Proyecto. Reconocimiento Estadístico de Patrones

Mercè Nachón Moreno

11 de Junio

Resumen

El objetivo del presente trabajo es poder encontrar un modelos estadísticos de tipo clasificatorio y predictivo, para contestar diversas preguntas hechas previamente en utilizando como referencia una base de datos de *MiBici*, un sistema de bicicletas públicas que presta servicio todos los días del año, basado en la renta de bicicletas disponibles en estaciones ubicadas en los puntos más importantes de la ciudad.

Se orientó el trabajo a las diferencias que pudiera haber entre el género femenino y masculino, además de las edades de estos grupos. Dado que se trabajo con datos naturales, obtuvimos en algunas preguntas malos resultados y en otras regulares. Al final se llegó a la conclusión que dependiendo de la categoría, el género de la usuario hace diferencia.

1. Introducción

La movilidad ecológica está echando raíces con la implementación de los sistemas de préstamo de bicicletas que los gobiernos de algunos estados de la república han implementado en sus localidades, teniendo como resultado, un aumento en los viajes realizados por medio de la bicicleta y cumpliendo con el propósito fundamental de desincentivar el uso del automóvil como medio de transporte.

En México existen cuatro sistemas de bicicleta pública. La Ciudad de México fue pionera con la implementación del sistema ECOBICI, el cual arrancó en febrero de 2010; de ahí, Guadalajara se unió a la marcha sustentable con su sistema **MiBici**, arrancando en diciembre de 2014; sumándose a estas dos, posteriormente se instalaron en Toluca el sistema Huizi, en noviembre de 2015, y muy recientemente, en febrero del año en curso, la ciudad de Pachuca con Bici Capital, innovando el sistema con bicicletas híbridas.

Aunque a paso más lento, en la Zona Metropolitana de Guadalajara, la propuesta de bici pública como parte de un sistema de transporte en las ciudades, también se ha replicado a través de **MiBici**, que es el segundo sistema de préstamo de bicicletas en el país y que logró en su primer año de implementación, más de 450 mil viajes.

Considerando la saturación vehicular que ha alcanzado a la fecha, la aceptación del sistema en Guadalajara ha sido favorable, pues según una encuesta realizada por el Instituto de Movilidad y Transporte de Jalisco, en un año de operaciones de **MiBici**, de los usuarios que migraron a esta modalidad para trasladarse, 52 % lo hacía antes en el transporte público; 22 % viajaba en automóvil, 12 % lo hacía en su propia bici, 11 % caminaba y 3 % se trasladaba en taxi.

En este trabajo estaremos utilizando con la base de datos de el sistema de bicicletas públicas **MiBici**, por lo que comenzaremos realizando un análisis exploratorio de este. Proseguimos con las secciones de análisis, planteamiento del problema y respuesta de las preguntas propuestas para los métodos de agrupación y predicción, terminando con conclusiones.

2. Análisis exploratorio.

Primero comencemos analizando los datos. En dichos datos abiertos contamos con bases de datos por mes comenzando en diciembre del año 2014 hasta el mes de mayo del presente año. Por el tipo de preguntas que

se plantearon, únicamente tomé 12 meses los que conformarían un año comenzado de junio del 2022 a mayo del 2023. Esto con el fin de contemplar la estacionalidad.

Al observar los datos contamos con información del usuario como su ID, genero y año de nacimiento. En cuanto al viaje como tal, tendremos la fecha en la que se realiza el viaje con hora inicial y final del trayecto, además del punto de partida y llegada. Se hicieron algunos cambios con la información que se tenía para un mejor trato de ella. Primero, se cambió el año de nacimiento por la edad, esta se calculó restándole a 2022 el año de nacimiento. Por otro lado en lugar de tener fecha y hora de inicio y fin del trayecto, se calculó la duración del trayecto restando ambas columnas y se agregó una columna con el mes en el que se realizó el trayecto. Por último se contempló el ID del usuario agregando la información de la frecuencia de uso. En todos los casos se quitaron los outliers como edades y duraciones poco lógicas.

Dicho lo anterior, obtenemos el gráfico de dispersión (1) de una muestra del 5 % conjunto de datos pintados en dos colores, azul para mujeres y naranja para hombres. Tomemos en cuenta que la cantidad de mujeres en el dataset completo es menos de un tercio de la cantidad de hombres, por lo tanto tiene sentido que las gráficas en la diagonal muestren que la distribución en mujeres sea menor que la de hombres. Con lo anterior, podemos observar que las horas de inicio y finalización de los trayectos se mantienen constantes en ambos grupos a lo largo de todos los grupos de edad, teniendo la mayor cantidad de viajes en bicicleta al rededor de las 18 horas. Mientras tanto, la duración y frecuencia de uso comienza a decrecer cuanto más edad tiene la persona, tiendo que la mayoría de viajes es de 0 a 25 minutos con una frecuencia de 50 viajes al mes. Por otro lado en cuestión de la edad de los usuarios, hay mayor concentración en los 20 a 35 años.

El boxplot en la figura (2), el cual contiene todo la base de datos, nos confirma que las conclusiones hechas con el gráfico de dispersión son correctas. Primero vemos una gran correlación entre hora de inicio y fin, teniendo a la mediana en las 17 horas. Por otro lado, podemos deducir que la duración promedio esta al rededor de los 12 minutos, teniendo la mayor concentración de datos entre los 9 y 17 minutos. Por último la mayor concentración de edades la muestra entre los 28 a 40 años. Con los usuarios realizando entre 25 a 50 viajes al mes. Con el mes más concurrido siendo Febrero.

3. Métodos de agrupamiento

Después del análisis hecho, comencemos con el planteamiento de la primer parte del proyecto. Para la parte de métodos de agrupamiento, decidimos realizar el siguiente listado de preguntas:

1. ¿Existen grupos o segmentos claramente distinguibles de usuarios según sus características demográficas?
2. ¿Se pueden identificar patrones de comportamiento de uso de las bicicletas compartidas en función de la hora de inicio y finalización de los viajes?
3. ¿Los usuarios se pueden agrupar en función de la duración promedio de sus viajes?
4. ¿Que relación existe entre la edad y las determinadas épocas del año?
5. ¿Es posible identificar grupos de usuarios con comportamientos de uso similares en términos de frecuencia de uso de las bicicletas?

En cada resultado de las preguntas anteriores, después de aplicar el método de segmentación, decidí hacer una reducción de dimensión para mostrar las agrupaciones con un *Análisis de componentes principales* (*PCA*), esto dado que esta parece aportar la mayor información visual de la clasificación. Notemos que al ver la tabla (1), en la que se expone la información que aporta cada una de las columnas del *dataframe* a los componentes, notamos que las variables que aportan mayor información son la frecuencia uso al mes, edad del usuario y duración del trayecto. En este caso la primer componente aporta ya un 81 % de la información y la varianza acumulada de las 3 componentes nos estaría dado 95 % de la información. Pese a que con la primera ya contamos con casi toda la información, me pareció importante y más visual mostrar en las tres componentes como se muestra en la figura (4).

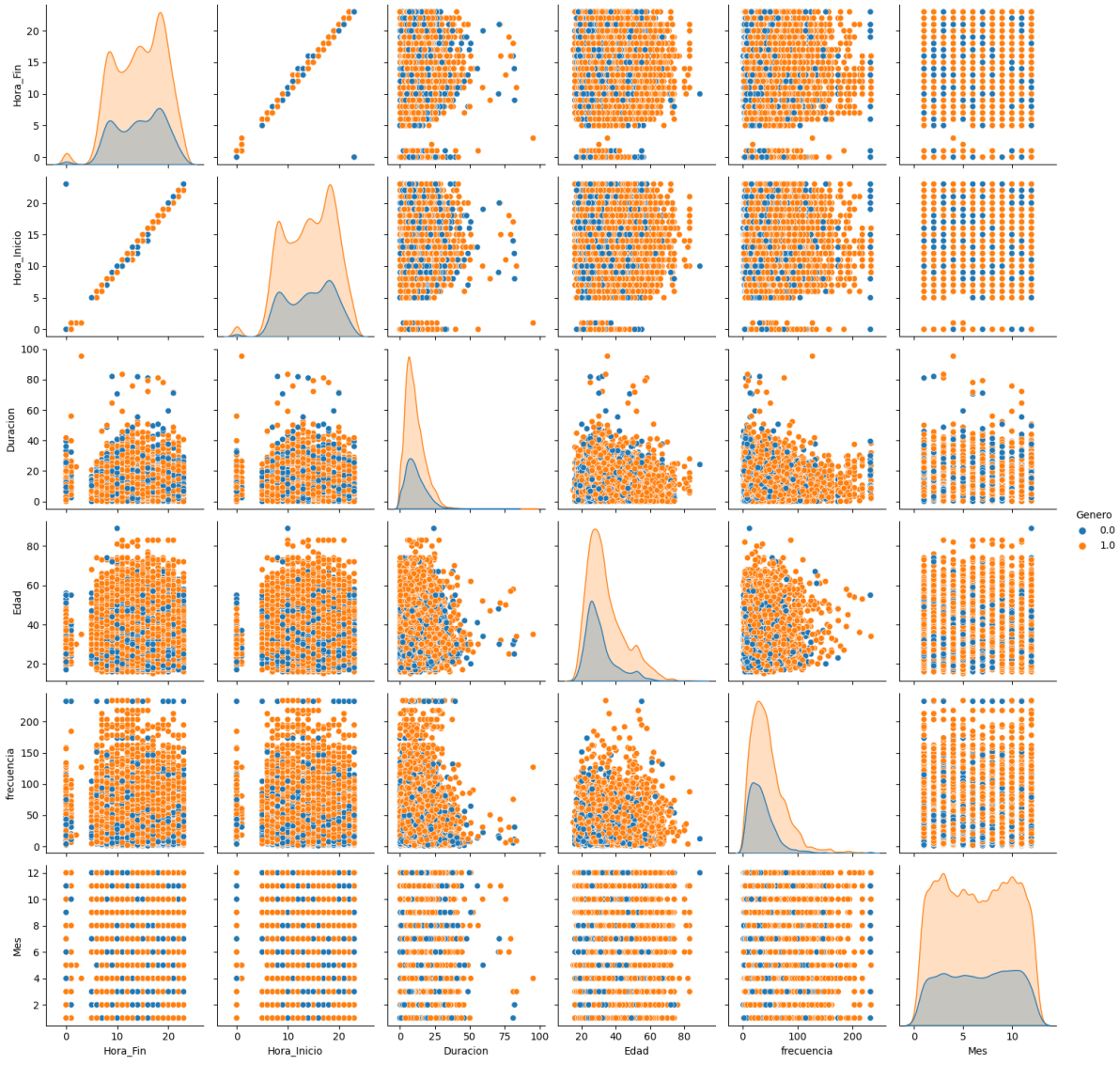


Figura 1: Gráfico de dispersión del 5 % de la base de datos. El color azul representa al género femenino, mientras que el naranja al masculino.

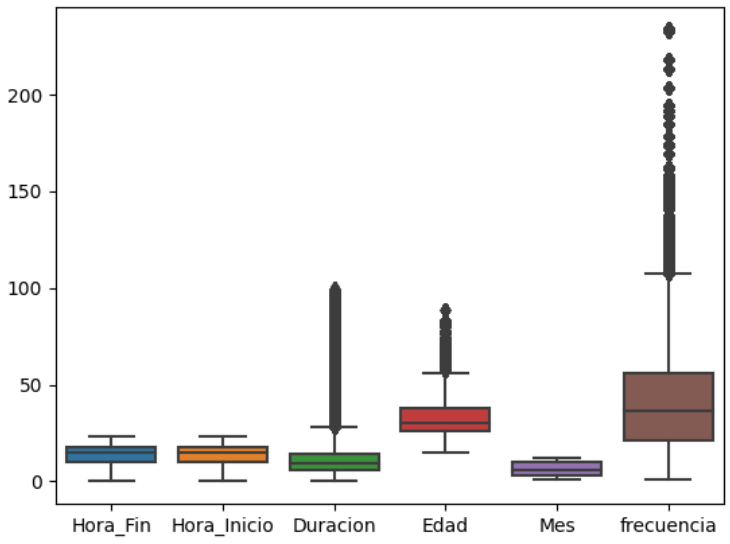


Figura 2: Boxplot de toda la base de datos seleccionada.

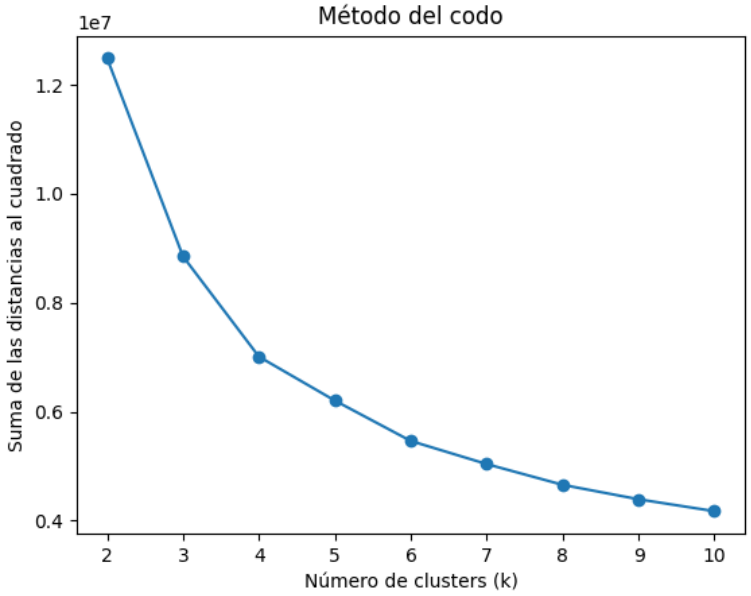


Figura 3: Método del codo, utilizado para saber que número de cluster es el más óptimo.

	Genero	Edad	Mes	Hora inicio	Hora fin	Duración	Frecuencia al mes
1	0.002500	0.054599	-0.001515	-0.003171	-0.003407	0.002009	0.998491
2	0.004298	0.996472	0.003055	-0.035797	-0.034629	0.039142	-0.054806
3	-0.001301	-0.025234	0.000826	0.174199	0.185467	0.966756	0.000625

Cuadro 1: Carga de información de cada variable en cada componente.

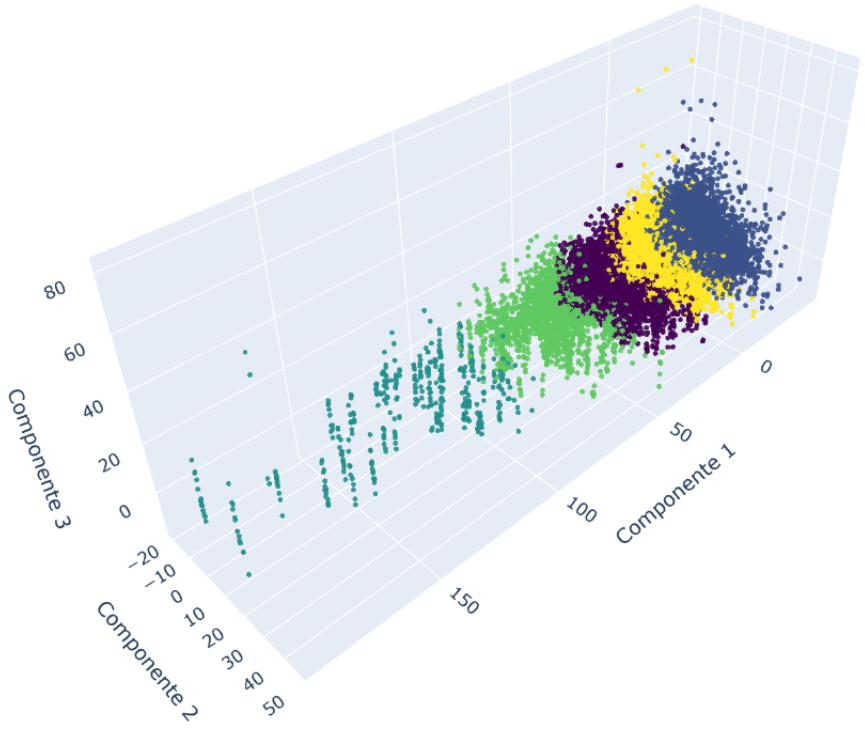


Figura 4: *PCA* de tres componentes de las categorías género, edad, mes de viaje, hora de inicio y finalización, duración y frecuencia de uso al mes. Se ven los clusters de la quinta pregunta.

Primero contestemos la última pregunta utilizando *K-medias* evaluando únicamente la frecuencia. Para saber que variable k utilizar, implementé el método del codo. En este caso obtuve como resultado el observado en la figura (3). Como no se muestra una respuesta clara de que número de cluster sería el mejor, dado que no contamos con un valor que defina claramente un 'codo', tomaremos $k = 5$ dado que a partir de este en todos casos comienza a realizar un descenso más uniforme. Ahora, con la k anterior obtenemos las separaciones vistas en (4). En estas podemos notar que en la figura los clusters están hechos perpendicularmente al eje de la componente 1, indicándonos que únicamente toma en cuenta esta variable para realizar separación, esto es más obvio al evaluar los centroides de cada grupos. En estos la única categoría en la que realmente hay un cambio es en la frecuencia, el resto se mantienen con los promedios generales de la base de datos. Así podríamos contestar la pregunta 5 concluyendo que no es posible identificar grupos en relación a ninguna otra categoría cuando tomamos en cuenta la frecuencia de los usuarios al mes.

Si tomamos en cuenta las categorías demográficas, es decir la edad y el género, para realizar *K-medias*, tomando a $k = 6$, obtendremos como resultado la figura (5). En este observamos que se categoriza en función a la edad, pero contrario al caso anterior veremos un cambio por grupo de frecuencia de uso. Tendremos los grupos con promedios de edad de 21, 27, 33, 40, 50 y 61 años los cuales cuentan con un promedio de frecuencia de uso de 40, 40, 42, 43, 53 y 54 veces al mes, respectivamente. El resto de categorías se mantienen constantes

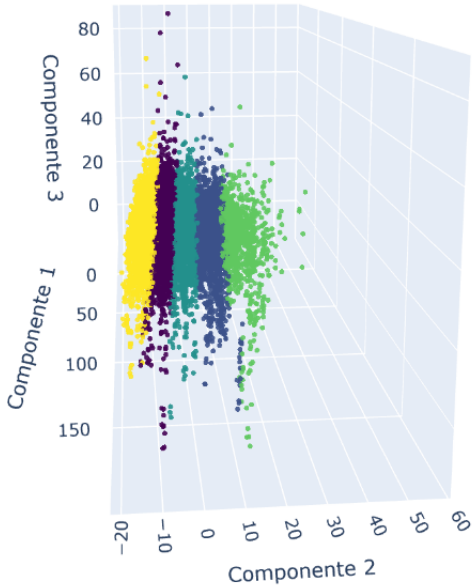


Figura 5: *PCA* de los clusters de la primer pregunta.

con los promedios generales de cada una. Con esto podríamos contestar la primer pregunta diciendo que no existen grupos claramente distinguibles, aún que se podría concluir que cuando la edad aumenta, mayor la frecuencia en el uso de bicicletas al mes.

Para contestar la segunda pregunta, evaluamos con un modelo de *agrupamiento espectral*, la hora de inicio y finalización de los viajes. En este caso para elegir k implementamos el método del codo se obtuvo como mejor número de clusters 6. En este caso llegamos visualmente a algo no distinguible con *PCA*, pero con *T-SNE* obtenemos algo aparentemente más seccionado como se ve en la figura (6). En este caso los horarios muestran relación no muy fuerte, con la duración, frecuencia de uso y edad, como lo mostrado en la tabla (2).

	Hora inicio	Hora fin	Duración	Edad	Frecuencia
0	17.060120	17.239518	10.970488	32.849518	41.069209
1	23.000000	0.000000	13.925000	31.210526	46.895804
2	0.077348	0.176796	10.616759	31.569061	51.714785
3	7.731617	7.881654	9.803553	33.585541	44.661786
4	20.929423	21.051027	10.481069	31.960901	45.109476
5	12.007142	12.187213	10.925282	33.895596	42.033774

Cuadro 2: Tabla resultado de clasificación *agrupamiento espectral* con horarios de inicio y fin de viaje

Para la cuarta pregunta hacemos nuevamente *agrupamiento espectral* con $k = 4$, con la elección hecha con el método de la silueta. Para dicho método únicamente tomamos en cuanta las categorías de edad y mes. La imagen del resultado esta realizada con *T-SNE* y es la mostrada en la (7). Observemos que se tiene una notoria división entre los grupos amarillo y morado del verde y azul, aún que entre cada uno de ellos no hay una diferencia clara. Creo que la razón de esto es más clara cuando vemos los centroides ya que contamos con dos grupos con promedio de mes en 3 y los otros dos en 9, pero las edades promedio estarán en 30,5 y 54,6 para el mes de marzo mientras que para el mes de septiembre tendremos 52,9 y 29,5. Las frecuencias de cada grupo son: 41,6, 53,5, 58,2 y 40,1, respectivamente. Esto nos podría estar indicando que existen dos grandes grupos de personas en la primera y segunda parte del año que comprenden edades de 30 y 53 años,

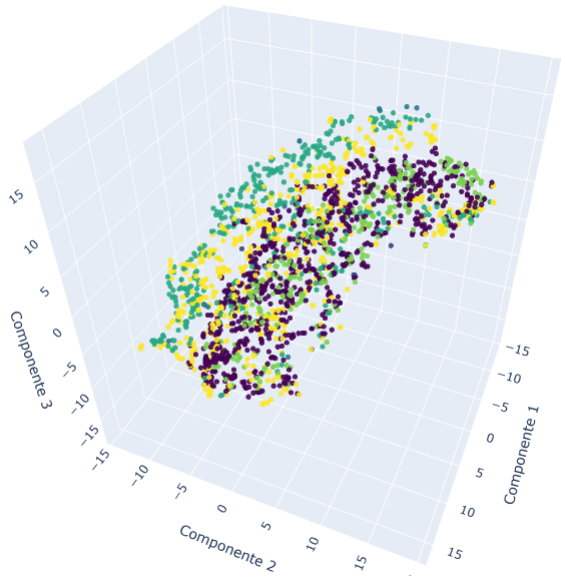


Figura 6: *T-SNE* del tercer *K-medias*.

donde los de mayor edad tienen una mayor frecuencia de uso a lo largo del año.

4. Métodos de predicción

Pasemos ahora a los métodos de predicción. En este caso utilizamos modelos como *K-nearest neighbors*, *Regresión Lineal* y *Regresión Logística*. Para los cuales propusimos contestar a las siguiente preguntas:

1. ¿Existe una diferencia significativa en la duración promedio de los viajes entre géneros? Para la primera, implementamos regresión lineal dado que la variable dependiente (duración promedio de los viajes) es una variable continua y numérica en lugar de una variable categórica o binaria. Obtenemos los resultados mostrados en la tabla (8).

   En el resultado de la regresión lineal, el Coeficiente constante (const) siendo de 10,8494 representa la duración promedio de los viajes para las mujeres (cuando Genero es 0) en minutos. Se sigue que el coeficiente *Genero* cuyo valor es $-0{,}2675$, indica que, en promedio, los hombres (cuando Genero es 1) tienen una duración promedio de viaje 0,2675 minutos menor que las mujeres. La duración promedio de los viajes de los hombres es menor en comparación con las mujeres. En cuanto al valor p ($P > |t|$) asociado al coeficiente de Genero es de 0, lo que indica una significancia estadística. Esto sugiere que la diferencia observada en la duración promedio de los viajes entre géneros no es simplemente el resultado del azar, sino que hay una relación estadísticamente significativa. El valor *R-cuadrado* es 0.000, lo que significa que el modelo de regresión lineal no explica una cantidad significativa de la variabilidad en la duración de los viajes. Esto indica que otros factores pueden influir más en la duración de los viajes además del género.

   En resumen, de acuerdo con los resultados de la regresión lineal, existe una diferencia significativa en la duración promedio de los viajes entre géneros. Los hombres tienden a tener una duración promedio de viaje menor en comparación con las mujeres. Sin embargo, la variabilidad en la duración de los viajes no se explica en gran medida por el género, lo que sugiere la presencia de otros factores influyentes en la duración de los viajes.

2. ¿Hay diferencias en los patrones de uso de las bicicletas compartidas entre géneros en términos de la hora de inicio y finalización de los viajes?

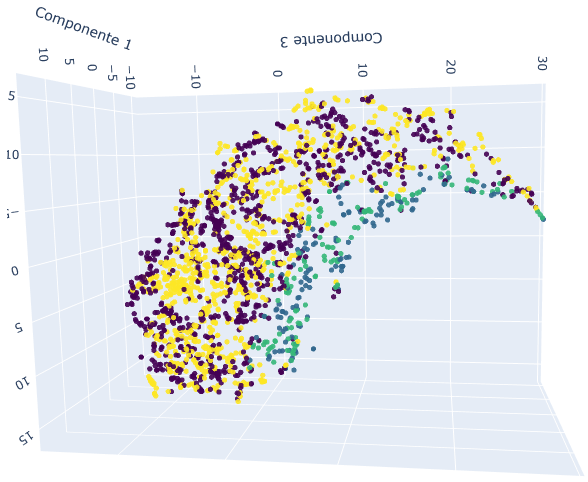


Figura 7: *T-SNE* en respuesta a la cuarta pregunta.

Utilizamos regresión logística, dado que queremos predecir mediante los horarios de inicio y fin de los trayectos si el usuario es mujer o hombre. En este caso el resultado del modelo se puede ver en la tabla (9).

La cual nos indica que la hora de inicio y final afectan a la predicción del género de la siguiente manera:

- Hora de inicio: Un incremento en la hora de inicio del viaje se asocia con un ligero aumento en la probabilidad de que el usuario pertenezca al género femenino, manteniendo constantes las otras variables. Sin embargo, la magnitud de este efecto es relativamente pequeña.
- Hora de finalización: Un aumento en la hora de finalización del viaje está asociado con una ligera disminución en la probabilidad de que el usuario pertenezca al género femenino, teniendo en cuenta las demás variables. Nuevamente, este efecto es de magnitud pequeña.

En resumen, en el modelo de regresión logística, la hora de inicio y finalización del viaje tienen un impacto estadísticamente significativo pero pequeño en la predicción del género.

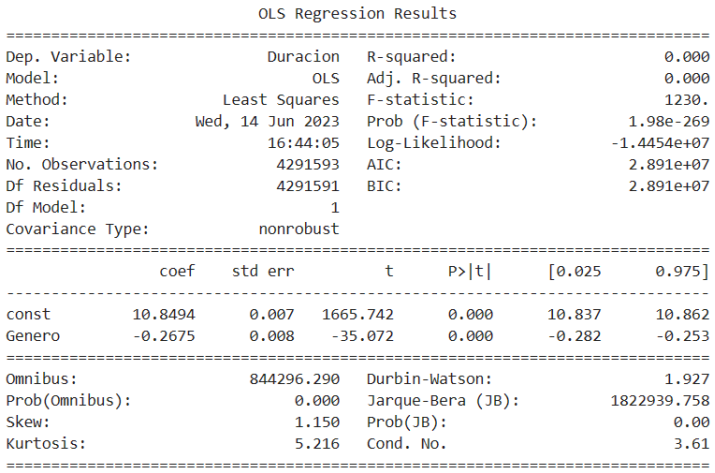
3. ¿Existe alguna relación entre la edad y la duración del trayecto por género?

   Volví a implementar regresión lineal con la edad de cada grupo, es decir de las mujeres y los hombres, para predecir la duración. Obtuve la siguientes tablas (**??**) como resultado.

   Los resultados obtenidos de los modelos de regresión lineal para hombres y mujeres indican que la edad tiene una relación significativa con la duración del recorrido para ambos géneros. Veamos lo por género:

   - Para el caso de las mujeres, el coeficiente de regresión para la edad es de 0.0209, lo que significa que, en promedio, se espera un aumento de 0.0209 en la duración del recorrido por cada año adicional de edad. El valor p asociado al coeficiente es prácticamente cero, lo que indica que la relación es estadísticamente significativa.
   - En el caso de los hombres, el coeficiente de regresión para la edad es de 0.0242. Esto implica que, en promedio, se espera un aumento de 0.0242 en la duración del recorrido por cada año adicional de edad. Al igual que en el caso de las mujeres, el valor p asociado al coeficiente es prácticamente cero, lo que indica una relación estadísticamente significativa.

   Es importante tener en cuenta que el coeficiente de determinación (R-cuadrado) es muy bajo en ambos modelos (0.001), lo que indica que la edad explica solo una pequeña proporción de la variabilidad en la



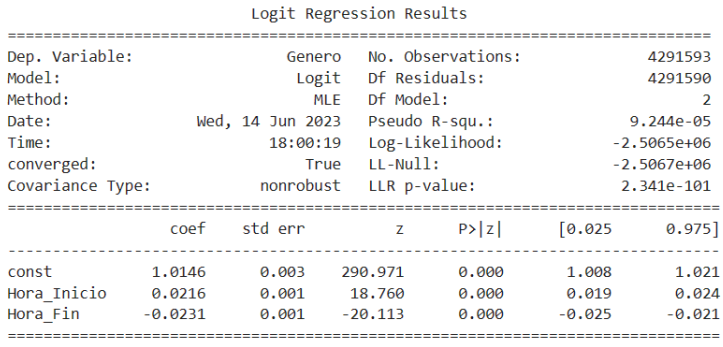
OLS Regression Results

Dep. Variable:	Duracion	R-squared:	0.000
Model:	OLS	Adj. R-squared:	0.000
Method:	Least Squares	F-statistic:	1230.
Date:	Wed, 14 Jun 2023	Prob (F-statistic):	1.98e-269
Time:	16:44:05	Log-Likelihood:	-1.4454e+07
No. Observations:	4291593	AIC:	2.891e+07
Df Residuals:	4291591	BIC:	2.891e+07
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	10.8494	0.007	1665.742	0.000	10.837	10.862
Genero	-0.2675	0.008	-35.072	0.000	-0.282	-0.253

Omnibus:	844296.290	Durbin-Watson:	1.927
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1822939.758
Skew:	1.150	Prob(JB):	0.00
Kurtosis:	5.216	Cond. No.	3.61

Figura 8:



Logit Regression Results

Dep. Variable:	Genero	No. Observations:	4291593
Model:	Logit	Df Residuals:	4291590
Method:	MLE	Df Model:	2
Date:	Wed, 14 Jun 2023	Pseudo R-squ.:	9.244e-05
Time:	18:00:19	Log-Likelihood:	-2.5065e+06
converged:	True	LL-Null:	-2.5067e+06
Covariance Type:	nonrobust	LLR p-value:	2.341e-101

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.0146	0.003	290.971	0.000	1.008	1.021
Hora_Inicio	0.0216	0.001	18.760	0.000	0.019	0.024
Hora_Fin	-0.0231	0.001	-20.113	0.000	-0.025	-0.021

Figura 9:

duración del recorrido. Esto sugiere que otros factores también pueden estar influyendo en la duración del recorrido, además de la edad. En resumen, según los resultados de estos modelos de regresión lineal, existe una relación significativa entre la edad y la duración del recorrido tanto para hombres como para mujeres. Sin embargo, la edad por sí sola no explica la mayor parte de la variabilidad en la duración del recorrido, por lo que pueden haber otros factores que estén influyendo en esta relación.

4. ¿El género influye en la preferencia de las estaciones de origen y destino?

   Por último contestemos la pregunta con *k-nn*. Evaluamos el modelo con el id del origen y el destino además del horario. El modelo obtuvo una exactitud del 74 % con la siguiente matriz de confusión vista en la figura (11).

   Podemos notar que hace una buena predicción en hombres, pero no en mujeres. Esto puede ser por la naturaleza de los datos dado a que se tienen menos información de mujeres que hombres en la base de datos. En el momento en el que equilibramos los datos con el mismo número en ambas categorías reduce la exactitud a un 65 % aun que la predicción es más equilibrada ya que ahora en ambos casos se equivoca más o menos en el mismo porcentaje, que antes la equivocación en las mujeres era mucho mayor. En cualquier caso, podemos inferir que el género puede influir en la preferencia de las estaciones de origen y destino, ya que el modelo muestra un desempeño superior al 50 %.

OLS Regression Results

Dep. Variable:	Duracion	R-squared:	0.001
Model:	OLS	Adj. R-squared:	0.001
Method:	Least Squares	F-statistic:	834.4
Date:	Wed, 14 Jun 2023	Prob (F-statistic):	2.04e-183
Time:	19:33:34	Log-Likelihood:	-3.9258e+06
No. Observations:	1162349	AIC:	7.852e+06
Df Residuals:	1162347	BIC:	7.852e+06
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	10.2042	0.023	438.251	0.000	10.159	10.250
Edad	0.0209	0.001	28.887	0.000	0.019	0.022

Omnibus:	225105.601	Durbin-Watson:	1.909
Prob(Omnibus):	0.000	Jarque-Bera (JB):	489339.789
Skew:	1.130	Prob(JB):	0.00
Kurtosis:	5.236	Cond. No.	114.

OLS Regression Results

Dep. Variable:	Duracion	R-squared:	0.001
Model:	OLS	Adj. R-squared:	0.001
Method:	Least Squares	F-statistic:	4431.
Date:	Wed, 14 Jun 2023	Prob (F-statistic):	0.00
Time:	19:33:01	Log-Likelihood:	-1.0525e+07
No. Observations:	3129244	AIC:	2.105e+07
Df Residuals:	3129242	BIC:	2.105e+07
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	9.7619	0.013	754.616	0.000	9.737	9.787
Edad	0.0242	0.000	66.566	0.000	0.023	0.025

Omnibus:	619238.130	Durbin-Watson:	1.942
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1335496.277
Skew:	1.156	Prob(JB):	0.00
Kurtosis:	5.212	Cond. No.	117.

Figura 10: Resultados regresión lineal Mujeres (izquierda) y Hombres (derecha).

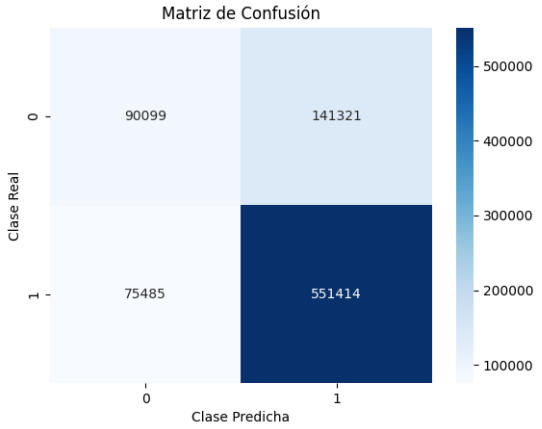


Figura 11:

5. Conclusiones

Se trato de darle un enfoque al proyecto en relación a las diferencias del uso del servicio entre géneros. Pese a que al principio pensé que podríamos encontrar mayores discrepancias, principalmente en horarios y zonas en las que se hacia uso del servicio, estoy bastante satisfecha en concluir que aun que si las pueden haber no son diferencias grandes. Además fue interesante trabajar con datos reales y tratar de encontrar los mejores modelos que se ajustaran a los objetivos de cada pregunta.

Referencias

[1] MiBiciPublica. Acerca de Mibici, MiBici. Disponible en: https://www.mibici.net/es/acerca-de-mibici. (Visitado: 12 June 2023).

[2] Robles, A. Los sistemas de bicicletas públicas en México, Noticias Pasajero7. Disponible en: http://www.pasajero7.com. (Visitado: 14 June 2023).