Lab 8: Flight Delay Prediction
Name: Bonaobra, Perez, Jazmin
Date: April 11, 2025
GitHub/Colab Link: https://github.com/arnthegreat/Bonaobra_Elec2.git

Objective

The main goal of this project was to predict whether a flight would be delayed, based on basic flight details. We also wanted to see how different machine learning models compared in terms of accuracy and performance.

Introduction

Flight delays can be frustrating for both passengers and airlines. In this project, we used Apache Spark to analyze and predict flight delays using just a few key pieces of information: the airline, the airport, and the month the flight took place. We tested three different machine learning models to figure out which one performed the best.

Methodology

Here's a step-by-step overview of what we did:

1. **Set Up Spark:** We started by launching a Spark session.
2. **Loaded the Data:** We brought in our flight delay data from a CSV file.
3. **Filtered the Flights:** To keep things manageable, we focused only on flights going to four specific airports.
4. **Labeled the Data:** We marked a flight as "delayed" if it arrived more than 15 minutes late.
5. **Created Features:** We used the month, airline, and destination airport as the input features for our models.
6. **Built the Pipeline:** We converted text into numbers, assembled feature vectors, and set up a machine learning pipeline.
7. **Split the Data:** We divided our dataset into 80% training and 20% testing sets.
8. **Trained the Models:** We trained three types of models—Logistic Regression, Random Forest, and Gradient Boosted Trees.

Results & Analysis

All three models were able to predict delays with pretty good accuracy. Gradient Boosted Trees had the best performance, but it also took the longest to train. Logistic Regression was the quickest, while Random Forest gave solid, balanced results.

Challenges & Solutions

One challenge was converting text data (like airline names and airport codes) into numerical formats that the models could understand. Thankfully, Spark has tools that made this easier. Another challenge was narrowing down which airports to include—

we had lots to choose from, but picked four to keep the process simple and the tests faster. Some models also took a while to train, so we tweaked their settings to speed things up.

Conclusion

This project showed that it’s totally possible to predict flight delays using Spark and some basic flight data. Gradient Boosted Trees gave the best results, but each model had its strengths. With more data and by adding extra features—like the flight time or weather—we could probably make even better predictions in the future.