**APACHE SPARK**

**Background**

Flight delay is a significant problem that negatively impacts the aviation industry and costs billions of dollars each year. Most existing studies investigated this issue using various methods based on applying machine learning methods to predict the flight delay. However, due to the highly dynamic environments of the aviation industry, relying only on single route of airport may not be sufficient and applicable to forecast the future of flights. The purpose of this project is to analyze a broader scope of factors which may potentially influence the flight delay it compares several machine learning-based models in designed generalized flight delay prediction tasks. In this project we have used flight delay dataset from US Department of Transportation (DOT) to predict flight delays. We have used supervised learning algorithms to predict flight departure delay and then model evaluation is done to get best model and our model can identify which features were more important when predicting flight delays.

**Theoretical Framework**

Apache Spark (Spark) is an open-source data-processing engine for large data sets. It is designed to deliver the computational speed, scalability, and programmability required for Big Data—specifically for streaming data, graph data, machine learning, and artificial intelligence (AI) applications.

Spark's analytics engine processes data 10 to 100 times faster than alternatives. It scales by distributing processing work across large clusters of computers, with built-in parallelism and fault tolerance. It even includes APIs for programming languages that are popular among data analysts and data scientists, including Scala, Java, Python, and R.

Spark is often compared to Apache Hadoop, and specifically to MapReduce, Hadoop’s native data-processing component. The chief difference between Spark and MapReduce is that Spark processes and keeps the data in memory for subsequent steps—without writing to or reading from disk—which results in dramatically faster processing speeds.

**Methodology**

The goal of this project is to create a machine learning model (logistic regression and random forest) and predict if a flight will be delayed over 15 minutes using Apache Spark.

The steps we are going to perform are:

1. Loading the data into Apache Spark
2. Reading and processing data with Pyspark
3. Using Logistic Regression and Random Forest to predict delayed flights

We also aim to build a machine learning pipeline with PySpark. For that we have used pyspark.ml.

At the core of the pyspark.ml module are the Transformer and Estimator classes.

Transformer classes have a.transform() method that takes a Data Frame and returns a new Data Frame; usually the original one with a new column appended. For example, you might use the class Bucketizer to create discrete bins from a continuous feature or the class PCA to reduce the dimensionality of your dataset using principal component analysis.

Estimator classes all implement a .fit() method. These methods also take a DataFrame, but instead of returning another DataFrame they return a model object. This can be something like a StringIndexerModel for including categorical data saved as strings in your models, or a RandomForestModel that uses the random forest algorithm for classification or regression.

Pipeline is a class which now can be used in the pyspark.ml module that combines all the Estimators and Transformers

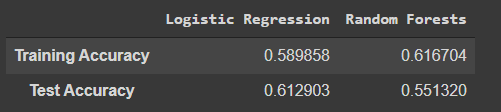
**Results:**

LINK TO CODE: [(here)](https://github.com/ashish-3916/academic_nsut/blob/main/sem%206/computer%20H-S/Apache%20Sparks/flight_delay_prediction.ipynb)

LINK TO COLLAB: [(here)](https://colab.research.google.com/drive/1gSF1pOma_ce6r7-YDUUE2ZzHh3Ok0YLy)

**OUTPUT:**

Without cross validation:



With cross validation:

