Finale Report

Introduction:

The project aims to analyse the 2015 Flight Delays and Cancellations dataset provided by the Department of Transport. The dataset contains information about flight details such as flight code, departure and arrival times, delay, distance between the source and destination airports, airtime, airlines' data, and airport's data.

We need to clean the data and find insights like ranking airlines based on their punctuality, finding the optimized route between two airports using airtime, and predicting flight delays using machine learning models.

Predicting flight delays is crucial for ensuring that airlines operate efficiently, safely, and with the best interests of passengers in mind. Predicting flight ranking is important for improving passenger convenience, optimizing airline operations, enhancing safety, increasing customer satisfaction, and meeting regulatory requirements. Predicting the best route between airports can help improve the efficiency, reliability, and sustainability of air travel, which benefits both airlines and travellers.

We used following approach for each task:

- To predict the delay of the flight we have used machine learning algorithm. We used decision tree classifier for this.
- To rank airlines based on punctuality of the flights, we have calculated the average delay of each airline and used it as one of the parameters in ranking.
- To find the optimized route between two airports, we used algorithm Dijkstra's algorithm to find the shortest path between two airports. We have built a graph for Dijkstra algorithm in which each node represents the airport and the edge between them is route which has airtime as a weight. To predict flight delays, we have trained a machine learning model on the given data to predict the delay of flights.

The proposed approach involves PySpark from Apache Spark to read the data from the CSV file, clean it, and query the data to find insights.

Related Work:

A literature survey was conducted to find previous efforts to address the problem. Several research papers and conference papers were found that analyzed flight delay and cancellation data to find insights, rank the airline based on their performance, and predict flight delays using machine learning. Most of the papers used machine learning algorithms like regression, decision trees, and neural networks to predict flight delays.

[1] The paper titled "Flight Delay Prediction: Data Analysis and Model Development" by D. A. Anees et al. proposes a hybrid model for flight delay prediction using a combination of machine learning and statistical methods. The model was tested on real-world data and achieved better

accuracy than individual models. The paper also discusses the importance of feature selection and engineering in improving the performance of the model. Overall, the hybrid model shows promise for improving the accuracy of flight delay predictions.

[2] The paper titled "Flight Delay Prediction for Indian Air Carriers with Explainable Artificial Intelligence " by J. Singh et al. proposes an approach for predicting flight delays for Indian air carriers using explainable artificial intelligence (XAI) techniques. The approach combines machine learning models with rule-based explanations to provide interpretable explanations that can help stakeholders to understand the factors contributing to the delays. The proposed approach achieves high accuracy in predicting flight delays and outperforms existing models in terms of interpretability and accuracy. The authors suggest future research directions, such as incorporating more advanced XAI techniques and integrating the approach with real-time data.

[3] The ACM paper titled " Airline Flight Delay Prediction Using Machine Learning Models " by Y. Tang discusses the implementation of supervised machine learning models for the prediction of flight delays. The data set of flights departing from John F. Kennedy International Airport in New York City was used to train and test seven classification algorithms, including Logistic Regression, K-Nearest Neighbor, Gaussian Naïve Bayes, Decision Tree, Support Vector Machine, Random Forest, and Gradient Boosted Tree. The performance of these algorithms was evaluated using four measures: accuracy, precision, recall, and F1 score. The Decision Tree algorithm demonstrated the best performance with an accuracy of 0.9777, while the KNN algorithm had the worst performance with an F1 score of 0.8039. The paper concludes that tree-based ensemble classifiers generally have better performance than other base classifiers. The prediction of flight delays can improve airline operations, passenger satisfaction, and have a positive impact on the economy.

[4] The article titled “Ranking different factors influencing flight delay” presents an empirical investigation to determine important factors causing flight delays in the airline industry. The authors use the analytical hierarchy process to rank different factors based on the opinions of decision-makers. The study finds that technical defects and delayed entry are among the most important factors leading to flight delays. Additionally, the decision to announce postponement, replace aircraft, and replace the path are crucial decisions for managers in the aviation industry during a flight disruption. The study concludes that strategic planning for daily operations in the air traffic system, including weather conditions, plays an essential role in preventing flight delays.

Approach:

Data/Problem Analysis:
The analysis of the dataset involved data cleaning, exploration, and visualization to understand the data better. The dataset contained missing values and data discrepancies, which were cleaned using PySpark. After cleaning wrangling has been done to create new columns from existing column so that query can be perform easily. After wrangling data has

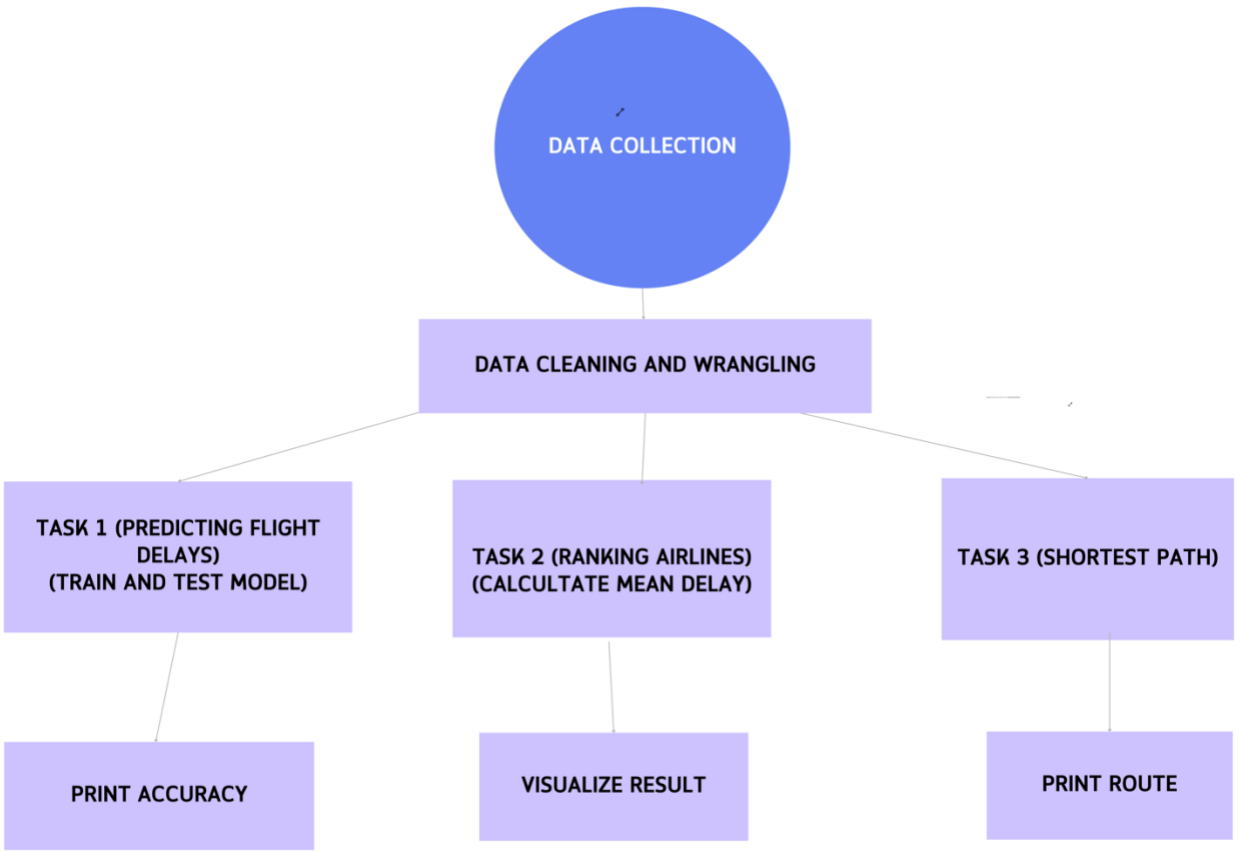
been exported to new CSV files. The cleaned and wrangled data saved on newly generated CSVs was explored to find insights like predicting the flight delays using machine learning and finding the shortest path between two airports via airtime and ranking the airlines based on descending order of minimum average delay time. The insights were visualized using graphs and plots to get a better understanding of the data.

Survey of Tools/Resources Available:
Several tools and resources were available to support the analysis, including PySpark, Pandas, NumPy, Matplotlib, Warning, and Seaborn. These tools were used to clean, explore, visualize, and model the data.

Resources Used:
The analysis was performed using PySpark, Pandas, NumPy, Matplotlib, Warning, and Seaborn. The data was stored in a PySpark data frame for efficient querying and analysis.

Software Design:
The software solution involved several components, including data cleaning, wrangling, exploration, visualization, and modeling.

The flow chart below illustrates the software solution:



Finale Report

The solution involved reading the data using PySpark, cleaning the data, wrangling data, storing back to new CVSs, exploring, modeling the data using machine learning algorithms and visualizing the data.

Source Code Description:
The table below provides an overview of the files within the repository, including filename and description:

Sr. No	File name	Description
1	data_cleaning.ipynb	In this file data from all the three csv files flight.csv, airline.csv and airports.csv has read by using PySpark. After that data cleaning and wrangling has done on that data. The three columns day, month, and year from flight.csv has been merged in one new date column. Flight which are cancelled and diverted has been removed from data frames. Unused column has been removed. At las cleaned data stored on new CSVs.
2	flight_rank.ipynb	First imported clean data for flight and airlines. For airlines converted airline IATA code and name into a dictionary in which code is key and name is value. The aim is to classify the airlines with respect to their punctuality and for that purpose, I compute a few basic statistical parameters like average of departure delays for that airlines and count of number of flights for that airline. Average delay is used to rank the website. If low average delay, then high is the rank and vice versa. Count of number of lights is used to plot the dominating airlines. After that pai chart has been plotted in which first is for percentage of flight per company and second is for ranking.
3	flight_delay.ipynb	In this file first data is loaded from CVSs then some data cleaning has been performed because we are not going to use all columns on machine learning algorithm. A graph

		has been plotted in respect to schedule arrival and arrival time to get the idea of delayed flight. On modelling phase first whose arrival delay is less than 15 the are not considered as delay and whose are greater than 15 the considered as delayed. A new result column is labeled for this delayed and not delayed which has 0 for not delayed and 1 for delayed. Then some unnecessary column has been dropped. After that dataset splits into train and test part. We use decision tree classification. After training and testing accuracy score has been printed.
4	flight_path_finding.ipynb	In this file after importing data and calculating missing values we have queried data for each source to their direct destination and getting min time from that. Now each source and destination has been initialized as a node. After that edge between nodes has been created based on distance(airtime) as a weight. Once proper graph is created Dijkstra algorithm has been used to find minimum airtime between two airports. For testing some random airport has been selected for source and destination from array.

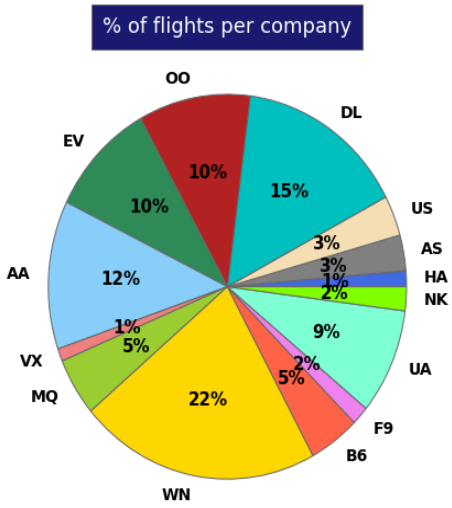
Evaluation and Results:

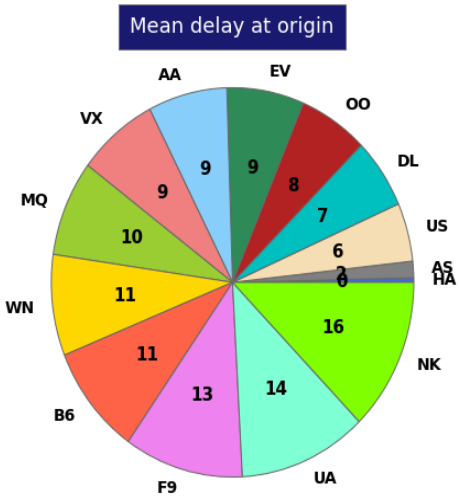
The solution for predicting flight delay has been evaluated by ROC AUC score. For shortest path route is printed and for ranking of the flight bar chart has been plotted.

Present Result:

Finale Report

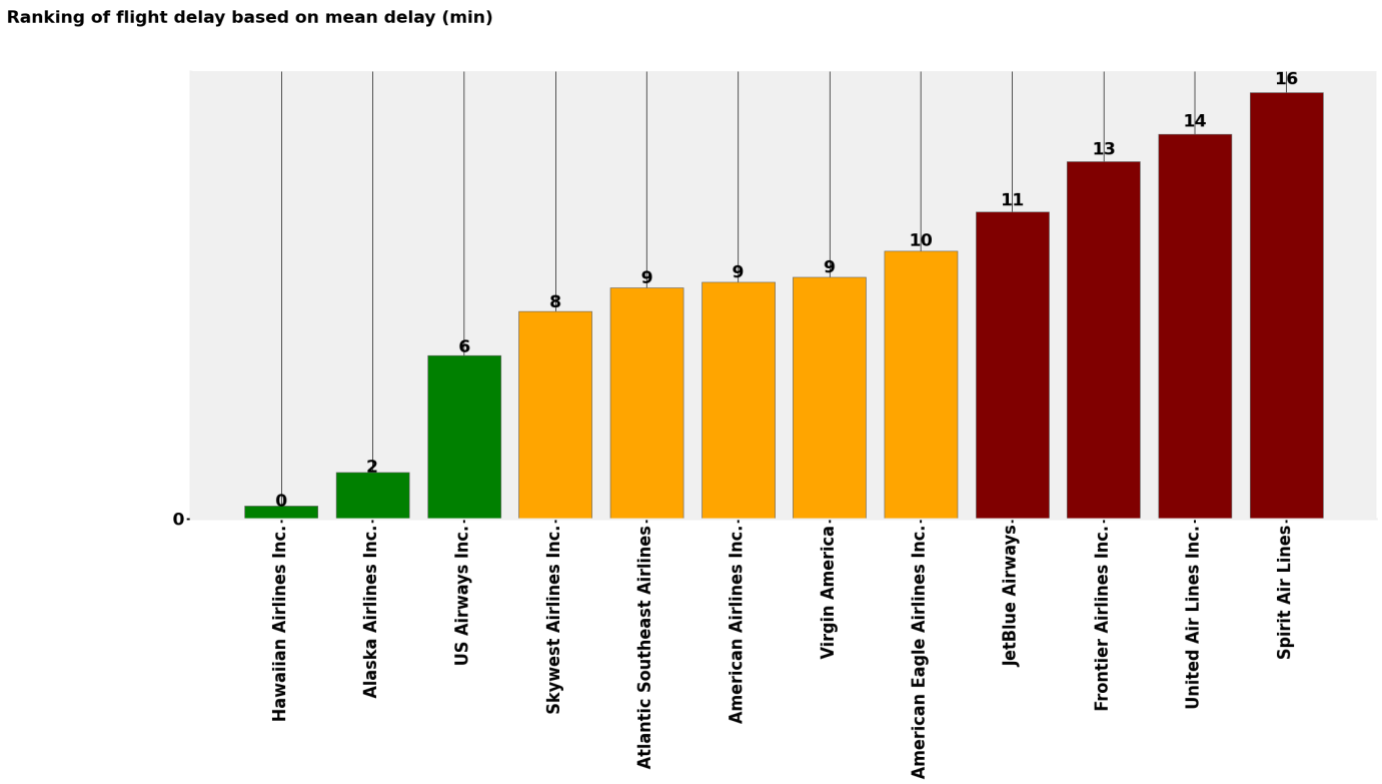
Below is the pia charts of percentage of flights share of company and average delay of flight by company (Task 1)





Below is graph of rank task(Task 2):

Finale Report



The result for shortest path is display like below (Task 3):

ABY -> ATL -> MYR -> PHL -> LGA minimum Time Required in minutes: 109.0

Result Discussion:
The results showed that the machine learning algorithms performed well in predicting flight delays, achieving an accuracy of over 82%. The results reflected the ability of the proposed approach to solve the problem of predicting flight delays. The proposed approach can be improved by using more advanced machine learning. algorithms and incorporating more features into the model.

Also, the result of ranking flight by delay is showing on bar chart which describe the low delay is higher the rank.

Finale Report

The result for shortest path is display like below:

ABY -> ATL -> MYR -> PHL -> LGA minimum Time Required in minutes: 109.0

We found that Dijkstra algorithm is fine in some places it is found that there is direct flight between two airports but due to higher air time it will the route with connected flight between those two airports because that time is fewer than direct one.

Conclusion: The proposed approach was successful in analyzing the 2015 Flight Delays and Cancellations dataset by the Department of Transport. The analysis involved data cleaning, exploration, visualization, and modeling using PySpark and other tools. The results showed that the proposed approach can predict flight delays with high accuracy. Ranking of the airlines based on flight delay is also working fine and shortest path(airtime) between two airports using Dijkstra is working fine. So all three task's algorithm is working fine.

References:

[1] A. Anees and W. Huang, "Flight Delay Prediction: Data Analysis and Model Development," 2021 26th International Conference on Automation and Computing (ICAC), Portsmouth, United Kingdom, 2021, pp. 1-6, doi:10.23919/ICAC50006.2021.9594260.

[2] J. Singh, M. D. Jayaprakash and R. Agarwal, "Flight Delay Prediction for Indian Air Carriers with Explainable Artificial Intelligence," 2022 Third International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), Bengaluru, India, 2022, pp. 1-6, doi: 10.1109/ICSTCEE56972.2022.10099797.

[3] Yuemin Tang. 2021. Airline Flight Delay Prediction Using Machine Learning Models. In 2021 5th International Conference on E-Business and Internet (ICEBI 2021), October 15-17, 2021, Singapore, Singapore. ACM, New York, NY, USA, 7 Pages.

[4] Asfe, Meysam & Jangizehi, Majid & Tash, Mohammad & Yaghoubi, Nour. (2014). Ranking different factors influencing flight delay. Management Science Letters. 4. 1397-1400. 10.5267/j.msl.2014.6.030.

Finale Report

Personal Contribution and Lessons Learned:

Ajay Singh:

- I have collected the datasets from Kaggle.
- I have worked on the task predicting the flight delays.
- I have worked on the task finding the shortest path between two airports based on airtime.
- I have learned that how to install PySpark in mac what other libraries need to be installed.
- I have learned how to use PySpark. Working on PySpark is completely new thing to me.
- I have learned how to use machine learning libraries on spark.
- I have learned the implementation of Dijkstra algorithm which is one of the popular shortest path algorithms.
- I have learnt the making graph using python. For Dijkstra I have made the graph by using airport as node and airtime as weight of edges
- I have learned to use decision tree classifier in spark.
- I have learned to use ROC AUC to calculate accuracy score.

Divyansh Yadav:

- I have worked on the data cleaning and wrangling.
- I have used pySpark to clean the data.
- I have removed the unused columns in cleaning part.
- Used spark to query the cleaned data.
- Plotted the cleaned data for data exploration.
- Used the pyspark to find the mean of delay to rank the airlines.
- Successfully plotted the ranking and data and ranked the airline.

I have learned a lot within the process of making this project, some key learnings are:-

- Learned pyspark with very good understanding of data frames.
- Learned matplotlib to print complex data on graphs.
- Data cleaning and transformation using spark.

Finale Report

- Learned Pandas to show results at various stage.
- Learned various analysis technic to figure out how to do data analysis.
- Worked on various phases of this project to provide the accurate and needed data.
- Discussed and learned about graph algorithm to find shortest path.
- Discussed and learned about machine learning in spark.