Twitter Airline Sentiment Analysis Report

1. Introduction

Project Overview
The goal of this project is to analyze the sentiment of tweets directed at various airlines. The sentiment analysis classifies tweets into three categories: positive, neutral, and negative.

Dataset Description
The dataset contains 14,640 tweets with the 15 columns:

2. Data Preprocessing

Data Cleaning

- autoclean() and klib.datacleaning() were used at the 1st attempt
- Removed columns that were missing 60% of the values.
- Handled missing values by filling them as appropriate.
- Converted the **airline_sentiment** column to numerical values using mapping as it is target value.
- at the 2nd attempt encoding was done manually
- scaling was done only on features assigning them to x variable and removed the target value temporarily to prevent it from scaling and becoming continuous value

3. Model Training

Splitted the dataset into training and test sets using an 80-20 split.

Model Selection
Two machine learning models were considered, including:

- RandomForestClassifier
- RandomForestRegressor

4. Model Evaluation

The following metrics were used:

- Mean squared error
- R2 score

- Classification report
- Accuracy
- Presicion
- F1 score

Results

Model	Accuracy	Precision	Recall	F1 score
Random forest classifier	0.85	0.86	0.72	0.85

Model	Mean squared error	R2 score
Random forest regressor	0.15280495390041446	0.8216019072875267

5. Conclusion

Summary
The Random Forest Classifier achieved the best performance with an accuracy of 85%, precision of 86%, recall of 72%, and an F1 score of 85%. The model effectively classified the sentiment of tweets directed at airlines.

6. Model improvement

Several attemps were done using different techniques, such as:

- Using klib, autoclean
- Feature transforming (creating new columns for date splitting into year, month, day)
- KFold Cross validation score tools