PROYECTO BIGDATA

CLAUDIA LOZANO PÉREZ

ÍNDICE:

1.INTRODUCCIÓN:

Nos muestran un escenario donde hemos conseguido nuestro primer trabajo de análisis de Big Data dentro de la empresa Tokio School Viajes, dónde debemos analizar una serie de datos sobre el tráfico en el aeropuerto de San Francisco, donde despegan muchos aviones con destino a Tokio y otras múltiples ciudades del país de Japón.

2.ESTUDIO DE DATOS:

Tenemos un Dataset donde recogen todos los datos para realizar el estudio de la compañía. En primer lugar, para realizar un buen estudio, debemos conocer cuantas columnas presenta el Dataset y que tipo de dato contienen cada una. Para ello he realizado la siguiente tabla:

NOMBRE	**TIPO**
Activity Period	Número
Operating Airline	String
Operating Airline IATA Code	String
Published Airline	String
Published Airline IATA Code	String
GEO Summary	String
GEO Region	String
Activity Type Code	String
Price Category Code	String
Terminal	String
Boarding Area	String
Passenger Count	Número
Adjusted Activity Type Code	String
Adjusted Passenger Count	Número
Year	Número
Month	String

Presentamos 16 columnas, de las cuales 12 contienen datos tipo “Cadena de caracteres” y 4 “Números”.

3.RESOLUCIÓN DE ALGUNAS CUESTIONES CON RELACIÓN AL DATAFRAME:

Una vez que conocemos la naturaleza de los datos y en que columna se encuentran podemos responder a las siguientes cuestiones que se nos plantean:

1. ¿Cuántas compañías diferentes presenta el fichero?

El fechero presenta un total de 77 compañías diferentes.

2. ¿Cuántos pasajeros tienen de media los vuelos de a compañía?

Para poder responder a esta pregunta he realizado un diccionario donde la clave es el nombre de la aerolínea y el valor es la media de pasajeros, donde he obtenido lo siguiente:

'ATA Airlines': 8744.636363636364

'Air Canada ': 4407.183673469388

'Air China': 5463.822222222222

'Air France': 2320.75

'Air New Zealand': 18251.560109289618

'AirTran Airways': 294.2142857142857

'Alaska Airlines': 6618.335907335907

'All Nippon Airways': 11589.077519379845

'American Airlines': 2834.5

'American Eagle Airlines': 7452.339768339768

'Asiana Airlines': 10569.238938053097

'Atlantic Southeast Airlines': 17251.637816245006

'BelAir Airlines': 6385.523255813953

'British Airways': 1516.8125

'Cathay Pacific': 127164.38970588235

'China Airlines': 4006.5283018867926

'Delta Air Lines': 5.0,

'EVA Airways': 5902.961240310077

'Frontier Airlines': 2176.909090909091

'Hawaiian Airlines': 34.0

'Horizon Air ': 415.3636363636364

'Icelandair': 18.0

'Independence Air': 17625.124031007752

'Japan Airlines': 3418.0714285714284

'KLM Royal Dutch Airlines': 17121.325581395347

'Korean Air Lines': 9857.51550387597

'Lufthansa German Airlines': 5498.402777777777

'Mesa Airlines': 4321.4375

'Mexicana Airlines': 23358.55681818182

'Midwest Airlines': 68498.49740932643

'Northwest Airlines': 13116.356589147286

'Philippine Airlines': 9070.866666666667

'Singapore Airlines': 6476.088235294118

'SkyWest Airlines': 2.0

'Sun Country Airlines': 5631.84375

'TACA': 17787.676923076924

'US Airways': 8282.186046511628

'United Airlines': 5577.583333333333

`United Airlines - Pre 07/01/2013': 2799.7

'Virgin Atlantic': 6391.3

'WestJet Airlines': 6470.332046332046

'Boeing Company': 4280.3125

'Miami Air International': 35261.13963963964

'Air Canada Jazz': 9221.813953488372

'Qantas Airways': 5678.461240310077

'Ameriflight': 2786.011111111111

'Spirit Airlines': 19301.96511627907

'Xtra Airways': 3710.5811965811968

'Evergreen International Airlines': 2864.7272727272725

'Aeromexico': 7993.806451612903

'JetBlue Airways ': 107.375

'ExpressJet Airlines': 3883.0

'Southwest Airlines': 26109.25

'Virgin America': 160.0

'Aer Lingus': 10248.635658914729

'Allegiant Air': 4991.2164179104475

'Jet Airways': 2452.5

'Emirates ': 5865.847222222223

'Mesaba Airlines': 90.05555555555556

'World Airways': 14746.647286821706

'Air Berlin': 37083.83904465213

'Republic Airlines': 81188.15857605178

'Servisair': 2921.0416666666665

'Pacific Aviation': 3992.652

'Swiss International': 6061.640287769784

'LAN Peru': 258.6

'Swissport USA': 5066.197674418605

'XL Airways France': 8162.416666666667

'China Eastern': 55317.81578947369

'SAS Airlines': 72732.05829596413

'Atlas Air, Inc': 48915.46750232126

'Compass Airlines': 74405.35359116022

'Etihad Airways': 9847.10465116279

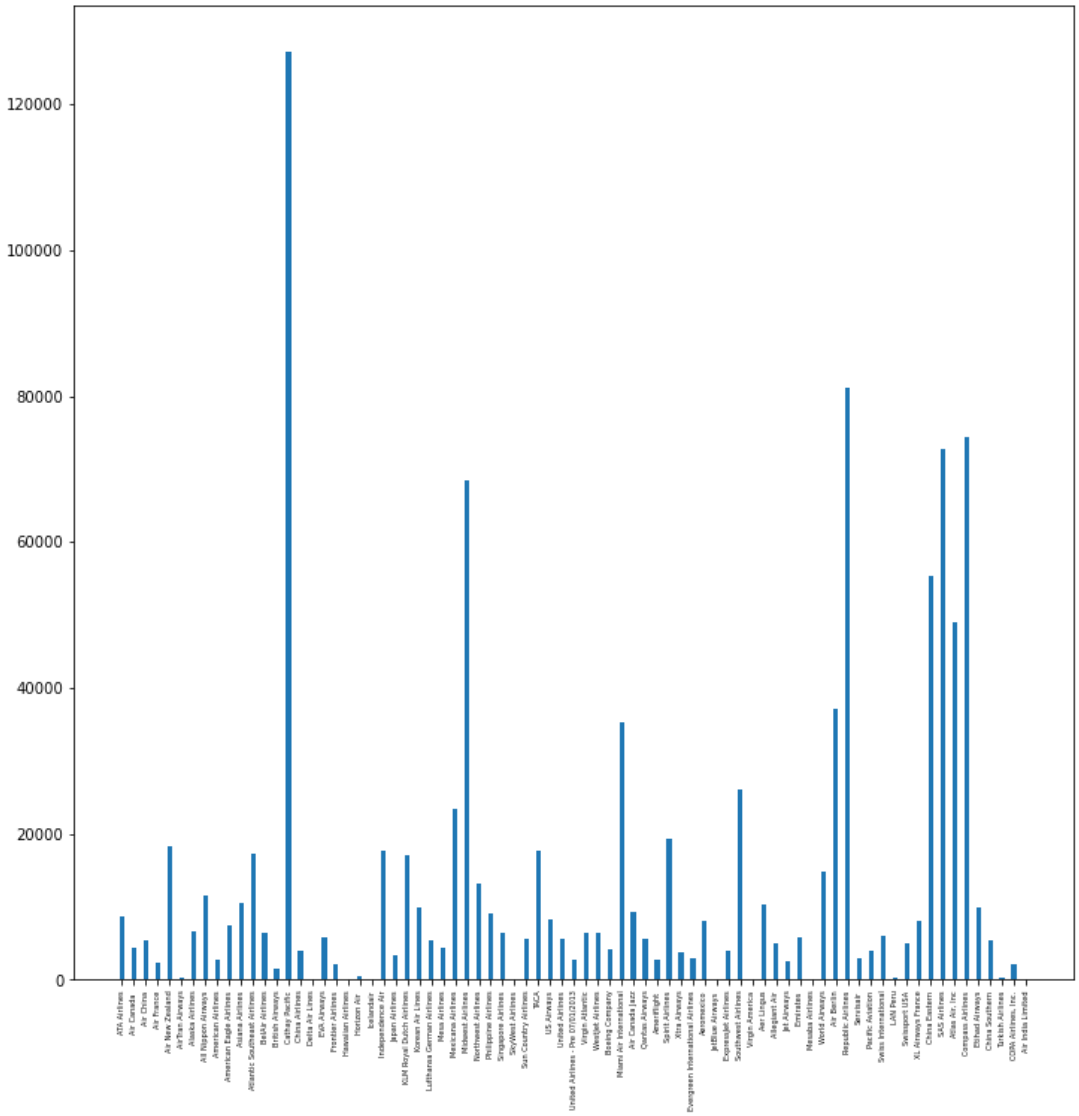
'China Southern': 5338.155339805825

'Turkish Airlines': 261.6666666666667

'COPA Airlines, Inc.': 2223.1612903225805

'Air India Limited': 73.0

Todas estas predicciones se recogen en el siguiente gráfico de barras:



3. Eliminación de los valores duplicados por el campo “GEO Región”, manteniendo el valor más alto con relación al número de pasajeros.

La columna GEO Región presenta datos duplicados por lo que he eliminados los duplicados y he mantenido el valor más alto.

4.Guardar los datos en archivos CSV.

Una vez analizado lo anterior he recopilado toda la información en dos archivos csv. Uno donde encontramos la media de los pasajeros por aerolínea que lo he llamado “aerolíneas_totales” y en el otro encontramos el dataframe sin los datos duplicados de GEO Región que lo he llamado “ aerolíneas_región”.

4.ANÁLISIS DESCRIPTIVO DEL DATAFRAME:

Aquí he realizado la media y la desviación típica del resto de las columnas que presentan datos del tipo “Int”.

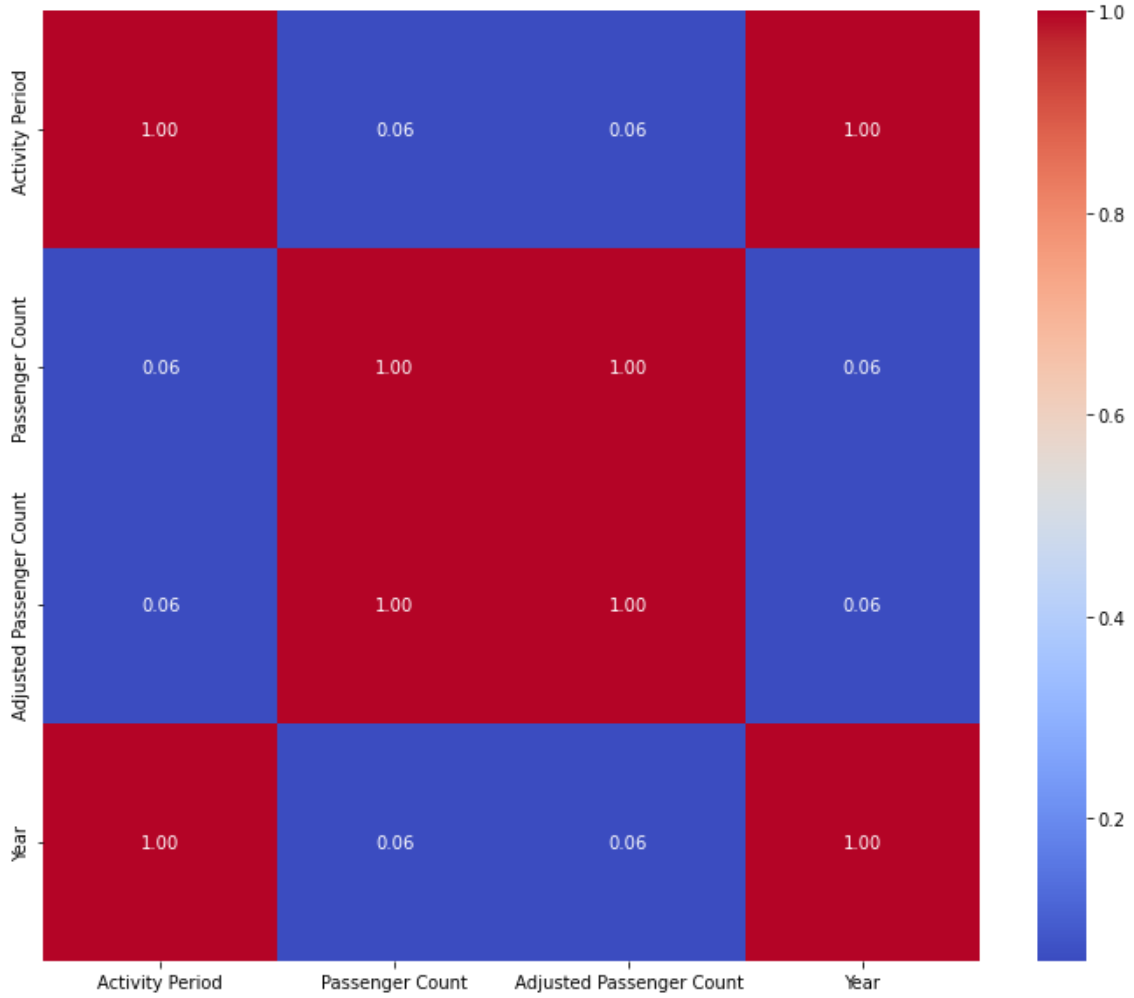
Las medias obtenidas son:

- Activity Period: 201045.07336576266
- Adjusted Passenger Count: 29331.917105350836
- Year: 2010.385220230559

Las desviaciones típicas obtenidas son:

- Activity Period: 313.33619609971413
- Adjusted Passenger Count: 58284.18221866232
- Year: 3.1375890431679667

Luego he realizado una matriz de correlación entre las cuatro columnas para ver cuanto estaban relacionadas entre ellas. He obtenido lo siguiente:

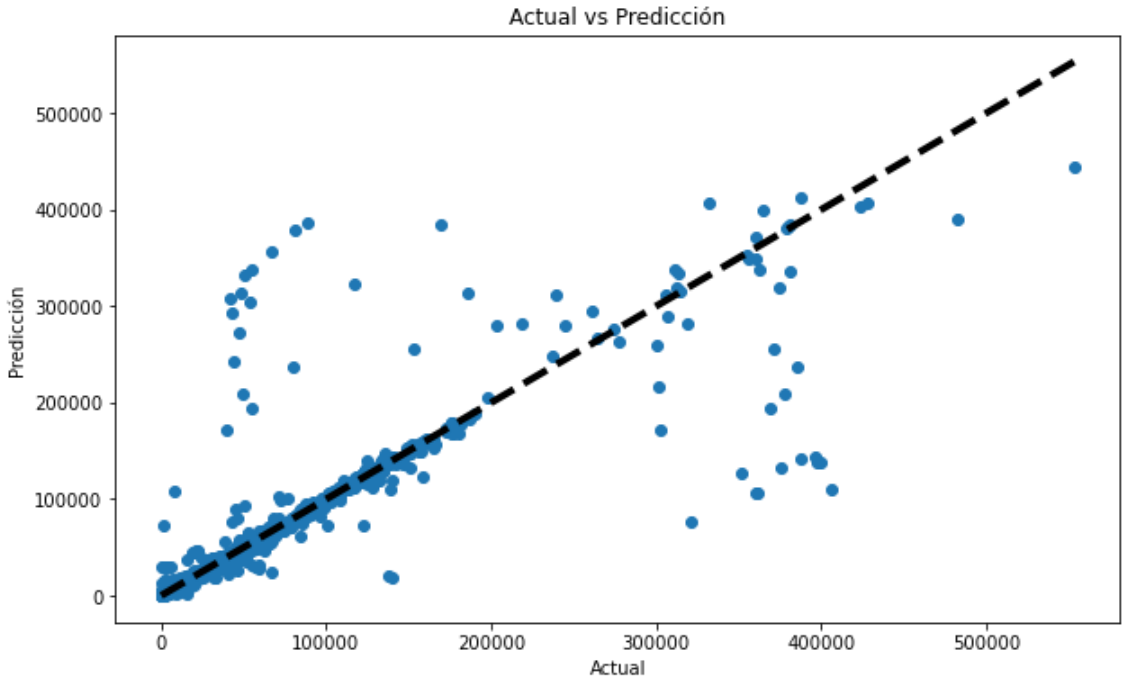


Esta matriz va de 0 hasta 1. Podemos observar que Passenger Count y Adjusted Passenger Count tienen una relación de 1.00, cómo Year y Activity Period. En cambio Passenger Count y Adjusted Passenger Count cuando se relacionan con Year o Activity Period, muestran muy poca relación, tan solo de un 0.06.

Por último, he limpiado el Dataframe inicial, eliminando los valores nulos. Luego he seleccionado las columnas con las que quiero quedarme, para estudiar el número de pasajeros, que han sido 'Operating Airline IATA Code', 'GEO Summary', 'GEO Region', 'Activity Type Code', 'Terminal' y 'Boarding Area', 'Month'

Una vez que tego el Dataframe como yo quiero divido los datos en variables dependientes e independientes, es decir los valores que voy a predecir y los que voy a usar para predecir.

Finalmente he calculado el R2 obteniendo una cifra de 0.7827318301600451.



En conclusión obtenemos que para valores de pasajeros más bajos el modelo funciona mejor llegando a ser muy bueno, sin embargo para valores más alejados de los cuales disponemos menos datos falla nuestro modelo. Esto último podría corregirse con más datos de los valores más altos.

Si tuvieras que escalar este proyecto a un conjunto de datos más grande que no cabría en la memoria de una sola máquina, ¿cómo distribuirías los datos y el trabajo entre diferentes máquinas usando Dask? Describe la estrategia que usarías y por qué crees que sería efectiva.

Si presentamos un conjunto lo suficientemente grande de datos para caber en una memoria de una sola máquina, tenemos la posibilidad de distribuir los datos y el trabajo entre diferentes máquinas utilizando Dask de una manera efectiva. La estrategia que emplearemos de basa en la partición y distribución de datos en un clúster de máquinas. Seguiremos los siguientes pasos:

1. Configuramos un cluster con Dask que esté distribuido, utilizando múltiples máquinas a la vez.
2. Dividimos los datos en particiones más pequeñas qque puedan caber en la memoria de ce cada máquina. Dask manejará la distribución de estas particiones en el clúster.

3. Cargar los datos distribuidos desde el almacenamiento del disco en cada máquina del clúster, proporcionando Dask estructuras de datos distribuidas, como el DataFrame distribuido, permitiendo realizar operaciones de una manera distribuida.
4. Ejecutamos las operaciones paralelas distribuidas con los datos. Un ejemplo de esto sería eliminar datos duplicados.
5. Recopilar todos los resultados y consolidarlos en una única solución.

Esta estrategia sería muy efectiva ya que permite aprovechar la capacidad de procesamiento y memoria distribuida de múltiples máquinas para manejar conjuntos de datos muy grandes.

Cuando calculaste la matriz de correlación y aplicaste el algoritmo de tu elección, ¿cómo se benefició tu análisis de la programación paralela y distribuida? Explica cómo la paralelización y la distribución del trabajo mejoraron el rendimiento de estos cálculos y cualquier desafío que hayas encontrado en el camino.

Como ya he mencionado antes, Dask nos permite ejecutar cantidades masivas de datos en diferentes máquinas, por lo que al crear una matriz de correlación de manera paralela y distribuida, es mucho más eficaz, se están haciendo simultáneamente varias matrices, para que como he mencionado previamente, se llegue a una única solución.

La matriz de correlación la he calculado antes, mediante Randon Forest, y tenemos un gráfico mostrándola.