger satisfaction. Accurately predicting flight delays is crucial for mitigating these negative impacts and optimizing the overall efficiency of air transportation.

This project seeks to address this issue by comparing the performance of two machine learning classification algorithms in predicting flight delays. Using a dataset of flights departing from the airport over one year, the study evaluates the effectiveness of K-Nearest Neighbor (KNN) and Decision Tree. Performance metrics such as accuracy, precision, recall, and F1-score are used to assess each algorithm's predictive capabilities. By identifying the most effective models, this project aims to provide actionable insights for improving airline operations and passenger experiences, ultimately contributing to a more efficient and reliable air travel system.

DATASET DESCRIPTION

The dataset used for this study comprises flight information for departures from John F. Kennedy International Airport (JFK) over a one-year period, from November 2019 to December 2020. Sourced from an open data repository on Kaggle, it includes 28,820 records with 23 attributes per flight. Key attributes include the month, day of the week, carrier code, flight number, destination, departure delay, scheduled and actual departure times, scheduled arrival time, taxi-out time, and various weather conditions such as temperature, humidity, wind speed, and pressure. To ensure the dataset's suitability for machine learning analysis, minor adjustments were made, such as converting categorical variables to numerical values and removing rows with missing data, which comprised only two entries, thus minimally impacting the dataset's distribution. A detailed description of data set attributes is presented in Table:

Table 1. Attribute description for the data set.

Attribute Name	Description	Type
MONTH	Month	Integer
DAY_OF_MONTH	Date of flight	Integer
DAY_OF_WEEK	Day of the week	Integer
OP_UNIQUE_CARRIER	Carrier code that represents the carrier company	Object
TAIL_NUM	Air flight number	Object
DEST	Destination	Object
DEP_DELAY	Departure delay of the flight	Integer
CRS_ELAPSED_TIME	Scheduled journey time of the flight	Integer
DISTANCE	Distance of the flight	Integer
CRS_DEP_M	Scheduled departure time	Integer
DEP_TIME_M	Actual departure time	Integer
CRS_ARR_M	Scheduled arrival time	Integer
Temperature	Temperature	Integer
Dew Point	Dew Point	Object
Humidity	Humidity	Integer
Wind	Wind direction	Object
Wind Speed	Wind speed	Integer
Wind Gust	Wind gust	Integer
Pressure	Pressure	Floating Point
Condition	Condition of the climate	Object
sch_dep	Number of flights scheduled for departure	Integer
sch_arr	Number of flights scheduled for arrival	Integer
TAXI_OUT	Taxi-out time	Integer

DATA PROCESSING

To prepare the dataset for machine learning analysis, several data processing steps were undertaken. Initially, the dataset was cleaned by addressing missing values; specifically, two rows with missing data were removed, which had a negligible impact on the overall distribution given the dataset's size. Categorical variables were converted to numerical values using integer encoding to ensure compatibility with algorithms that only handle numerical data. This included transforming the "DEW_POINT" variable from an object type to an integer type. Variables with minimal influence on predicting flight delays, such as "TAIL_NUM," were dropped to streamline the dataset.

The target variable "DEP_DELAY" was used to create a binary classification label "IS_DELAY," where flights delayed by more than 15 minutes were marked as 1 (delayed) and those with lesser or no delays were marked as 0 (not delayed). Based on the variable “IS_DELAY”, it can be seen that the data set consists of 3873 delayed flights and 24945 non-delayed flights, showing an imbalanced distribution since the majority of flights were not delayed. Furthermore, weighted precision, recall, and F1-score were used to provide a more accurate evaluation of the model performance, considering the imbalanced nature of the dataset. These preprocessing steps ensured that the data was in an optimal format for training and evaluating the chosen machine learning algorithms. The formulas for weighted precision, recall, and F1 are:

$$Precision = \frac{TP}{TP+FP} \times \frac{TP+FN}{P+TN+FP+FN} + \frac{TN}{TN+FN} \times \frac{TN+FP}{P+TN+FP+FN}$$

$$Recall = \frac{TP}{TP+FN} \times \frac{TP+FN}{P+TN+FP+FN} + \frac{TN}{TN+FP} \times \frac{TN+FP}{P+TN+FP+FN}$$

$$F1 - Score = 2 \times \frac{\frac{TP}{TP+FP} \times \frac{TP}{TP+FN}}{\frac{TP}{TP+FP} + \frac{TP}{TP+FN}} \times \frac{TP+FN}{P+TN+FP+FN} + 2 \times \frac{\frac{TN}{TN+FN} \times \frac{TN}{TN+FP}}{\frac{TN}{TN+FN} + \frac{TN}{TN+FP}} \times \frac{TN+FP}{P+TN+FP+FN}$$

IMPLEMENTATION

Decision Tree Algorithm:

The Decision Tree model was created using the scikit-learn library in Python, which involves a series of steps to train and evaluate the model effectively. Initially, the preprocessed dataset was split into training and testing sets to ensure the model's robustness and mitigate the impact of data imbalance. The Decision Tree algorithm was chosen for its ability to handle both numerical and categorical data, making it well-suited for the diverse attributes in the flight delay dataset. The model construction began by selecting the best attribute to split the data at each node, starting from the root, based on criteria like entropy to maximize information gain. Each node represents a decision point, and branches signify the outcomes leading to further splits or terminal leaf nodes, which indicate the final prediction. The trained Decision Tree model was then tested on the unseen data, and its performance was evaluated using metrics such as accuracy, precision, recall, and F1-score. The results demonstrated that the Decision Tree model achieved the highest scores among all the algorithms tested, indicating its strong predictive capability for flight delays in this dataset.

K-Nearest Neighbors (KNN) Algorithm:

The K-Nearest Neighbors (KNN) model was developed using the scikit-learn library in Python, focusing on classifying flight delays based on the closest training examples in the feature space. The preprocessing steps ensured the dataset was ready for analysis, converting categorical variables to numerical ones and normalizing the data to ensure all features contributed equally to the distance calculations. The model creation involved selecting an optimal value for k, the number of nearest neighbors, which was determined through cross-validation to balance bias and variance effectively. During the training phase, the KNN algorithm memorized the entire training dataset without explicitly learning a model. For each flight in the test set, the algorithm calculated the Euclidean distance to all training examples, identifying the k closest neighbors. The majority class among these neighbors was then assigned as the predicted class for the test instance. The KNN model's performance was assessed using accuracy, precision, recall, and F1-score, revealing that while it provided reasonable predictions, its accuracy and precision were lower compared to Decision Tree algorithm, highlighting its limitations in handling the imbalanced and diverse nature of the flight delay dataset.

RESULTS

First 5 rows of the dataset:

	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	DEP_DELAY	CRS_ELAPSED_TIME	DISTANCE	\
0	11	1	5	-1	124	636	
1	11	1	5	-7	371	2475	
2	11	1	5	40	181	1069	
3	11	1	5	-2	168	944	
4	11	1	5	-4	139	760	

	CRS_DEP_M	DEP_TIME_M	CRS_ARR_M	Temperature	...	Wind_S	Wind_SE	\
0	324	323	448	48	...	False	False	
1	340	333	531	48	...	False	False	
2	301	341	482	48	...	False	False	
3	345	343	513	48	...	False	False	
4	360	356	499	46	...	False	False	

	Wind_SSE	Wind_SSW	Wind_SW	Wind_VAR	Wind_W	Wind_WNW	Wind_WSW	IS_DELAY
0	False	False	False	False	True	False	False	0
1	False	False	False	False	True	False	False	0
2	False	False	False	False	True	False	False	1
3	False	False	False	False	True	False	False	0
4	False	False	False	False	True	False	False	0

[5 rows x 2204 columns]

Decision Tree:
Accuracy: 0.9583
Precision: 0.9573
Recall: 0.9583
F1 Score: 0.9576

KNN:
Accuracy: 0.9223
Precision: 0.9282
Recall: 0.9223
F1 Score: 0.9093

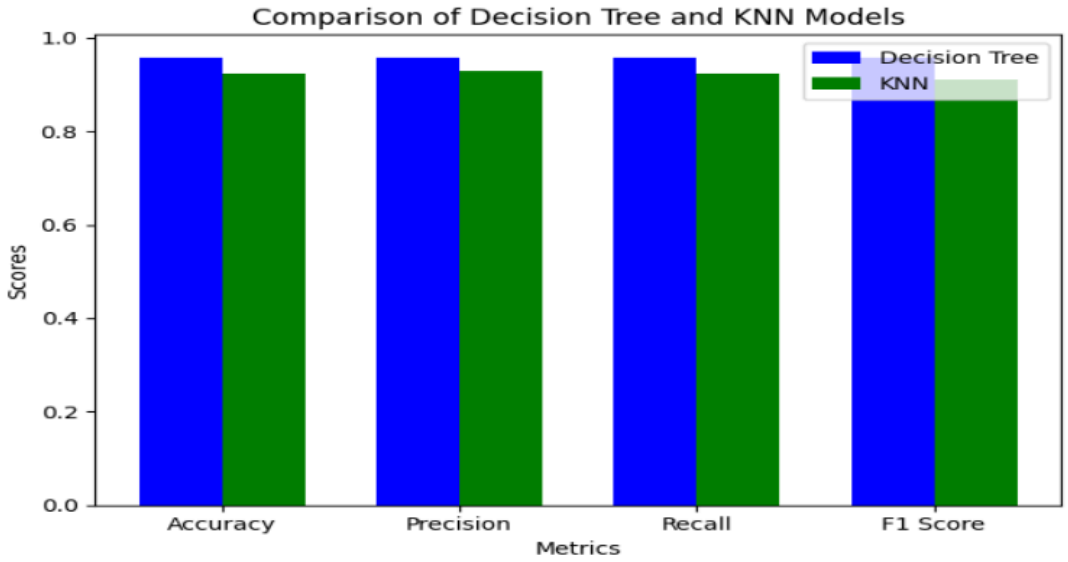

Fig1: Results of Airline Flight Delay Prediction using Decision Tree and K-NN Algorithm

CONCLUSION

This study aimed to predict flight delays at John F. Kennedy International Airport by comparing the performance of two machine learning algorithms: K-Nearest Neighbor (KNN) and Decision Tree Algorithms. Using a one-year dataset of flights departing from JFK, the models were evaluated based on accuracy, precision, recall, and F1-score, with a particular focus on handling the imbalanced nature of the data.

The results demonstrated that the Decision Tree model achieved higher performance across all metrics, indicating its strong capability in predicting flight delays. In contrast, the KNN model exhibited a relatively lowest performance, suggesting they were less suited to the complexities of the dataset.

In conclusion, the Decision Tree model is highly effective for predicting flight delays in the given dataset. Future research could explore additional ensemble techniques and address the data imbalance using advanced methods such as Synthetic Minority Over-sampling Technique (SMOTE) to further enhance prediction accuracy. These findings can significantly improve airline operations and passenger experiences by providing more reliable predictions of flight delays.