

BIG DATA

Victor Di Salvatore

PROYECTO FINAL

(tu foto)



Victor Di Salvatore

RESPONSABLE DEL DESARROLLO DE LA APLICACIÓN

**Fecha:** 07, Dic, 2024

**DESARROLLO**

**Introducción.**

En este informe, se presenta un análisis detallado del tráfico aéreo utilizando una base de datos denominada "Trafico\_aereo.csv". El objetivo principal del estudio es desarrollar y evaluar modelos predictivos para estimar la cantidad de pasajeros a partir de variables como aerolíneas, meses, años, regiones y cantidad de pasajeros. Para llevar a cabo este análisis, se emplearon diversas técnicas estadísticas y de machine learning, tales como la correlación de Pearson, ANOVA (Análisis de Varianza) y modelos predictivos basados en Árboles de Decisión y Random Forest.

El procesamiento de datos fue realizado íntegramente con la librería PySpark de Python, asegurando la transformación, limpieza y preparación de los datos para su posterior análisis. En este proceso, se ajustaron los tipos de datos y se crearon variables categóricas para mejorar la compatibilidad con los modelos estadísticos seleccionados. Además, se exploraron las medias y desviaciones estándar de diversas variables para entender mejor la distribución de los datos y detectar posibles disparidades significativas, como las observadas entre regiones y aerolíneas.

Uno de los aspectos más destacados del análisis fue la identificación de patrones relacionados con la estacionalidad de los meses y la disparidad entre aerolíneas grandes y pequeñas. Este informe también aborda los desafíos y limitaciones que surgieron al predecir la cantidad exacta de pasajeros, proponiendo una solución mediante la categorización de los datos en cuatro grupos. Al final, se evaluó la precisión de los modelos predictivos, mostrando que Random Forest superó al Árbol de Decisión en términos de efectividad predictiva.

En resumen, este informe busca proporcionar una comprensión profunda de cómo diversas variables influyen en el tráfico aéreo y cómo se pueden emplear técnicas estadísticas avanzadas para predecir la cantidad de pasajeros, destacando las diferencias entre las aerolíneas y la estacionalidad del tráfico aéreo.

**Metodología y análisis de los resultados.**

Procesamiento de datos.

La preparación, transformación y pulido de los datos se hicieron en su totalidad con la librería PySpark de Python, se crearon variables categóricas y se ajustaron los tipos de datos para asegurar la compatibilidad con los modelos estadísticos.

Modelos y técnicas usadas:

1. Media y Desviación Típica.
2. ANOVA
3. Árbol de Decisión y Random Forest.

Utilizando una base de datos de tráfico aéreo llamada “Trafico\_aereo.csv” hemos realizado un análisis en base a datos relevantes como Aerolíneas, meses, años, regiones y cantidad de pasajeros.

En cuanto a la manipulación de esta, nos hemos visto en la obligación de convertir a entero algunas de las variables antes mencionadas para su posterior uso en modelos predictivos bien sea árbol de decisión o Random Forrest, ya que utilizamos ambos para observar la eficiencia de cada uno a la hora de predecir cantidad de pasajeros. Ahora bien, me he tomado la libertad de hacer esto de diferentes formas la primera extrayendo la idea de conocimientos de R studio donde le asignamos etiquetas y un nuevo nombre de variable a una ya existente, como se puede observar en la siguiente imagen.

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza baja

En cuanto al otro método que lo mencionaremos más adelante, estaríamos hablando del indexado que no es más que crear índices a cada uno de datos que tiene una variable.

Utilizando el método anterior de etiquetado para cada mes del año, nos dimos a la tarea de graficar separando las regiones en hemisferios (Norte, Sur y Centro) esto para facilitar la tarea de graficar y analizar las temporadas altas de cada una de ellas y para una posterior predicción que nos aumentara la taza de efectividad.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

En resumen, en las primeras dos graficas y debido a las estaciones climáticas de las mismas podemos ver picos pronunciados en la demanda, en la primera a mitad de año entre junio y septiembre y en la segunda entre noviembre y febrero, con respecto a la tercera vemos un cambio radical ya que no hay presencia de cambios de climas, pero si podemos ver un claro pico a la baja en la época de septiembre que concuerda con el inicio del año escolar muy común en países cercano al Ecuador, mas que todo en países de América central y algunos del sur.

Con respecto a otro de los puntos obligatorios pero necesarios para ir observando la dificultad que tendrá la predicción que llevaremos a cabo más adelante, tenemos las medias y las desviaciones estándar de diferentes variables con respecto a la cantidad de pasajeros, en resumidas cuentas al analizar cada una de ellas podemos ver cómo tanto las medias como las desviaciones son muy diferentes entre cada uno de datos dentro de una variable, podemos tomar el ejemplo más claro que es el de las regiones.

Texto

Descripción generada automáticamente

Texto, Tabla

Descripción generada automáticamente con confianza media

Si bien la diferencia entre regiones con respecto a la media era de esperarse por razones económicas, sociales, incluso poblaciones, etc. Nos llama la atención como la desviación estándar de cada una de las regiones con respecto a la cantidad de pasajeros es tan alta y mas aun en un lugar como Estados Unidos, esto podría deberse a que en los Estados Unidos por ejemplo hay una gran disparidad en el número de pasajeros entre estados y ciudades grandes (por ejemplo, Nueva York, Los Ángeles) versus zonas más pequeñas o rurales.

También podríamos hablar de la desviación estándar de Asia, pero la disparidad económica y social de ese continente hace que por ejemplo países como China, Japón, Corea del Sur, etc. Con respecto a los demás la distorsionen, pero claro está, no tanto como Estados Unidos.

Otro dato que nos llamó la atención fue el gran parecido en la media de la cantidad de pasajeros con respecto al mes y la altísima desviación estándar.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Texto

Descripción generada automáticamente

Ambas pueden ser atribuidas a las estaciones del año en los diferentes hemisferios, ya que, mientras en el hemisferio norte es verano en el sur es invierno y viceversa, llevándonos a datos tan dispersos como los que podemos ver en las imágenes anteriores.

Otro dato interesante es la desviación de las aerolíneas con respecto a la cantidad de pasajeros

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza baja

Tabla

Descripción generada automáticamente con confianza media

Podemos observar específicamente las aerolíneas que forman parte del mercado de Estados Unidos, la diferencia tan grande que hay tanto en desviación como en media de aerolíneas que atienden a mercados regionales con respecto a aerolíneas del mismo país que atienden a mercados tanto nacionales como internacionales, debiéndose estas al aprovechamiento de las grandes de la industria de factores como el turismo nacional e internacional, clave para el crecimiento en el área.

Ahora bien, luego del análisis de las medias y las desviaciones estándar de las diferentes variables que hemos considerado que tienen una importancia significativa en el modelo predictivo que realizaremos más adelante, hemos decidido realizar un análisis de varianza (ANOVA) que no es más que un estudio que se le hace a una variable con diferentes grupos de estudio con respecto a otra, para así poder decidir cuáles serán las que usaremos en nuestro en nuestro árbol de decisión o en nuestro Random Forrest.

Hemos decidido tomar las variables GEO\_Region, Year, Operating\_Airline y mes, debido a su efecto significativo sobre la cantidad de pasajeros, ya que en todos el análisis de estos nos hacen rechazar la hipótesis nula debido a la diferencia significativa que hay entre cada una de esas variables y la variable dependiente.

Algunos ejemplos como en análisis de las 9 regiones con respecto a la cantidad de pasajeros:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente con confianza media

En este ejemplo en especifico he tenido un pequeño problema y no he sabido como encontrar los valores de P ya que solo mostraba 0.0 y quería ver los valores exactos para ver qué tan lejos estaba de la zona de aceptación, así que me he tomado la libertad de extraer el resultado usando R Studio, que a mi parecer en temas estadísticos es más poderoso.

Captura de pantalla de un celular

Descripción generada automáticamente

Concluyendo que GEO\_Region influye significativamente en la variable dependiente, con un nivel de significancia extremadamente alto.

Colocar capturas de pantalla de los resultados de las variables antes mencionadas sería muy redundante ya que en todas ellas los resultados son aproximadamente los mismos, el rechazo claro de la hipótesis nula con una lejanía considerable de la zona de aceptación.

Visto lo anterior, y teniendo claro cuáles son las variables que deberíamos usar para poder realizar con eficiencia y eficacia nuestro modelo de predicción, aún tenemos un problema, que es la dificultad que tendría para predecir con exactitud la cantidad exacta de pasajeros que tendrá una aerolínea, en un mes, año y región determinados, por ende, nos hemos visto en la obligación de separar las cantidades de pasajeros en 4 categorías

Interfaz de usuario gráfica

Descripción generada automáticamente

Esto con dos fines el primero poder graficar para poder observar el comportamiento de los pasajeros de cada tipo de aerolínea que hemos llamados “Baja”, “Media”, y “Alta” con otras etiquetas con respecto a los meses del año concluyendo así, primero que nada la gran diferencia de pasajeros que hay entre las grandes aerolíneas las medianas y las pequeñas, ya que, las primeras mencionadas son las únicas que aprovechan las temporadas altas tanto en un hemisferio como en el otro, podemos observar picos tanto en el verano del norte como en el sur pero solo en las aerolíneas llamadas “Altas” en relación con su cantidad de pasajeros.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

En cuanto a la segunda utilidad de dichas etiquetas, el objetivo de esta es, facilitar el trabajo al modelo predictivo arrojándonos así en el caso del árbol de decisión un 75.26% de efectividad y en el caso del Random Forrest un 78.95% a la hora de predecir el aproximado de la variable dependiente con respecto a las diferentes variables que hemos venido mencionando.

Árbol de Decisión.

Tabla

Descripción generada automáticamente



Random Forrest.

Tabla

Descripción generada automáticamente



Con respecto a las variables de tipo cadena de texto, habíamos mencionado anteriormente como podíamos convertirlas por ejemplo a entero en el caso de los meses con etiquetas del 1 al 12, ahora bien, en el caso de las demás variables que utilizamos para entrenar los modelos hemos utilizado el método de indexación, obligatoriamente necesario para casos como el de las aerolíneas ya que hay una buena cantidad de ellas en la base de datos.

Interfaz de usuario gráfica, Texto, Aplicación, Chat o mensaje de texto

Descripción generada automáticamente

En un ultimo intento de aumentar la efectividad nos dimos a la tarea de filtrar por hemisferio, tomamos todas las regiones y las aglomeramos a cada una en su zona del planeta.

Texto

Descripción generada automáticamente

Esto para intentar facilitar aún mas la predicción de la cantidad de pasajeros, que como dijimos anteriormente las categorizamos del 1 al 4, dándonos cuenta de que por más que intentemos facilitar las cosas para nuestro modelo con las variables usadas es muy difícil pasar del 80% llegando a la conclusión que existe un 20% de variabilidad que es explicado con otras variables ajenas a las usadas, a continuación, presentaremos el ejemplo del caso del hemisferio norte:

Tabla

Descripción generada automáticamente



**CONCLUSIONES.**

Si bien tenemos una base de datos que nos puede dar datos muy relevantes y claros a la hora de entender cómo funciona la variabilidad de la cantidad de pasajeros y como está sujeta a factores como el mes del año y las diferentes estaciones, a la región del planeta y su hemisferio e incluso a la aerolínea, no dejan de ser muy generales, lo cual dificulta el objetivo fundamental de este estudio que era la predicción exacta de dicha variable que depende de una gran cantidad de factores incluso muchos que no están presentes, por ende, nos hemos visto obligados a categorizar las cantidades de pasajeros para hacerle más fácil al modelo predictivo su tarea logrando así un 78% en el caso más alto de efectividad.

Analizando en profundidad creemos que el objetivo del estudio no puede estar basado en números exactos de cantidad de pasajeros ya que es innecesario, sino más bien en poder predecir si la demanda de pasajeros en una determinada estación o mes del año, en una región o un año será alta, media o baja, para poder hacer los ajustes de precios, promociones y estrategias de mercadeo correspondientes.