**A Machine learning deployed model to predict**

**Flight Fare using decision tree cost functions to improve test data accuracy**

**Project Members:**

Yeshraj Maganti-1144105

Subash Chandra Bose Kondru-1140025

Amarnath Pohanerkar-1140818

Sri Lakshmi Srinivas Gandham-1120106

**Subject: Cloud Computing Project**

**AWS Services:**

* **EC2 Instances**
* **Security Group**
* **Amazon Sage Maker**
* **S3 bucket**

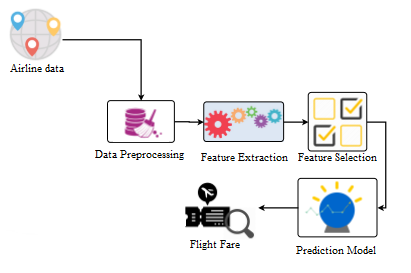
**Tools:**

* **Putty Generator**
* **Putty**
* **WinSCP**

**Abstract: -**

Flight transport system is playing vital role for every individual life when emergency takes place to travel from one place to another. The travelling of customers from source to destination at any time and they will come to know the chances of getting less prices of flight fare to plan their journey in a most economical where the charges are dynamic. So mainly the flight fare is the main challenging task to every customer will know about cost of the journey on priority basis. The machine learning model will suggest the best buy price of the flight ticket when to buy a ticket. The hyper parameters tuning will reduce the over-fitting of data and improves the accuracy on training data is with highest accuracy and on test data was improved as well on the prediction model. The data was collected from 16 routes among which 4 routes were used for analysis.

**Proposed System -**



**Airline Data**: The dataset with rows are 10683.

I) The Airline Feature has 12 distinct values

1. The Jet Airways are 4746 entries

2. The IndiGo is with 2564 entries

3. The Air India is with 2192

4. The Multiple Carriers are with 1543 entries

5. The SpiceJet with 1026 entries

6. The others with 7 entries

II) The Date of Journey Feature has 44 distinct values

1. The date 18.05.2019 has 633 entries

2. The date 06.06.2019 has 630 entries

3. The date 09.05.2019 has 628 entries

4. The date 12.06.2019 has 628 entries

5. The date 21.05.2019 has 615 entries

6. The other dates with 10220 entries

III) The Source Feature has 5 distinct values with categorical values

1. The Delhi with 5682 entries

2. The Kolkata with 3581 entries

3. The Bangalore with 2735 entries

4. The Mumbai with 883 entries

5. The Chennai with 456 entries

IV) The Destination Feature is a categorical with 6 distinct values

1. Cochin with 5682 entries

2. Bangalore with 3581 entries

3. Delhi with 1582 entries

4. New Delhi with 1170 entries

5. Hyderabad with 883 entries

V) The Route Feature is a categorical with 132 distinct values

1. The DEL->BOM->COK with 3000 entries

2. The BLR->DEL with 1941 entries

3. The CCU->BOM->BLR with 1232 entries

4. The CCU->BLR with 896 entries

5. The BOM->HYD with 785 entries

6. The Other values of 127 different routes with 5499 entries

VI) The Dep\_Time feature is categorical with 223 distinct values

1. The 18:52 with 288 values

2. The 17:00 with 275 values

3. The 10:00 with 265 values

4. The 07:05 with 240 entries

5. The 07:10 with 239 entries

6. The Different time values with 218 is 12047 entries

VII) The Arrival\_Time feature is categorical with 1451 distinct values

1. The 19:00 with 536 values

2. The 21:00 with 466 values

3. The 19:15 with 434 values

4. The 16:10 with 185 entries

5. The 12:35 with 159 entries

6. The Different time values with 1446 is 11574 entries

VIII) The Duration feature is categorical with 374 distinct values

1. The 2h 50m is 672 entries

2. The 1h 30m is 493 entries

3. The 2h 45m is 432 entries

4. The 2h 55m is 418 entries

5. The 2h 35m is 399 entries

6. The different duration values with 369 is 10940 entries

IX) The Total\_Stops feature is categorical with 5 distinct values

1. The nonstop with 4340 entries

2. The 1-stop with 7056 entries

3. The 2-stops with 1899 entries

4. The 3-stops with 56 entries

5. The 4-stops with 2 entries

X) The Additional\_info feature is categorical with 10 distinct values

1. The No info is 10493 entries

2. The In-Flight meal not included is 2426 entries

3. The No check-in baggage included is 396 entries

4. The 1 Long layover is 20 entries

5. The Change airports is 8 entries

6. The different duration values with 5 is 11 entries

XI) The price feature is a real number which is the dependent variable

1. The Distinct values are 1870 entries.

2. There are missing values.

**Security –**

**Security group**

**Public and private keys**

**Scalability –**

**You can add more data.**

**Hypervisor tuning provides more accuracy than decision tree.**