UNIVERSIDAD DE GUANAJUATO

DEPARTAMENTO DE MATEMÁTICAS
CAMPUS GUANAJUATO

Proyecto Final

Reconocimiento Estadístico de Patrones

INTEGRANTES:

Dan Heli Muñiz Sanchez
Bryan Calderón Rivera

PROFESOR:

Johan Van Horebeek

03 de junio del 2023

Introducción

¿Que es MiBICI?

El Sistema de Bicicletas Públicas MiBici es un servicio de transporte individual disponible todos los días del año, basado en la renta de bicicletas en estaciones dispuestas a manera de red en las centralidades urbanas más importantes de la ciudad.

Para este proyecto, pusimos énfasis en analizar los últimos meses del año en curso (2023)

¿Qué variables interfieren?

Hay algunas variables que podemos ignorar como lo son "*Id _ Usuario*z "*Id_ Viaje*"puesto que no ofrecen ningún valor informativo.

Las demás variables que si tomaremos encuenta serán

1. **Género**: Que en los datos originales estan etiquetados como "H"para Hombres y "F"para mujeres, por practicidad se hara el mapeo de "0"para Hombres "1"para mujeres
2. **Año de Nacimiento**: Simplemente el año en el que nacio el usuario
3. **Inicio del Viaje**: con formato AA- MM-DD y HH:MM:SS de cuando empezo el viaje (Fue tomada la bicicleta)
4. **Fin del Viaje**: Con el mismo formato que la anterior pero con el registro de cuando fue dejada la bicicleta
5. **Origen Id**: Es la etiqueta de la estacion de donde empezó el viaje
6. **Destino Id**: Es la etiqueta de la estación de donde terminó el viaje

Como archivo anexo a este conjunto de datos , también obtenemos información como "**Longitud**z "**Latitud**"de cada estación, además de a que "**Zona**"pertenece cada estación, ya sea **Zapopan Centro**, **Poligono Central** y **Corredoras Atlas** , con esto y con los las variables anteriormente enumeradas , agregamos al conjunto de datos original las siguientes variables

1. **Distancia** : Esta es la distancia Geodesica (Se elige esta metrica por una mayor presición a la Distancia Euclideana) la unidad esta en Kilometros
2. **Duracion del Viaje:** Tiempo de llegada desde la estación Origen a la Destino, esta dada en minutos.

Con estas variables podemos hacer un analizis superficial con los datos que tenemos disponibles.

Quienes usan más este servicio, Hombres o mujeres, El siguiente gráfico sugiere que el uso es predominante por parte de los hombres.

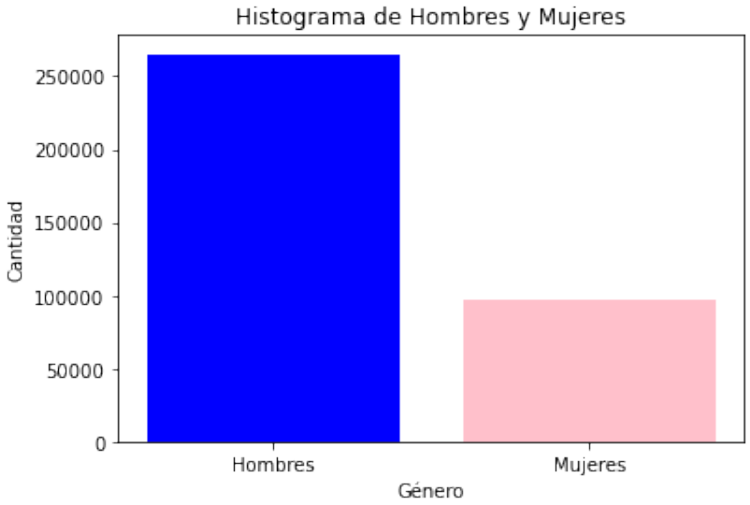


Fig 0.1 : Conteo de usarios hombres y mujeres.

Por otro lado , tambien no gustaria tener una idea si el servicio es princpialmente usado por ,**Adolecentes**, **Adultos Jovenes**, o **Personas Mayores** por mencionar algunos, de nueva cuenta usaremos la estrategia del histograma para hacernos de un panorama general.

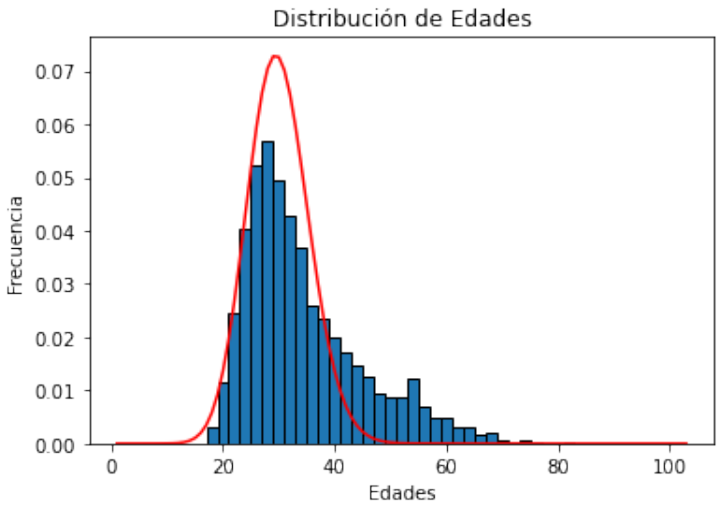


Fig 0.2 :Conteo de usarios por edades.

Observemos que en su mayoria son personas entre sus 20 años y 40 años de edad, ademas de que se ajusta muy bien a una distribución Poisson mas que a una normal, con $\lambda = 38{,}93$ que podriamos intepretar como el promedio en bruto de las edades.

Gracias a la Figura 0.1 y 0.2 , se puede concluir que el principal usuario de este servicio son hombres adultos y adultos jovenes.

Mas adelante veremos como esto se relaciona con el flujo de usuarios en horas laborales , puesto que ya que definimos a un grupo principa de participantes , es clave notar que dicho grupo esta en edad laboral.

Desarrollo

1. ¿En base a las características de un usuario, se puede predecir su destino?

Para esta pregunta, vamos a tratar de enfocarnos primero en una estación en particular y luego veremos el caso general.

Caso Particular

Vamos a utilizar como estación de origen la mas frecuente por los usuarios, para asi tener la mayor cantidad de información posibke. La estación de origen mas frecuente que es la número 51.

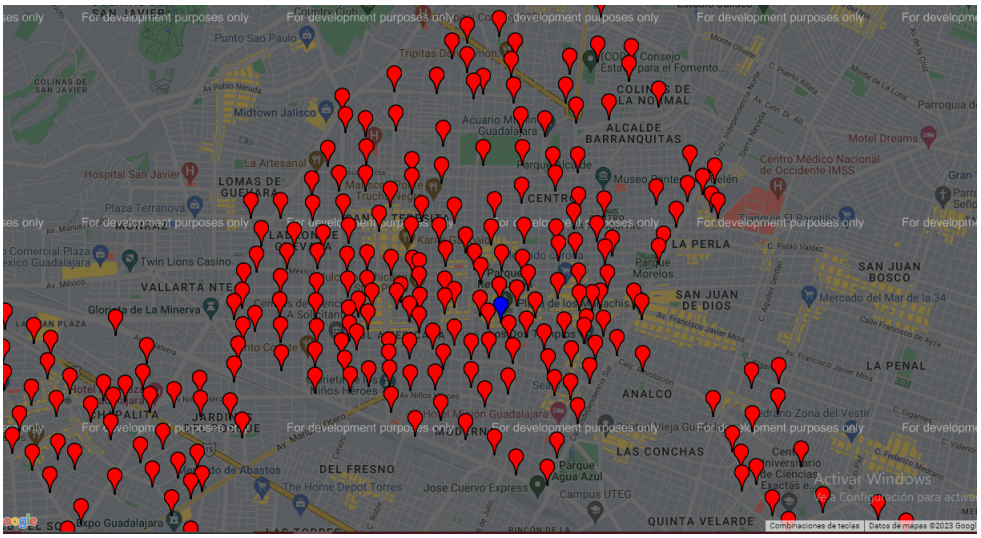


Fig 1.1: Estación Origen más frecuente

Tratar de predecir cada una de las estaciones de destino es algo muy complejo puesto que hay mas de 300 estaciones. Ademas, no existe gran correlación entre las caracteristicas de los usuarios y las estaciones de destino, por esto vamos a dividir las estaciones en zonas utilizando clustering con K-means. Al dividir nuestras estaciones en 6 grupos distintos obtenemos las siguientes 6 zonas:

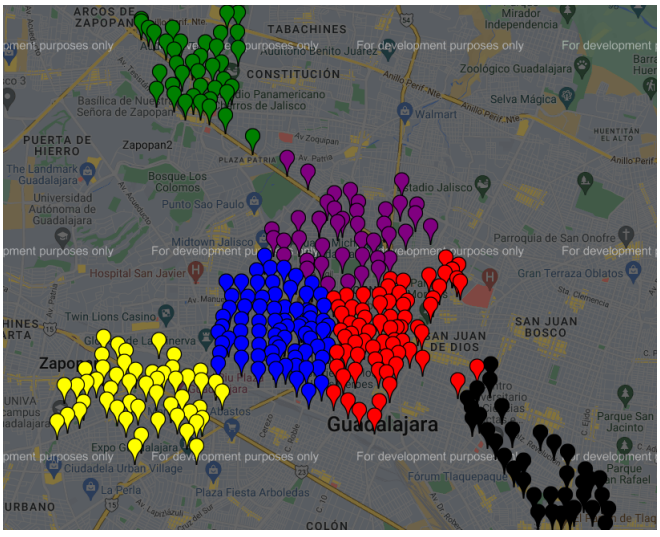


Fig 1.2: Grupos Generados por K-Means con K = 6

Añadimos a nuestros la columna 'Zona_destino' la cual indica en que zona se encuentra la estación destino de cada uno de los usuarios. Teniendo esto, vamos a tratar de predecir el destino de cada

uno de los usuarios en base a sus características y las características del viaje.

Tomamos como datos de entrenamiento un 75 % de los datos de nuestro dataframe(con estación de origen No. 51) de manera aleatoria. Ajustamos modelos de LDA y regresión logistica a nuestros datos. Luego, verficamos en nuestro conjunto de prueba y obtenemos una precisión de 0,634 y 0,636 respectivamente.

Utilizamos como caracteristicas predictoras: edad del usuario, genero del usuario, hora de inicio del viaje y duracion del mismo, siendo la duración del viaje la que mas afecta la predicción del destino. No se observa que las caracteristicas del usuario o la hora de salida afecten en gran manera el destino al que llega el usuario.

Caso General
Utilizamos las mismas 6 zonas de destinos. Como variables predictoras utilizamos las caracteristicas del usuario, la distancia recorrida, asi como la latitud y longitud de la estación de origen.

Ajustamos un modelo de regresion logistica al conjunto de entrenamiento. Al probar la precisión del modelo en nuestro conjunto de prueba obtuvimos un score de 0,42 siendo la latitud, la distancia recorrida y el genero los coeficientes que mas afectan a la predicción.

Ajustamos un modelo de LDA a nuestro conjunto de entrenamiento. al probar la precisión del modelo en nuestro conjunto de prueba obtuvimos un score de 0,58 siendo la distancia recorrida, la latitud y longitud los coeficientes que mas peso tienen al momento de predecir la zona de destino.

Ahora tomaremos otro acercamiento para intentar predecir "predecir.[el] destino de cada usuario, recordemos que cada estación pertenece a una de 3 zonas , que son POLÍGONO CENTRAL,,ZAPOPAN CENTRO, yTLQ-CORREDORATLAS, como se puede ver en la siguiente imagen.

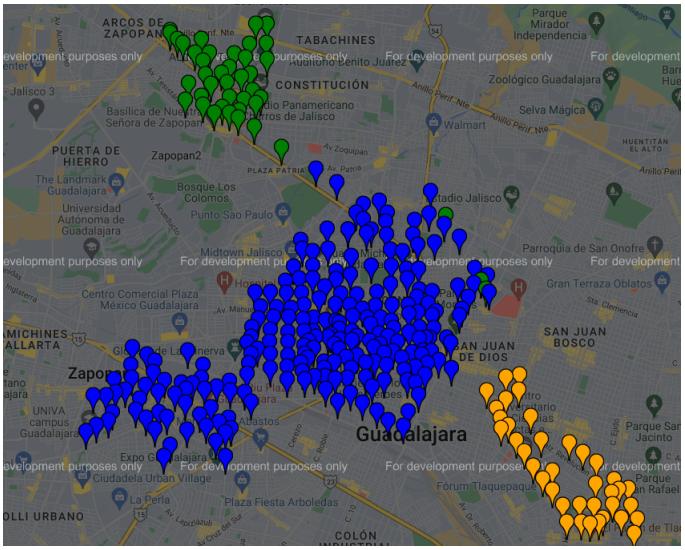
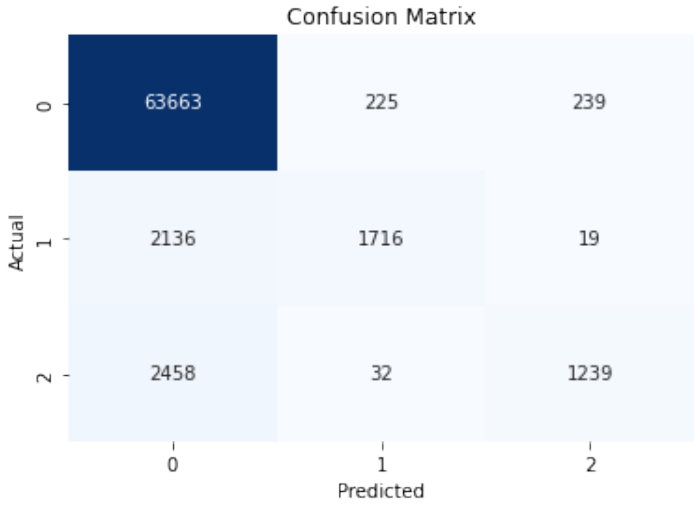


Fig 1.3: Las 3 Zonas principales de bicicletas por colores

Para ello utilizaremos el modelo de **RANDOMFOREST**, experimentando como afecta la exactitud de predicción cuando se toman diferentes variables. Entonces los casos siguientes evidencian el

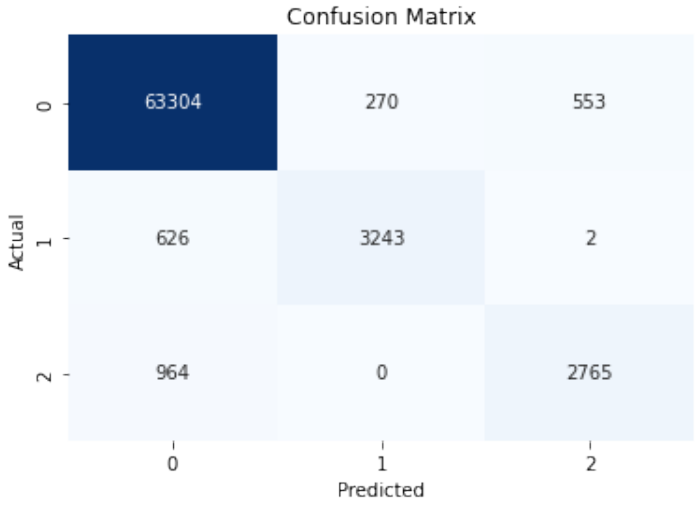
rendimiento en base a que variables tomamos.

1. **Genero**,**Edad**, **Duración del Viaje** y **Distancia**: obtenemos una precisión del modelo: 0.928771 y esta es la matriz de confución para cada una de las zonas



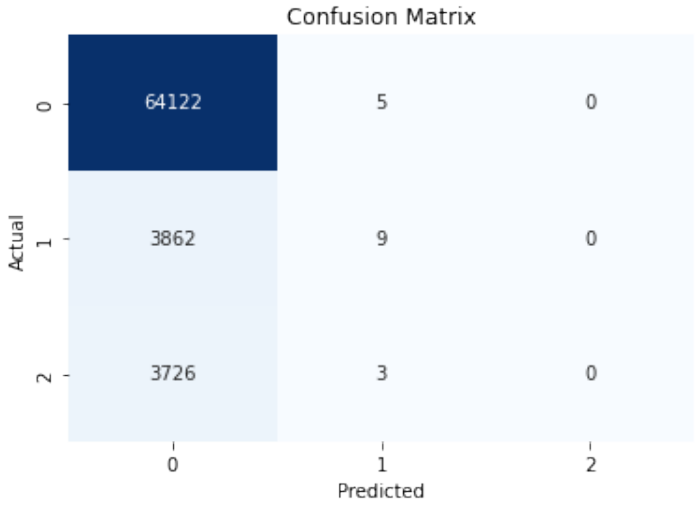
   Pero no debemos confiarnos de que este modelo sea el indicado para predecir a donde ira cada usuario, puesto que utiliza variables como **Duración del Viaje** y **Distancia** que estas variables las encontramos en base a las estaciones de Origen y Destino.

2. **Genero**,**Edad**,**Origen Id**: Ahora tomamos 2 variales que por el mapa de correlación antes mostrado, no tienen niguna relación entre si , y una variable aunque su naturaleza sea de etiqueta, nos da una idea de donde parte el usuario, Ahora tenemos un resultado de precisión: 0.9663, lo cual es impresionante porque nos dice que son saber dos atributos del usuario y su punto partida podemos predecir a cual de las 3 zonas llegara, esta es la matriz de confusión



   Pero como veremos mas adelante, los viajes suelen ser cortos en tiempo y por ende en distancia, por lo tanto la variable **Origen Id** es mas que suficiente porque el modelo predice basandose en que zona esta la estación de origen.

3. **Edad**,**Genero**: Ahora solo usaremos estas dos variables del usuario para ver si con ellas son suficientes para predecir su destino. Este caso es curioso porque obtenemos una precisión de 0.89409, que es mas que buena, pero al ver la matriz de confusión vemos un comportamiento extraño



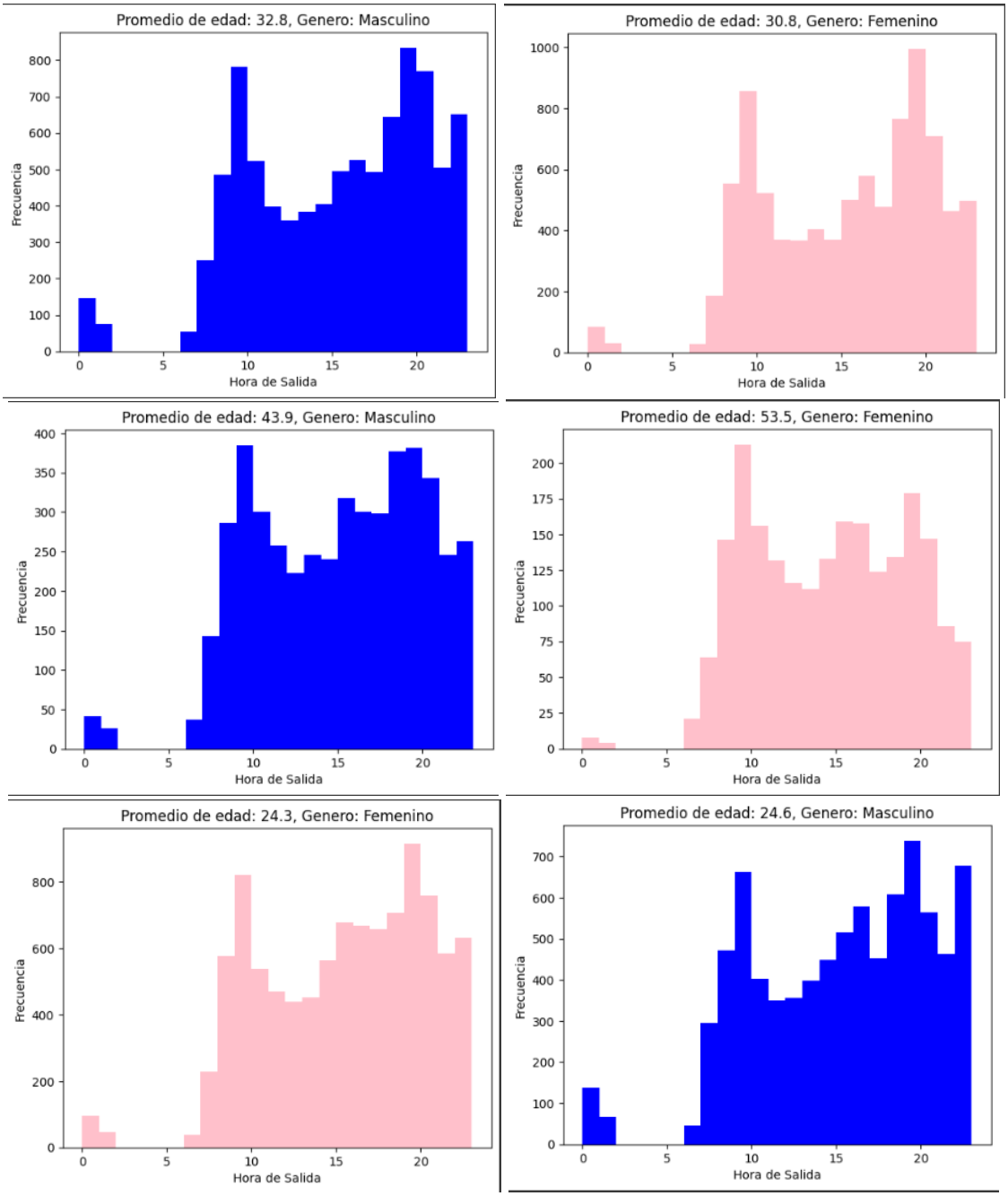
Veamos que ni siquiera clasifico usuarios para ir a la zona 2, esto significa que este modelo por la naturelza del mismo y de la variables, solo se fue a predecir en base a la mayoria de puntos que pertenecen a una zona, que en este caso seria **POLIGONO CENTRAL.**

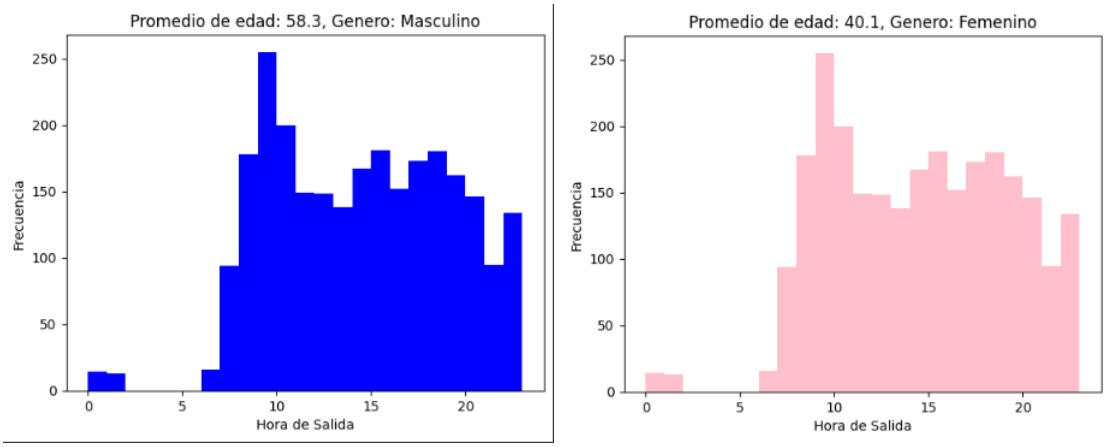
Conclusión

En los experimentos realizados tratamos de ajustar modelos para predecir los destinos de los usuarios tomando como datos predictores las características de los usuarios y del viaje. La predicción de la zona de destino fue decente, obteniendo un score de 0,6 aproximadamente en cada experimento. Al observar los coeficientes de dichos modelos pudimos observar que los que tuvieron mas peso fueron la distancia recorrida y el punto de partida, siendo asi las caracteristicas del usuario irrelevantes al momento de calcular el destino de este. Por lo que podemos concluir que no hay una correlación entre las características del usuario y el destino de este mismo.

2. ¿Que tipo de usuario utiliza este servicio a diferentes horas?

Las características de los usuarios con las que contamos en nuestro dataframe son la edad y el genero de cada uno de estos. Por lo que podemos agruparlos por el genero y por el rango de edad al que pertenecen los usuarios, utilizaremos un total de 8 clusters para asi tener 4 rangos de edades diferentes de cada genero.
Nos tomamos una muestra de nuestros datos que cuente con la misma cantidad de hombres y de mujeres. Luego, hacemos clustering con k-means tomando como parametros la edad y el genero de los usuarios(datos normalizados). Ya con esto, realizamos un histograma de cada uno de los clusters obtenidos y observamos la frecuencia con la que la gente usa el servicio a lo largo del dia. Los resultados obtenidos son los siguientes:





En general, podemos observar que los diferentes tipos de usuarios obtenidos suelen utilizar el servicio a lo largo del dia. Pero a pesar de esto podemos observar que en las personas con edad alrededor de los 30 años suele haber una densidad mas destacable de gente utilizando el servicio alrededor de las 9am y las 7pm, lo cual puede ser debido a los horarios laborales de entrada y de salida en dicha zona.

Esto tambien lo podemos ver en hombres y mujeres alrededor de los 24 años, donde los picos de gente a las 9am y 7pm destacan mucho sobre el resto de horas. Mientras que en hombres de edades cercanas a los 44 años y mujeres de edades cercanas a los 53 años tienden a utilizar de una forma mas uniforme a lo largo del dia respecto a los otros tipos de usuarios.

Luego, observamos a los hombres con edades cercanas a los 58 años y mujeres de edades cercanas a los 40 años. Estos usuarios tienen una mayor densidad a las 9pm y luego reducen el uso a lo largo del dia pero se mantiene casi constante el uso del servicio.

Conclusión

Los distintos grupos de personas obtenidos suelen utilizar el servicio a lo largo del dia, teniendo una mayor concentración de gente a las 9pm y 7pm. Aun asi, se puede destacar a los hombres con edad alrededor de los 43 años los cuales utilizan el servicio bastante a las distintas y no destacan tanto como en los otros grupos de gente los picos en las horas mencionadas. Ademas, de ser la gente con edades cercanas a los 30 años los que suelen utilizar mas el servicio en las horas pico y reducir considerablemente el uso de este a lo largo del dia, lo cual podria ser debido a los horarios laborales de la zona en la que se encuentran las estaciones.

Qué Patrón se Encuentra en la Saturación?

Una buena pregunta para ver cuan saturado esta el servicio es "**Cuanto tiempo duran las personas arriba de una bicicleta?**", y como se distribuye el tiempo de uso entre los usuarios.

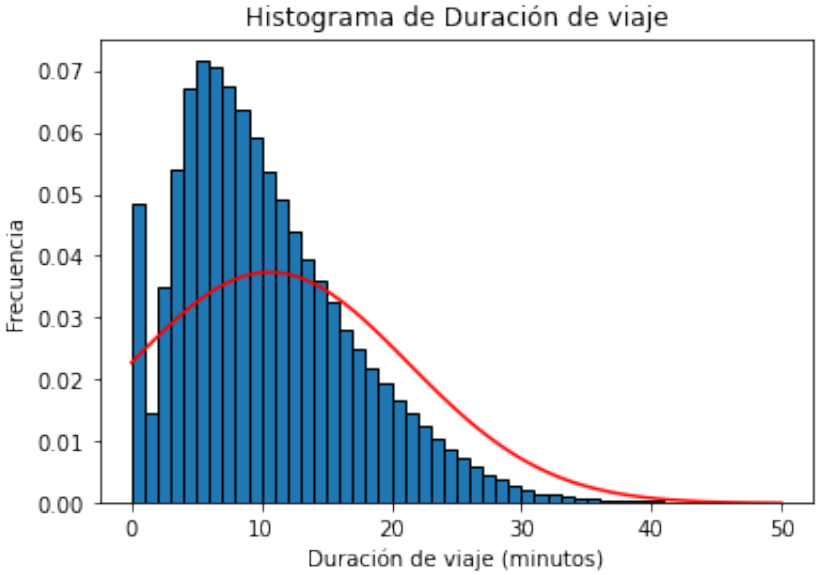


Fig 3.1 :Histograma y Distribución de la duración del viaje

tenemos una varianza de los datos de $\sigma^2 = 10{,}673$ y podemos ver gracias al histograma que la mayoria de los viajes no duran mas de 20 min. Entonces casi siempre seria ese el minimo a esperar para tomar una bicicleta en caso de que en una estacion no alla. Entonces como para tener una idea más amplia , nos deberiamos preguntar, "**A que hora el servicio se encuentra mas saturado?**"por lo que podemos ver en el siguiente histograma

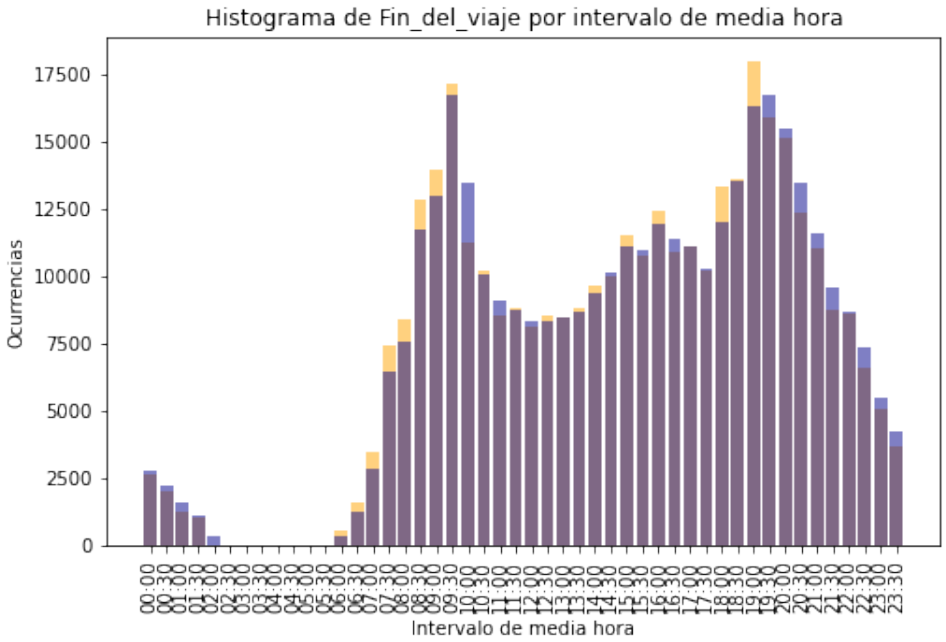


Fig 3.2 :Histograma Y Conteo de usuarios a cierta hora del dia

Para la figura 3.2 se aplicó el Algoritmo Para maximizar la experanza (EM) y obtner , los pesos, promedios, y varianzas de para este caso **2** curvas gaussianas que se ajusten mejor, el resultado fue el siguiente

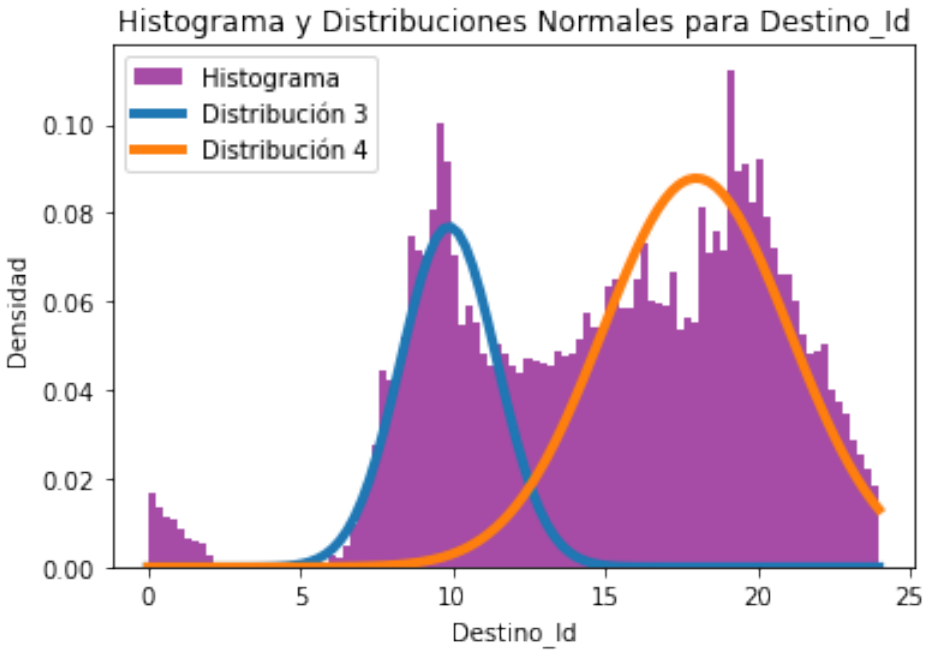


Fig 3.3 :Curvas gaussianas generadas por el algoritmo EM

Observemos como con este algoritmo de agrupamiento, obtenemos evidencia 2 grupos o como se puede interpretar 2 horarios de mayor saturación. **Qué podemos intuir por las curvas?** venamos que la distancia entre los 2 picos de las curvas es de aproximadamente 8 horas, y como el primer pico es en la mañana y el segundo por la tarde o noche se puede sacar la conclución que es debido a la entrada de los trabajos , el movimento sesa durante la jordana laborar y aumenta la saturación en los horarios de salida de los trabajadores.

Conclusiones

A pesar de la poca correlación que existe entre los datos, se pudo predecir de forma bastante descente las zonas a las que se dirigia cada uno de los usarios, siendo el tiempo de viaje la variable con mas peso para predecir la zona de destino. Así, como se pudo reconocer un patrón en las horas laborales que es donde solia haber una mayor densidad de gente utilizando las distintas estaciones.