|  |
| --- |
| **Experiment No. 9**  **Title: Case study: Big data platform / analytics as business need)** |

**Batch:B1** **Roll No.:1824008** **Experiment No.:9**

**Title: Case study: Determining whether there will be flight delayed or not.**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Resources needed: Microsoft Azure Machine Learning Studio**

Azure Gallery link : <https://gallery.cortanaintelligence.com/Experiment/Experiment-created-on-10-11-2020-Predicting-Flight-Delays-1824008>

Github link: https://github.com/trunesh/MLClassficationFlightArrivalDelayby15min\_1824008

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Describe the following points with respect to the business under consideration,**

1. **Problem faced by the business**

Flight delay creates major problems in the current aviation system. Methods are needed to analyze the manner how delay propagates in the airport networks. Traditional methods are inadequate to the task.

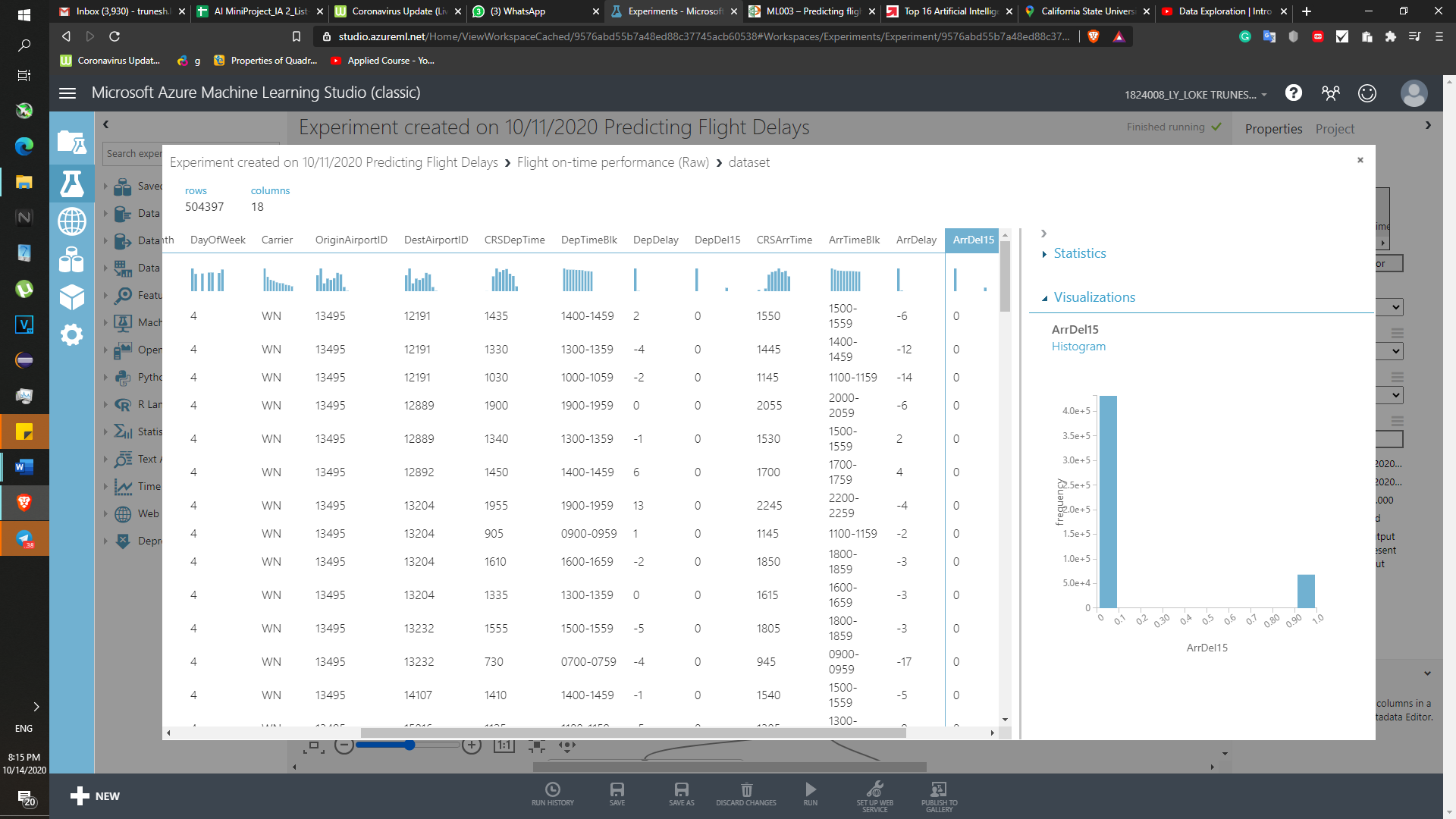
1. **Approach/ Methodology followed by the business**

A new machine learning based Arrival delay prediction model that uses the Two Class Decision tree classification and find approximated Flight Arrival delay.

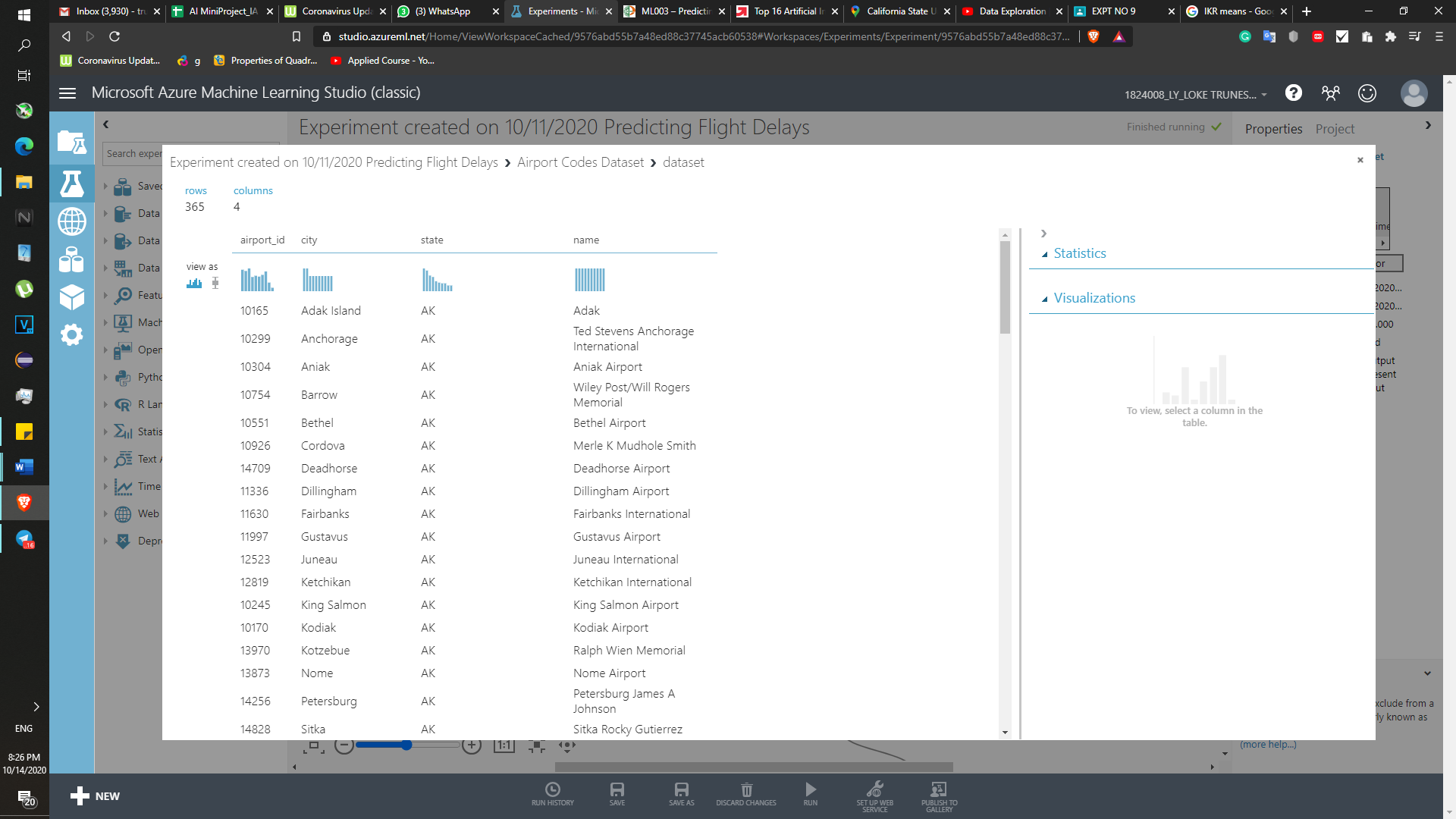
**Step 1: Selecting Databases**

To Calculate whether or not there will be Arrival delay on the destination point following two Datasets are used:

1. Flight on-time performance (Primary database )



1. Airport Codes Dataset

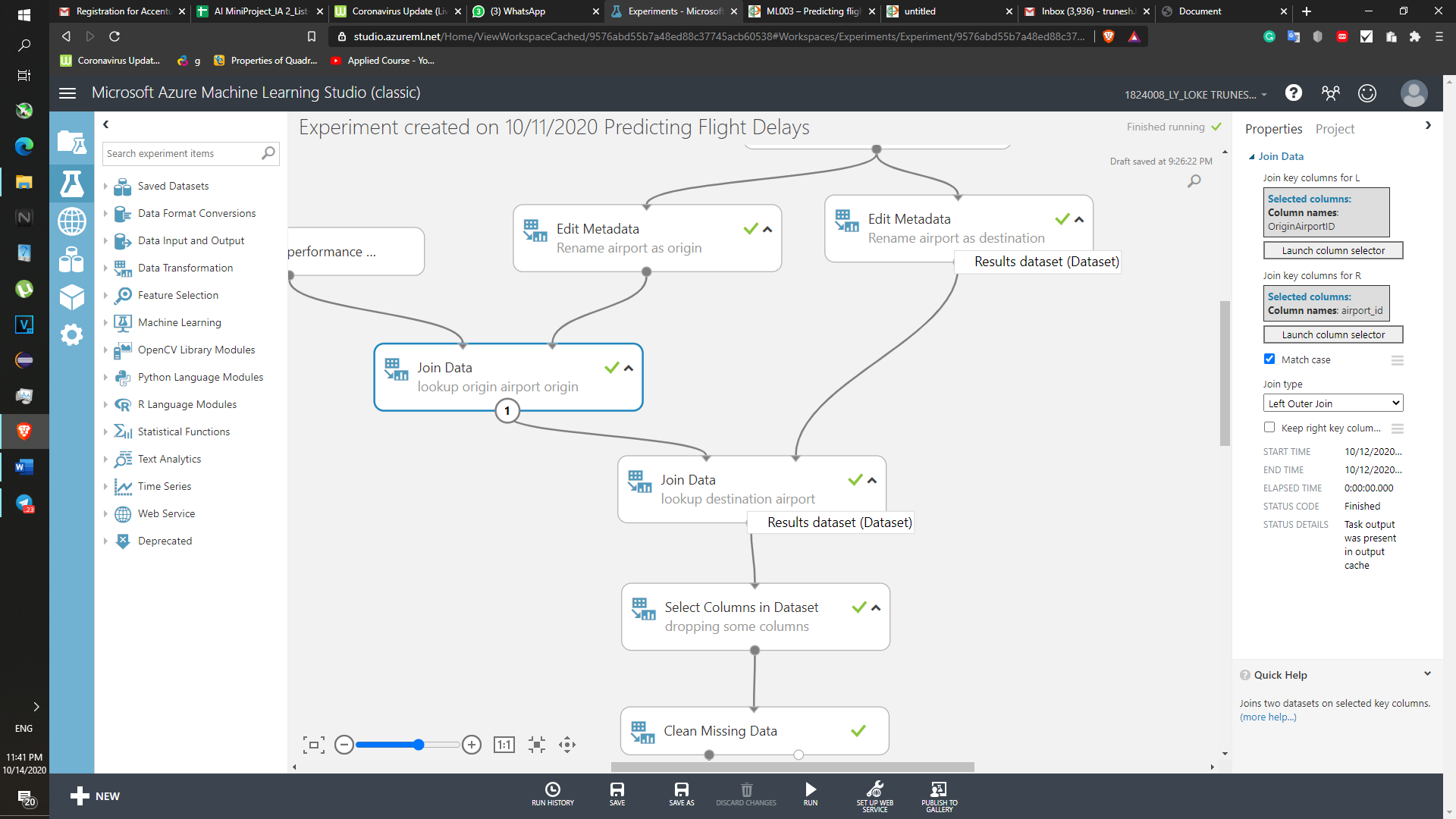


**Step 2 : Combining two databases to get the meaningful data.**

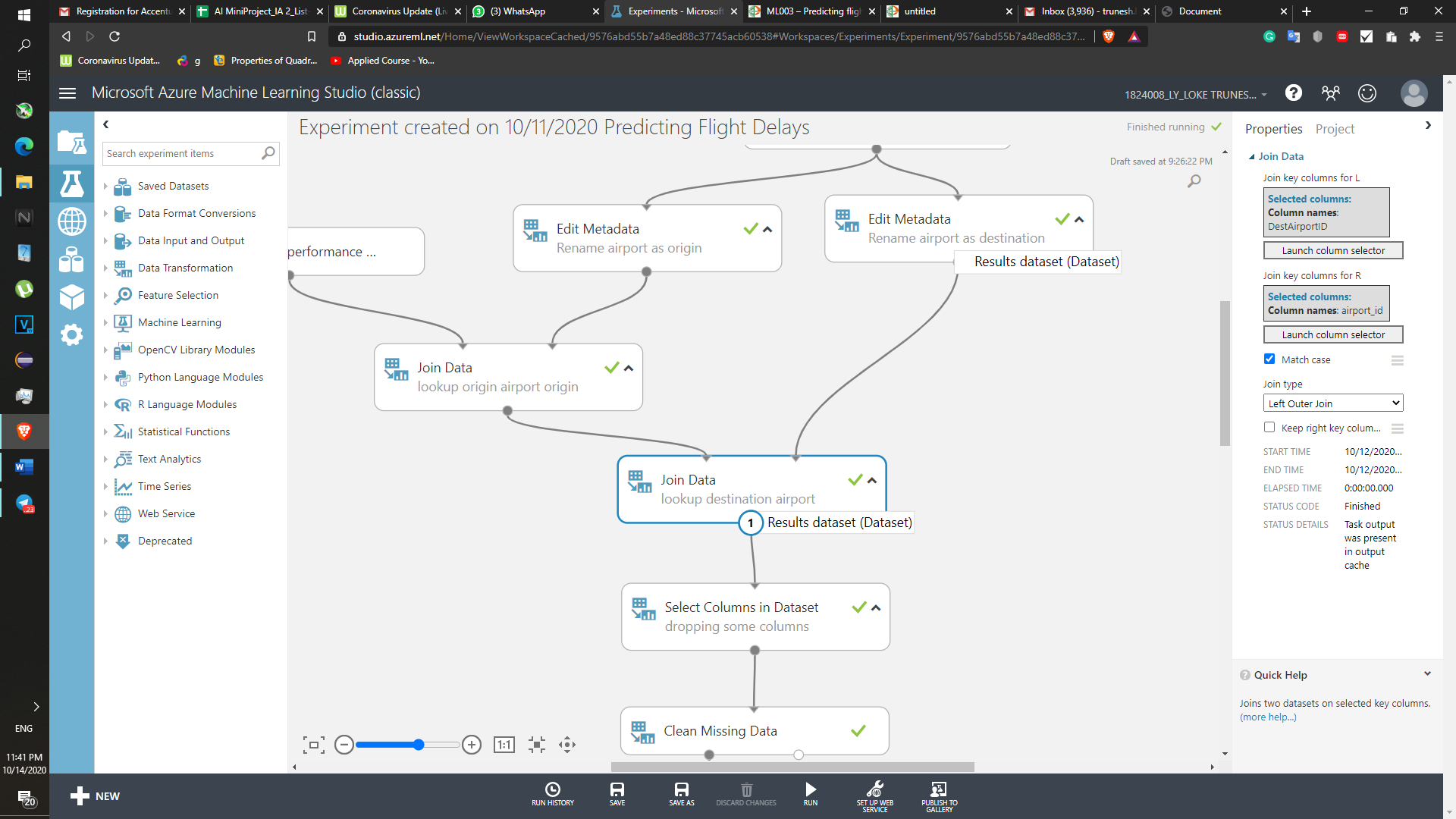
Airport Codes Dataset this dataset split into two having same columns but column name changes except Airport ID.

Later these two database which are split into two will be combined with the Flight on time performance using Left outer join.

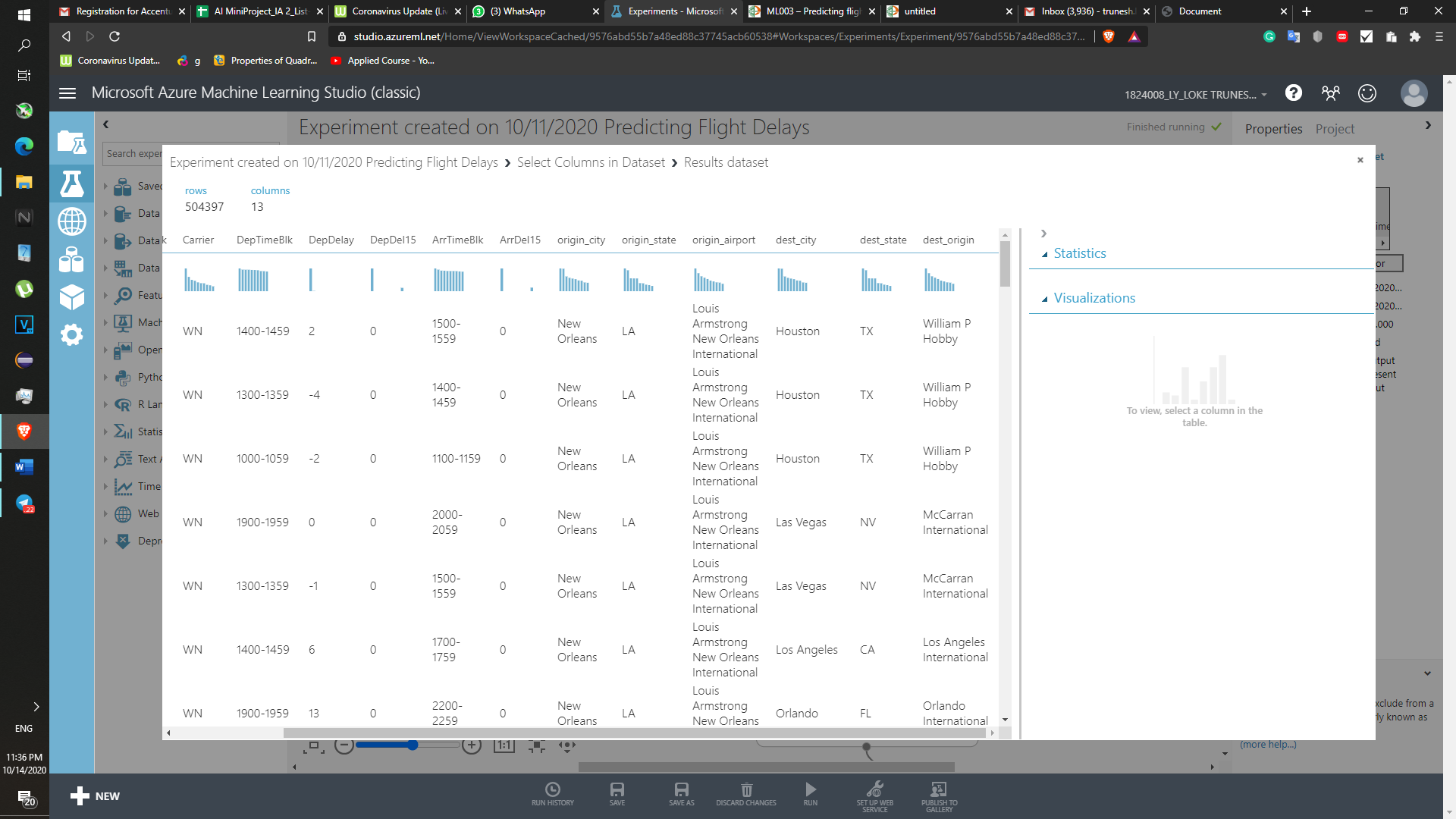
1. Flight on time performance is left joined with the Air Codes DataSet dataSet which columns renamed as the Origin City, Origin Port etc.



1. Again new Flight-on Time performance dataset is left join with Air Codes dataset which columns has been renamed as destination city,destination port etc.

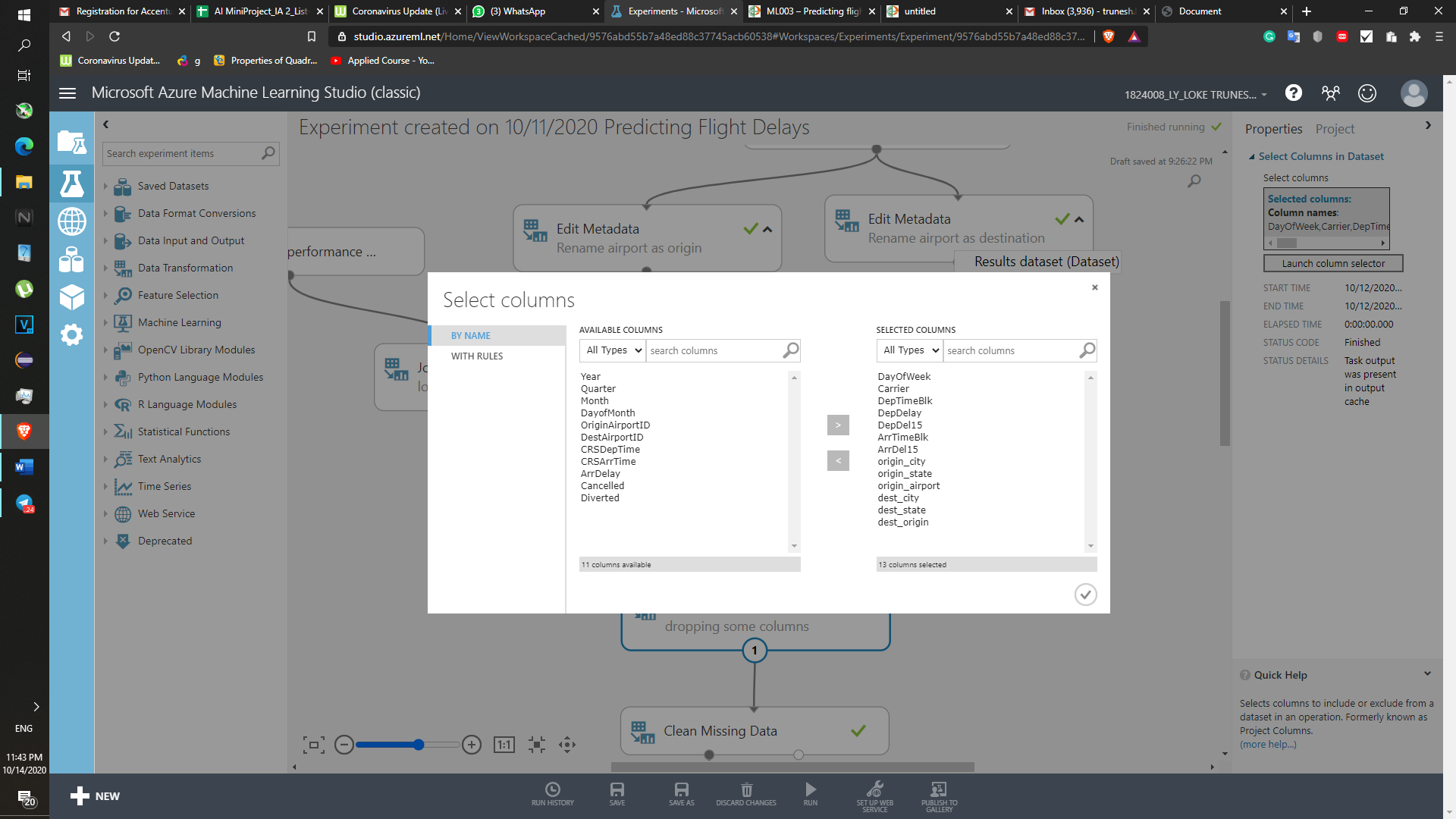


**Resultant database after combining**



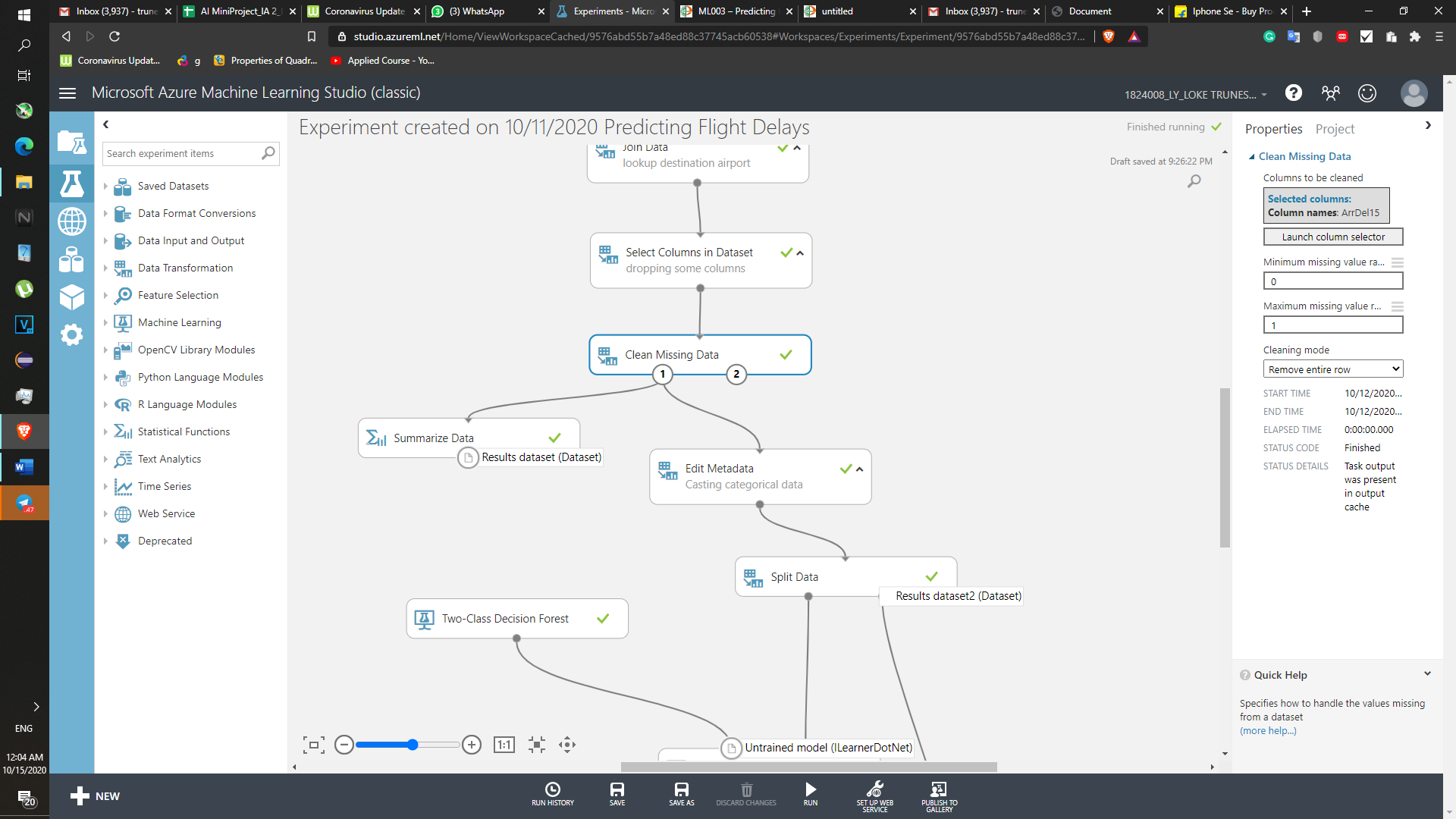
**Step 3: Selecting necessary Columns**

**‘**



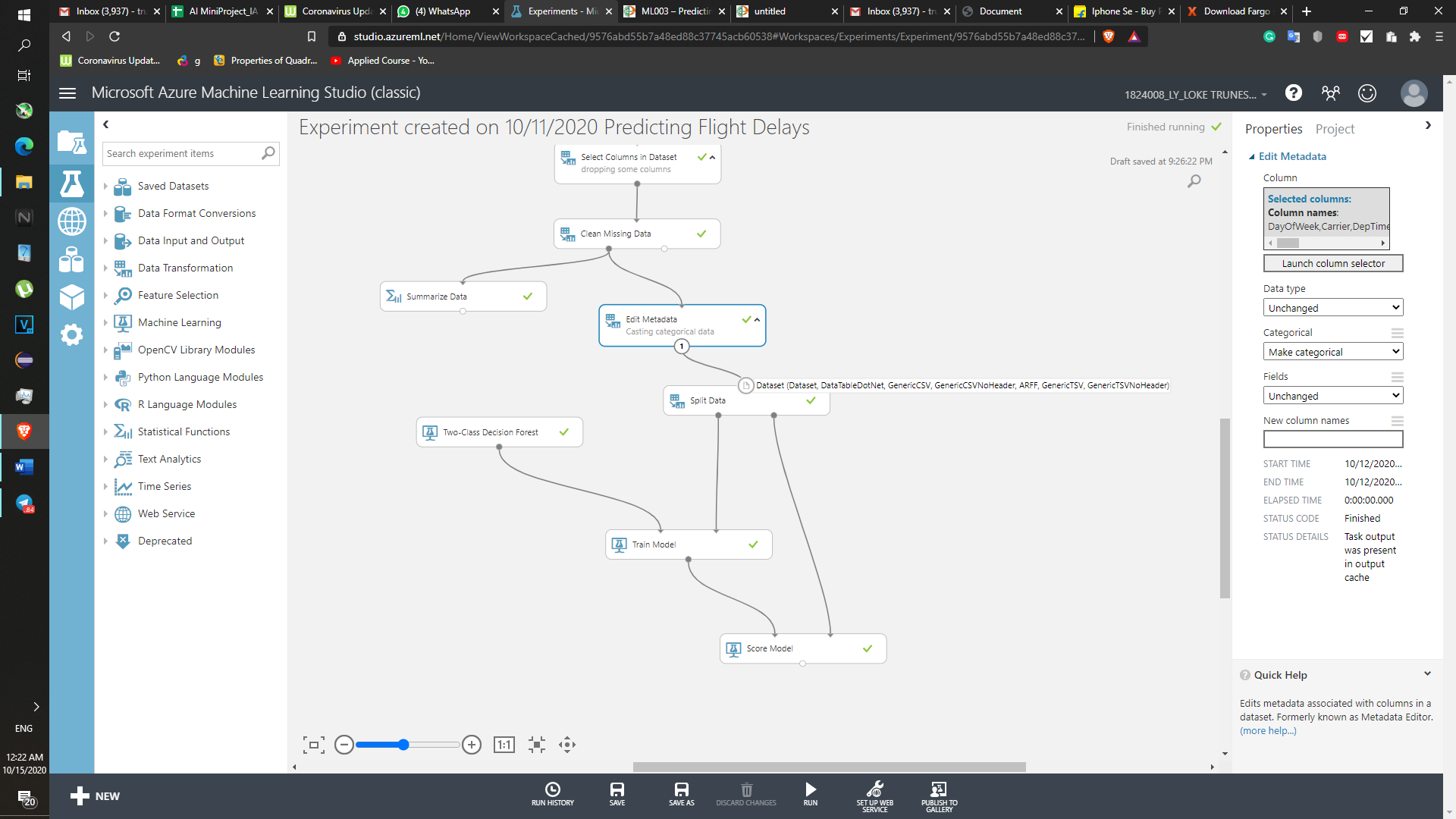
**Step 4: Cleaning Missing data**

Our task to find whether there will be arrival delay of the flight by 15 min or not, that’s why only eliminating the missing rows if column ‘ArrDel15’ column value is not present.

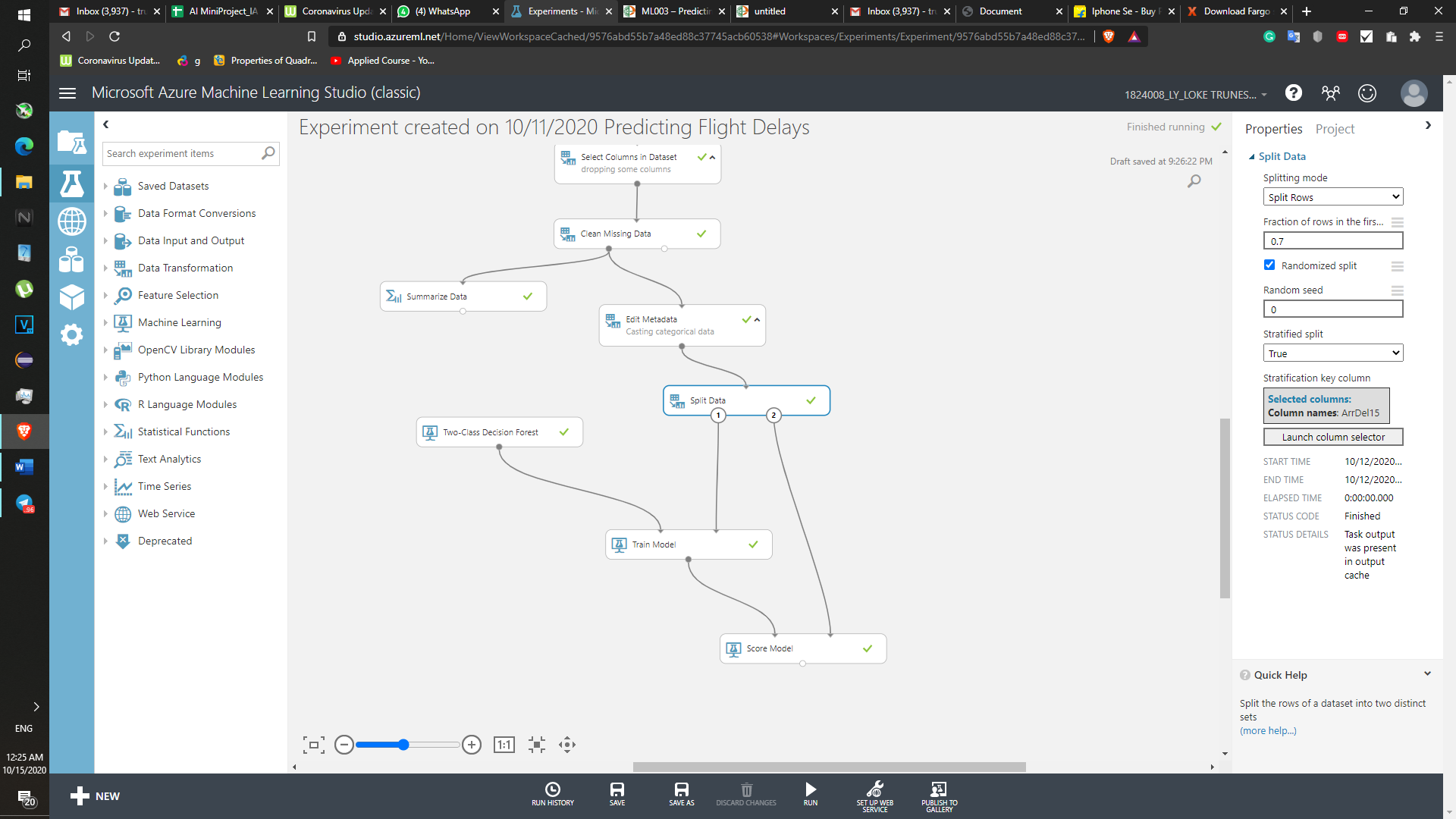


**Step 6:Casting to Categorical Data**

To make data suitable for machine learning model. Data has to be in categorical format. To make effective decision.



**Step 7: Split data 70% to train the model and 30 % to test the data . Stratified Split is set to true to divide the data in the equal ratio on the both trained dataset as well as the in the test data set.**



**Step 8: Selecting Machine learning Model**

**Current dataset is categorical , that means it’s a classification problem. We have to predict whether or not there will be arrival delay more than 15 minutes or not. So only two option we have to predict YES or NO, 0 or 1 . That’s why Two-Class Decision forest algorithm is selected to trained the model.**

**Two-class Decision Forest**

This decision forest algorithm is an ensemble learning method intended for classification tasks. Ensemble methods are based on the general principle that rather than relying on a single model, you can get better results and a more generalized model by creating multiple related models and combining them in some way. Generally, ensemble models provide better coverage and accuracy than single decision trees.

There are many ways to create individual models and combine them in an ensemble. This particular implementation of a decision forest works by building multiple decision trees and then voting on the most popular output class. Voting is one of the better-known methods for generating results in an ensemble model.

Many individual classification trees are created, using the entire dataset, but different (usually randomized) starting points. This differs from the random forest approach, in which the individual decision trees might only use some randomized portion of the data or features.

Each tree in the decision forest tree outputs a non-normalized frequency histogram of labels.

The aggregation process sums these histograms and normalizes the result to get the “probabilities” for each label.

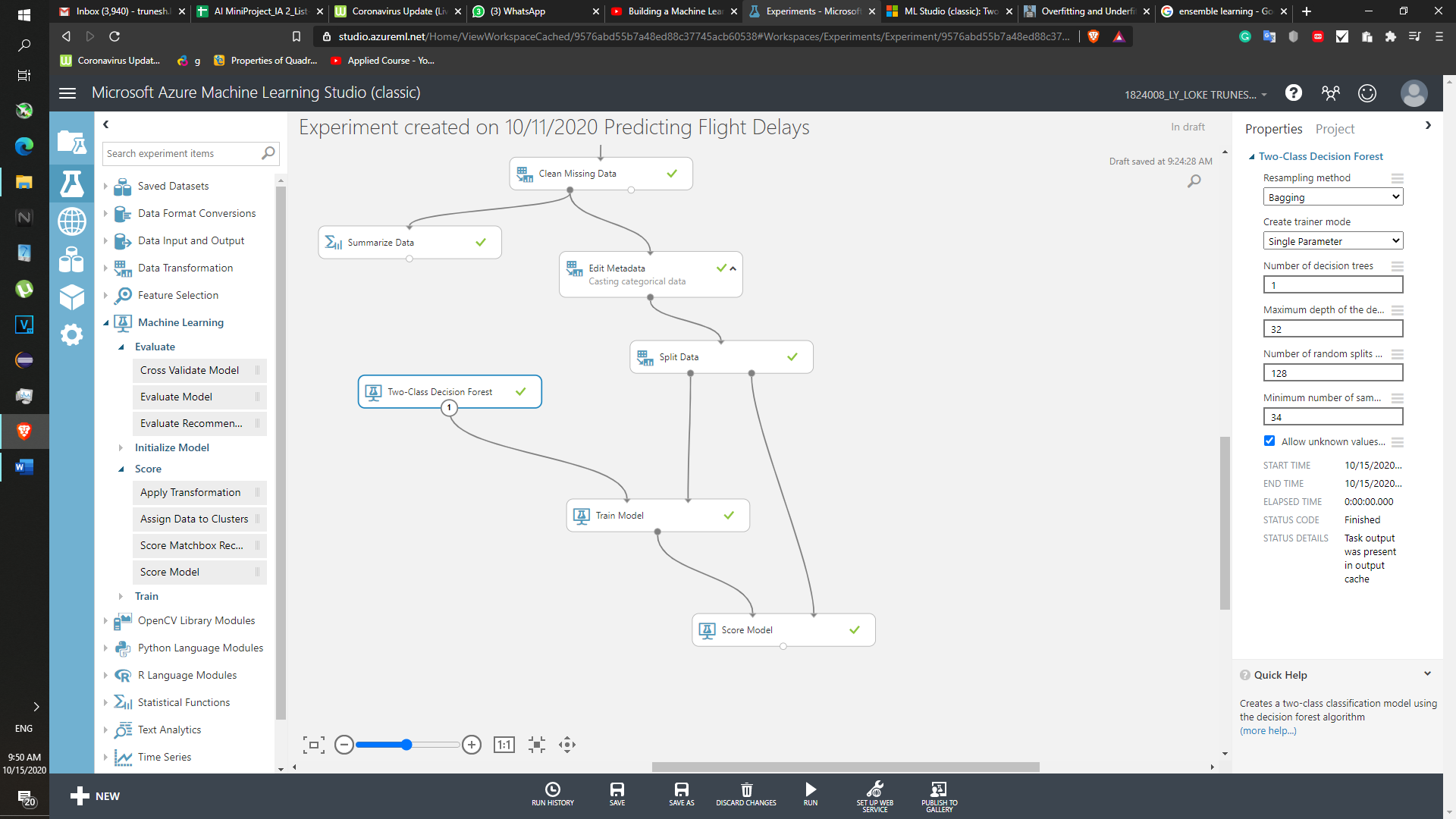
The trees that have high prediction confidence will have a greater weight in the final decision of the ensemble.

**Decision trees in general have many advantages for classification tasks:**

* They can capture non-linear decision boundaries.
* You can train and predict on lots of data, as they are efficient in computation and memory usage.
* Feature selection is integrated in the training and classification processes.
* Trees can acommodate noisy data and many features.
* They are non-parametric models, meaning they can handle data with varied distributions.

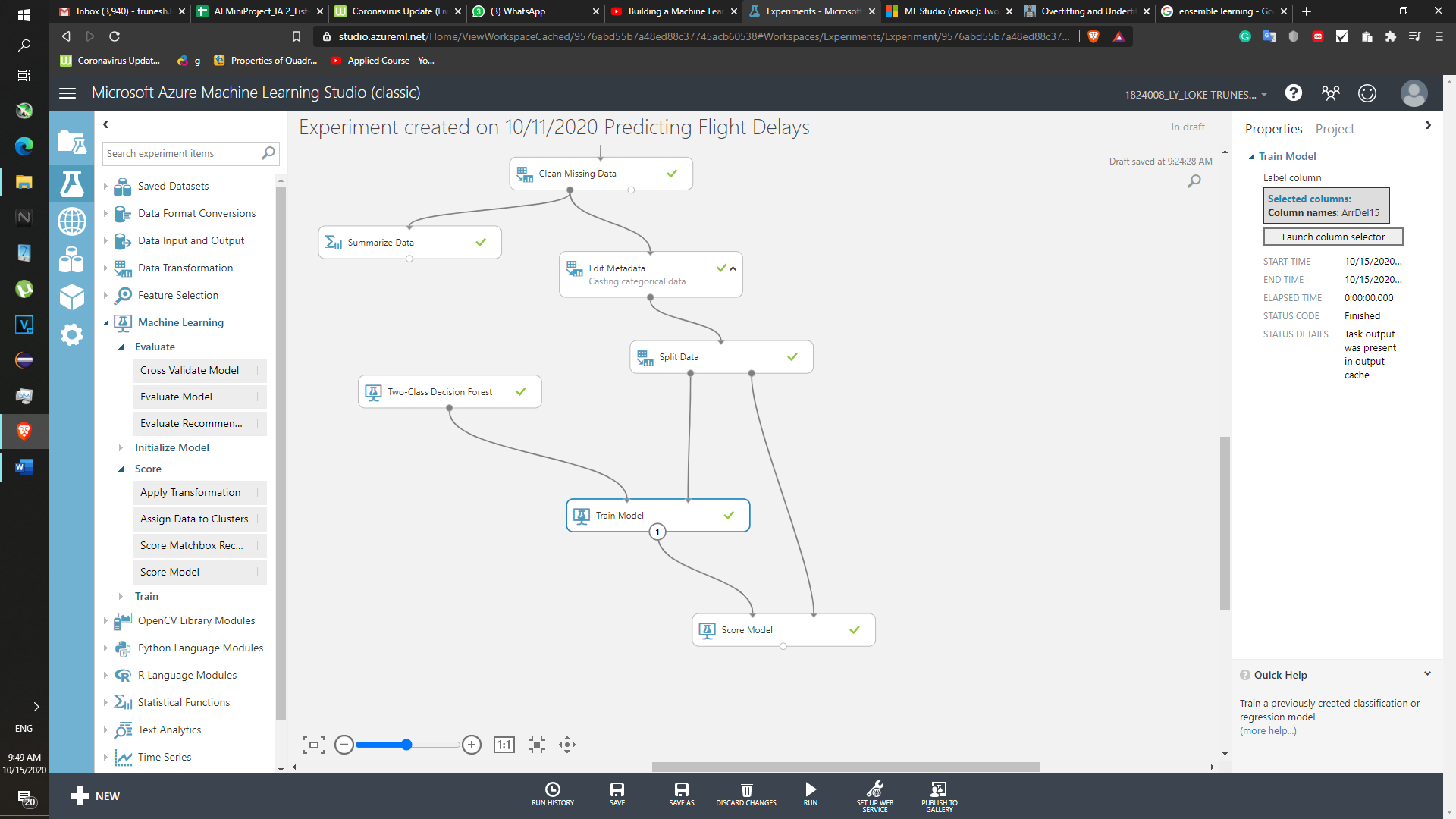
**Disadvantages:**

However, simple decision trees can overfit on data, and are less generalizable than tree ensembles.



**Step 9:  
To create a trained model we have to provide the data which selected to train the model (Train data set)**

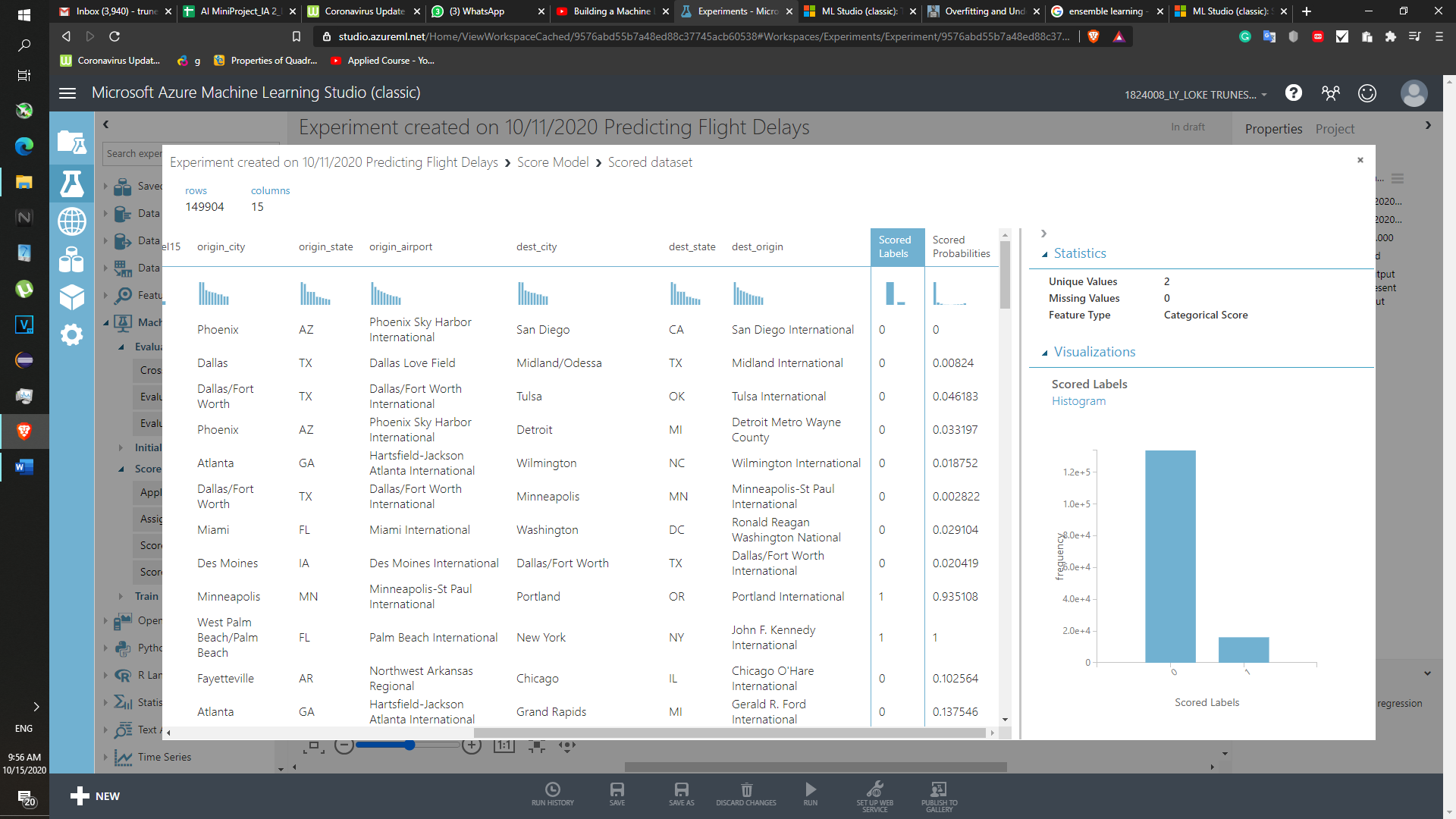
**Column selected to as ArrDel15 to obtain results.**



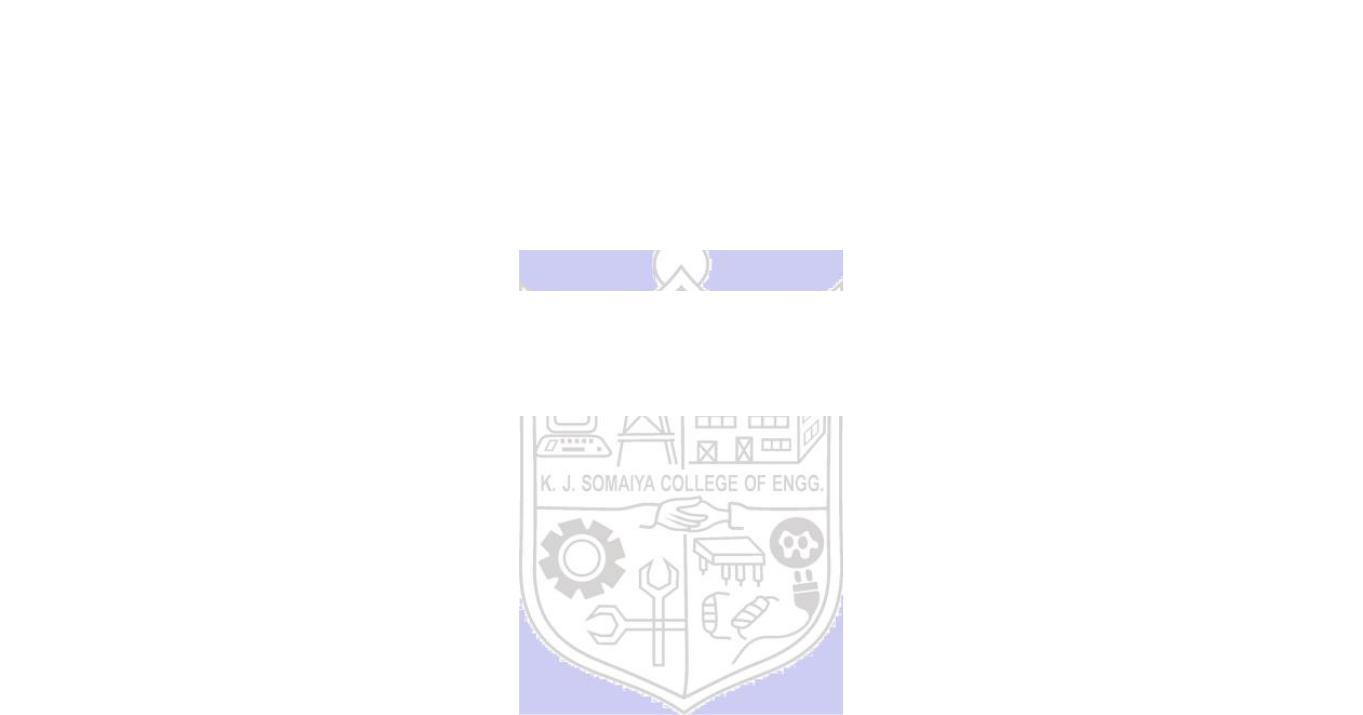
**Step 10: Final Step to Get result in Score model whether or not results are correct or not .**

**We pass the Train model data into Score model as well as the data to be tested.**

**Here we try to predict the Test data and compare with the actual results.**



As score labels are compared with the score probabilites one can say the model is almost working correctly. It ables to predict the for eg. From Monnepolis to Portland flight will be delayed by more than 15 min .



**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Questions:**

Discuss the tangible and intangible benefits the business has observed after the implementation.

**Budget Management:**

One of the major tangible benefits of project management is budgetary savings. Project managers control budgets and make decisions about how best to allocate resources in the process of working toward a project's objectives.

**Competative Advantage:**

Staying a step ahead of the competition is the key to success – especially in today’s global marketplace. Every company is chasing digital transformation and hoping to claim their spot at the head of the pack in their respective industry. The use of intelligent process automation can facilitate this transformation, not only be streamlining processes, but by empowering human workers.

**Decision Support:**

Because intelligent process automation is powered by machine learning, it is inherently capable of analyzing massive amounts of data and extracting value. Furthermore,ML can then turn that data into actionable insights that can be utilized by business leaders to make better decisions.

Incorporating advanced business automation technology into the mix enables the analysis of overall organizational performance. With these intelligent analytics, business leaders can more effectively identify and implement the right approaches to achieve improved performance over the long-term.

Via obtaining information whether or not there will be flight delay by 15 min or not . This can be used to inform the customers and relatives about the flight delay. Such decisions can help to develop trust between Business and Customers.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Outcomes:**

Realise adequate perspectives of big data analytics in various applications

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Conclusion: (Conclusion to be based on the objectives and outcomes achieved)**

Hence Successfully able to predict whether or not there will be Arrival delay of 15 min from one destination to another in the US Flights.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

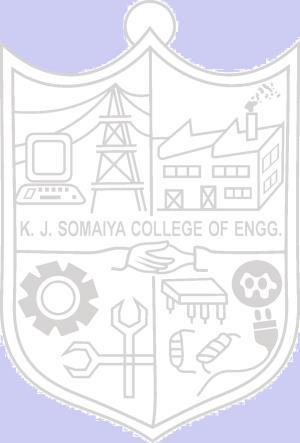


**Grade: AA / AB / BB / BC / CC / CD /DD**

**Signature of faculty in-charge with date**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**References:**



**Books/ Journals/ Websites:**