Proyecto final
Reconocimiento Estadístico de Patrones

Luis Ramón Guajardo — Universidad de Guanajuato
Jafet Castañeda — 6 de octubre de 2023

---

1. Introducción

El objetivo de este proyecto se centra en encontrar modelos estadísticos que describan las relaciones multivariadas que se presentan en el conjunto de datos abiertos Mibici (Sistema de Bicicletas Publicas del AMG) y usar dichos modelos para poder responder preguntas de interés sobre los datos en cuestión . Para lograr lo anterior se utilizaron distintos métodos de análisis de datos multivariados. Por ejemplo, primero se utilizó Pca (Análisis de Componente Principales) para la reducción de dimensiones con el fin de trabajar con datos proyectados en las direcciones de una mayor varianza.

Además implementamos métodos de agrupamiento con el fin de establecer una posible visualización sobre la futura clasificación o predicción de los datos con base una variable de interés. En específico hicimos uso de *dendogramas* y k-*means*. Por lo tanto, al final recaímos plenamente en modelos predictivos para encontrar soluciones a diferentes preguntas. Se consideraron modelos como Árboles de Decisión y el Clasificador Bayesiano Óptimo (ambos vistos en el curso).

Ahora, algunas preguntas que en las secciones anteriores procederemos a responder son del siguiente estilo:

- ¿Cuál es el poder predictivo de las variables para predecir el destino del usuario?
- ¿El poder predictivo para la duración del viaje depende del genero, la edad, entre otras?
- ¿Existe una relación entras las diferentes variable del viaje para clasificar o predecir la hora en la que se realiza el viaje?

Notemos que nos estaremos enfocando en preguntas sobre cómo cambia el poder predictivo para una característica de interés en función de las demás variables. Para poder encontrar una solución a dichas preguntas, primero debemos de realizar una descripción detallada del conjunto muestra. Es importante recalcar que los resultados obtenidos se basan en datos reales, por lo que, es posible obtener errores de predicción no deseables. Con esto establecido, realizamos una presentación corta de Mibici. Primero, se cumple que Mibici es un proyecto de la Agencia de Infraestructura para la Movilidad en la ciudad de Guadalajara, Jalisco., el cual consta de un Sistema de Bicicletas Publicas del Área Metropolitana de Guadalajara con el que un usuario puede moverse de manera libre y ecológica por la ciudad. Un usuario sigue los siguientes pasos para hacer uso de las bicicletas.

(1) Liberar la bicicleta en una estación, con una tarjeta o código

(2) Pedalear hasta el destino

(3) Regresar la bicicleta en cualquier estación cercana al destino

En la siguiente sección presentamos una descripción detallada de los datos registrados en Mibici.

2. Descripción general de los datos

Como se estableció en la sección anterior, trabajamos con los datos de Mibici, los cuales se presentan como un *dataframe* de registros mensuales. En dicho *dataframe* obtenemos como características: **ID del viaje**, **ID del usuario**, **Genero** (M o F), **Año de Nacimiento**, **Inicio del viaje** (día y hora), **Fin del viaje**, **ID de estación de origen** y **ID de estación de destino** . A continuación se presenta una imagen del formato anterior.

Viaje_Id	Usuario_Id	Genero	Ano_nacin	Inicio_del_viaje	Fin_del_viaje	Origen_Id	Destino_Id
28467098	70123	M	1967	1/5/2023 0:00	1/5/2023 0:22	64	141
28467099	2237235	M	1980	1/5/2023 0:00	1/5/2023 0:04	36	172
28467100	2051727	F	2002	1/5/2023 0:01	1/5/2023 0:10	96	296
28467101	2246225	M	1969	1/5/2023 0:01	1/5/2023 0:04	33	255
28467102	324247	M	1975	1/5/2023 0:01	1/5/2023 0:13	226	231

Figura 2.1: Dataframe Original

Notemos que del *dataframe* anterior podemos obtener nuevas variables de interés tanto continuas como categóricas. Dados el Inicio del viaje y el Fin del viaje calculamos la **Duración total del viaje**, también es posible calcular la **Edad del usuario**, el **Día del viaje** (Lunes - Domingo) y la **Hora del viaje** (0-23). Por lo tanto, obtenemos el siguiente *dataframe*

	Viaje_Id	Usuario_Id	Genero	Ano_nacimiento	Inicio_del_viaje	Fin_del_viaje	Origen_Id	Destino_Id	Edad	Duracion_de_viaje	Dia	Hora_del_viaje
0	28467098	70123	M	1967.0	01/05/2023 00:00	01/05/2023 00:22	64	141	56	22.266667	lunes	0
1	28467099	2237235	M	1980.0	01/05/2023 00:00	01/05/2023 00:04	36	172	43	4.050000	lunes	0
2	28467100	2051727	F	2002.0	01/05/2023 00:01	01/05/2023 00:10	96	296	21	9.266667	lunes	0
3	28467101	2246225	M	1969.0	01/05/2023 00:01	01/05/2023 00:04	33	255	54	3.116667	lunes	0
4	28467102	324247	M	1975.0	01/05/2023 00:01	01/05/2023 00:13	226	231	48	11.866667	lunes	0

Figura 2.2

Ahora, la pagina de Mibici también nos proporciona un *dataframe* con la nomenclatura de cada estación. A cada estación se le asigna una **Zona**, **Latitud** y **Longitud** (coordenadas). Con esto establecido, podemos construir las siguientes variables de interés

- **Genero - Binario** → En la Figura 2.2 podemos observar que obtenemos el Genero en formato M y F, por lo tanto, realizamos un cambio a un formato binario donde 0 representa Masculino y 1 Femenino. Hacemos el cambio para obtener una variable categórica que podemos representar de manera numérica.
- **Distancias** → Dadas las latitudes y longitudes de cada estación, calculamos las distancias geodésicas entre la estación de origen y la estación de destino de cada usuario. Así, sabemos una distancia aproximada que fue recorrida por el usuario. En las siguientes secciones veremos que existe una correlación entre la distancia y la duración.
- **Día (1 - 7)** → De nuevo, dado que en la Figura 2.2 obtenemos los días en formato de texto, realizamos un cambio a valores numéricos, donde 1 corresponde a Lunes y de manera sucesiva obtenemos que 7 corresponde a Domingo.
- **Intervalo de duración** → Debido a que una variable de interés es la duración, queremos categorizar dicha variable, por lo que, realizamos dicha categorización al dividir la duración (mins) en intervalos con IDs continuos. Es importante mencionar que obtuvimos 8 intervalos los cuales tienen una relación sucesiva, es decir 1 corresponde al intervalo de tiempo más pequeño y 8 al intervalo más grande.

- **Zona de Origen y Destino** → En el mismo *dataframe* de la nomenclatura podemos obtener la siguiente clasificación de las zonas por nombre y adjuntarles una etiqueta o clase correspondiente al id de la estación.

Clase	Zona
0	POLÍGONO CENTRAL
1	ZAPOPAN CENTRO
2	TLQ-CORREDORATLAS

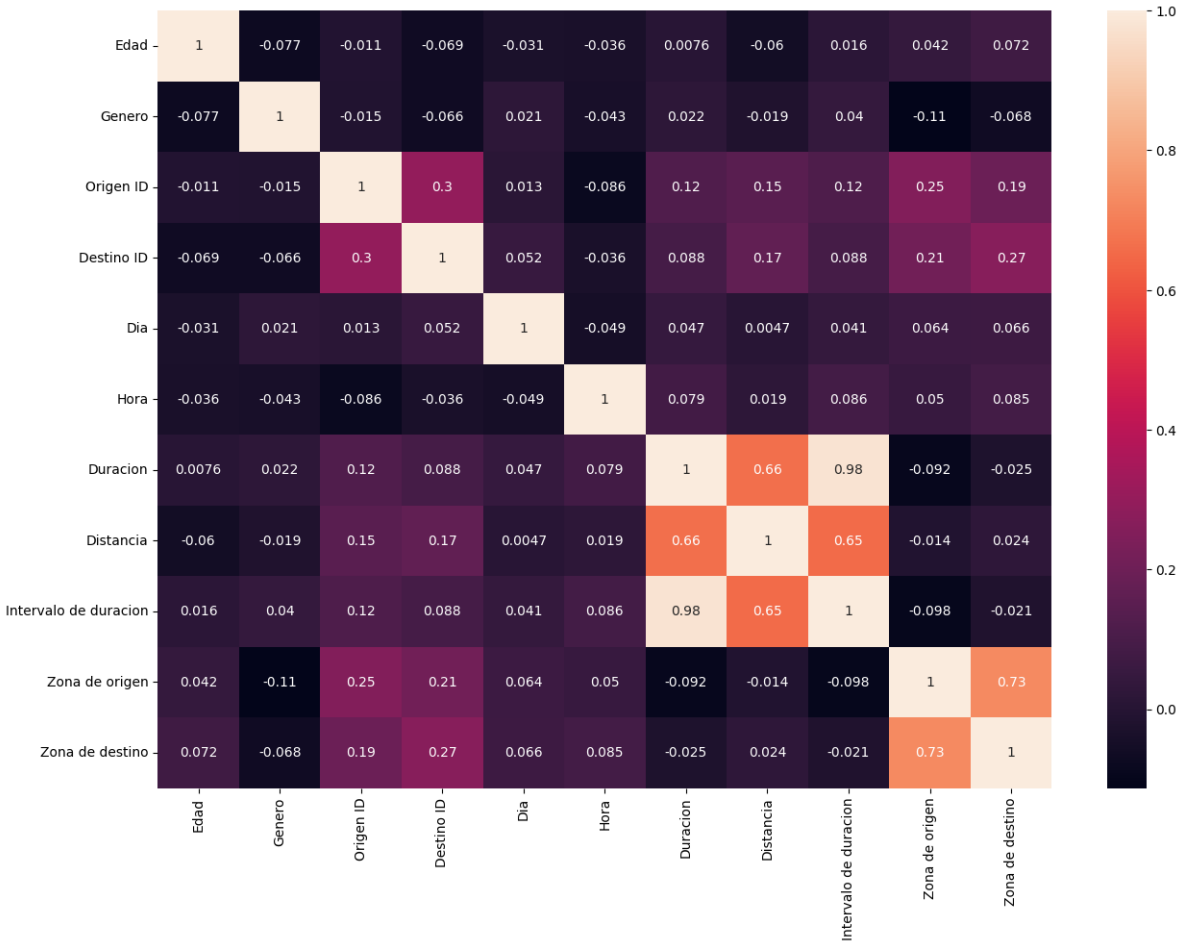
Así, obtuvimos un *dataframe* definitivo con variables continuas y categoricas con las que procederemos a realizar el análisis estadísctico. A continuación se presenta dicho *dataframe*

	Edad	Genero	Origen ID	Destino ID	Dia	Hora	Duracion	Distancia	Intervalo de duracion	Zona de origen	Zona de destino
0	36	0	11	35	1	18	6.916667	791.957123	2	1	1
1	28	0	260	272	5	19	11.383333	1802.525253	3	1	1
2	24	0	106	293	2	11	12.333333	1997.349364	3	2	2
3	53	1	4	51	4	19	8.866667	1430.854395	2	1	1
4	26	0	117	286	7	21	7.533333	1316.957751	2	2	2

Figura 2.3: Dataframe final

3. Algunas relaciones interesantes en los datos

Antes de pasar a analizar variables de interés en específico o profundizar en aspectos particulares como el poder predictivo o la influencia sobre otras variables, consideramos importante entender a grandes rasgos las relaciones que tienen las variables entre sí, al menos entre pares. Para ello, empleamos algunas herramientas de visualización vistas en la **primera parte del curso**. Podemos comenzar por la matriz de correlación, la cual es una excelente herramienta para evaluar, precisamente, la correlación entre pares de las variables que usaremos a lo largo del proyecto. La matriz obtenida para este caso es la siguiente:



En primera instancia, podemos observar que la mayoría de las variables **tienen muy poca correlación** con el resto de variables, exceptuando los casos que podemos inferir de forma lógica, tales como

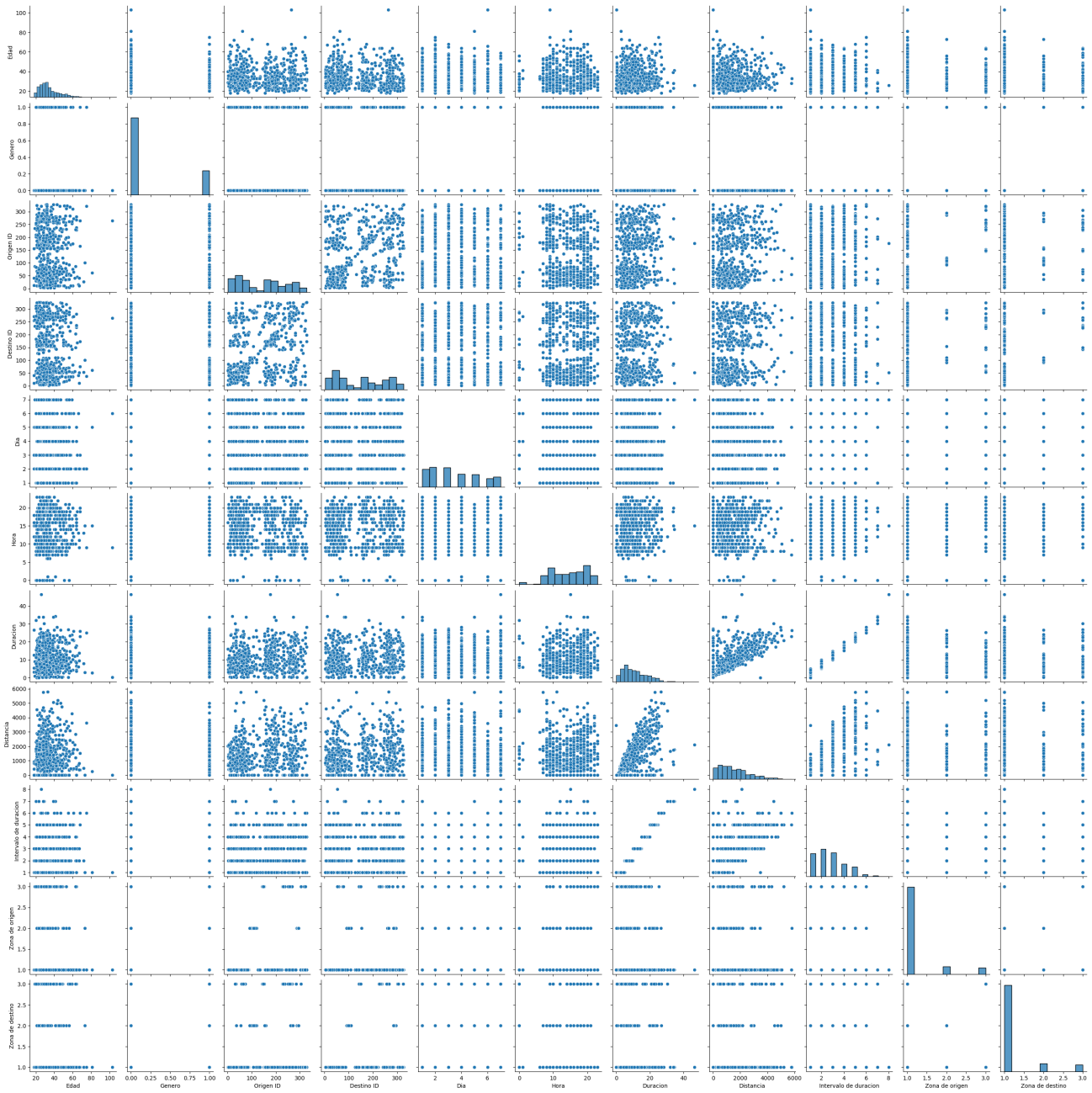
- La duración del viaje con la distancia. Es lógico que a mayor distancia recorrida, más dure el viaje, y si el viaje dura mucho, probablemente es por que se recorrió mucha distancia (aunque este ultimo caso no es tan fuerte como el primero).
- Las variables construidas a partir de otras variables. Por ejemplo, la duración con el intervalo de duración, los IDs de origen/destino con las zonas de origen/destino, etcétera.

Algo que podemos notar también y que quizás de primera mano no sea muy evidente el saber por qué, es que la zona de Origen y la zona de Destino están altamente correlacionadas ¿Es esto un indicio de que podemos predecir o describir fácilmente la zona de Destino con la zona de Origen, y viceversa? En términos generales sí, sin embargo, debemos ser cuidadosos al afirmar esto, ya que, como veremos más adelante, la gran mayoría de los viajes se realizan dentro de la misma zona (Véase Secciones [5] y [6]), es decir, tenemos

en su mayoría tuplas (zona origen - zona destino) de la forma (1 - 1), (2 - 2) o (3 - 3).

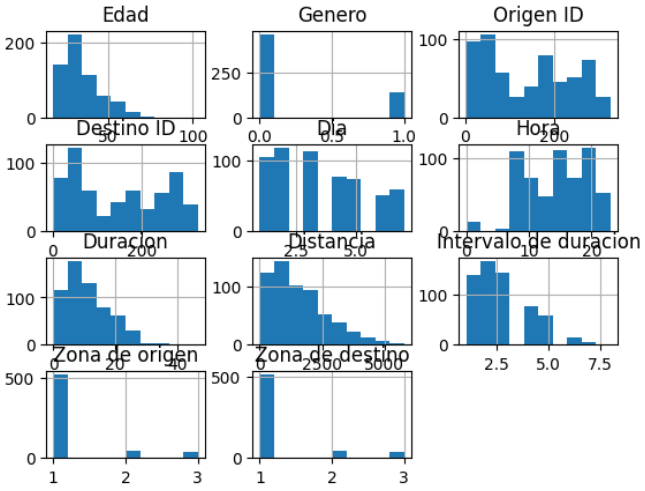
Otra cosa que hay que observar con cierto cuidado es que los IDs de origen y de destino son **variables nominales**, es decir, meramente etiquetas correspondientes a las estaciones, las cuales, cabe destacar, no siguen algún orden en específico. Esto se puede notar de mejor manera consultando el mapa de estaciones en el siguiente link: https://www.mibici.net/es/mapa/ y observando que hay varios casos de estaciones cercanas con IDs muy diferentes o lejanos entre sí.

La segunda herramienta utilizada para la visualización de las relaciones entre variables es el **pairsplot**. A través de la librería seaborn, generamos la siguiente gráfica para los datos.



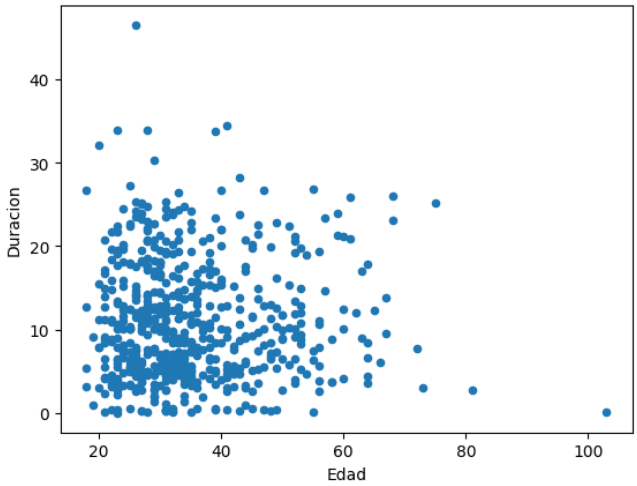
Desde luego, dado el tamaño de la imagen, se recomienda ver el pairsplot desde el notebook de python adjunto o abriendo la imagen en una pestaña nueva. No obstante, creemos importante puntualizar algunos

aspectos que podemos ver gracias a esta gráfica. El primero consiste en los histogramas de las variables, desplegados de una mejor manera en la siguiente imagen



A través de estos histogramas podemos notar la naturaleza de las variables. Por supuesto, la información correspondiente a las variables categóricas o discretas va a distribuirse solamente en ciertos tantos del histograma, por lo cual evidentemente no es posible suponer una distribución normal para estos casos, razón por la cual, más adelante, dichas variables no se incluirán en la representación de los datos por PCA. Por otra parte, con las variables continuas tampoco podemos ver de forma evidente una distribución normal, sin embargo, los histogramas si nos dan cierta información sobre los usuarios, los horarios y otras características de los viajes, aspectos que desarrollaremos en los párrafos a continuación.

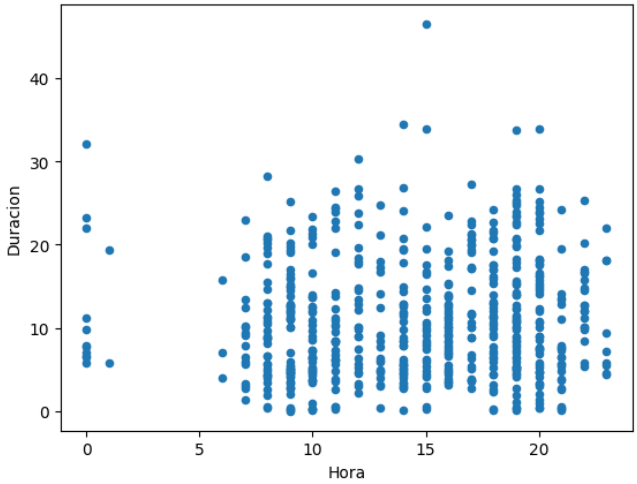
En este y los siguientes párrafos, describiremos algunas gráficas en particular contenidas en el pairsplot, las cuales consideramos nos ayudan a inferir información valiosa acerca de los datos. La primera de ellas es la de la duración del viaje con respecto a la edad:



Esta gráfica nos revela primeramente un aspecto muy importante acerca de la edad de los usuarios de la muestra, y es que **la mayoría de los usuarios se ubica entre los 20 y 40 años de edad**. Además, al comparar la edad con la duración del viaje, también podemos ver que las personas más jóvenes son las que tienen ligeramente una mayor duración de viaje, probablemente debido a las capacidades físicas de la edad, aunque la diferencia de duración con respecto a personas mayores tampoco es muy marcada. Por último, también podemos confirmar la presencia de datos atípicos, pues hay un dato que refleja más de 100 años de

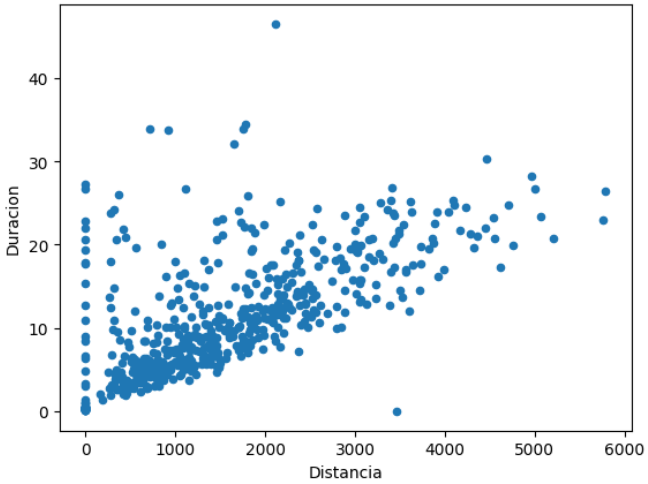
edad, lo cual evidentemente no es posible dentro del contexto (a menos que sea un caso bastante excepcional).

La siguiente gráfica que nos resultó de interés fue la de la duración con respecto a la hora del día (en su formato de 0 a 24).



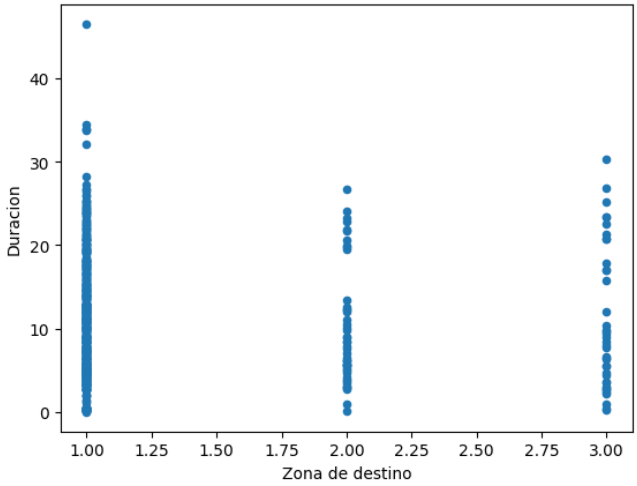
En esta gráfica vemos que la mayoría de los viajes se realizan durante el día (lo cual es de esperarse), pero también una menor pero no despreciable cantidad se hacen más tarde hasta la media noche. Notamos también que prácticamente no hay viajes durante la madrugada y que los viajes con mayor duración se hacen durante la tarde después del mediodía, aunque es de resaltar que la influencia de la hora sobre la duración no es muy marcada.

La siguiente gráfica refleja una correlación un tanto esperada pero que será objeto de interés durante las siguientes secciones, que es la de la duración respecto a la distancia del viaje.



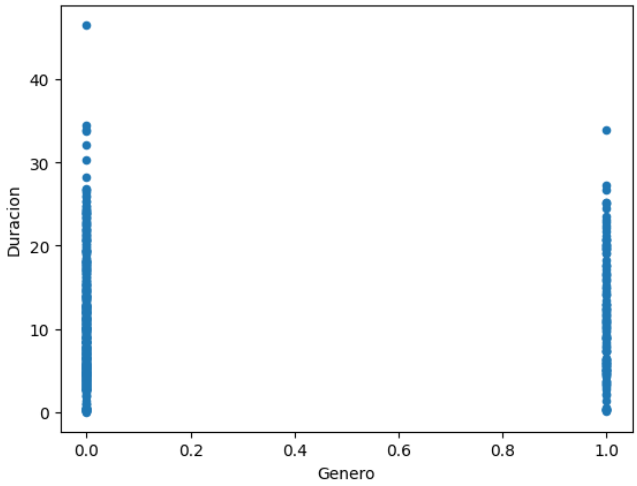
Como se mencionó antes, esta relación viene por el hecho de que a mayor distancia recorrida, más dura el viaje. No obstante, la gráfica deja en evidencia que una mayor duración del viaje no en todos los casos implica una mayor distancia recorrida, esto probablemente debido a otros factores como el tráfico, la condición física del usuario, etcétera.

Otra gráfica que refleja una relación que será retomada más adelante (Véase Sección [5] y [6]) es la de la duración con respecto a las zona de Origen del viaje.



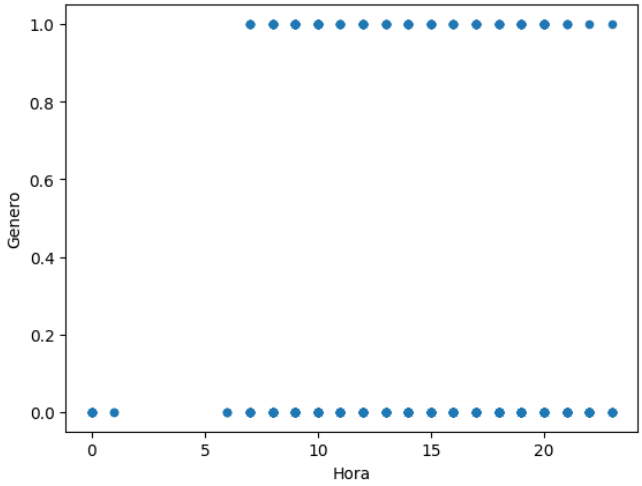
Con esta gráfica podemos darnos una idea de lo pequeñas o grandes que son las zonas de estaciones, pues para una zona con estaciones más cercanas entre sí (más pequeña) se espera que los viajes no sean tan largos y por lo tanto duren menos, cosa que podemos ver ligeramente para la zona número 2.

Ya para terminar con variables relacionadas a la duración, tenemos la gráfica de esta variable con respecto al género.

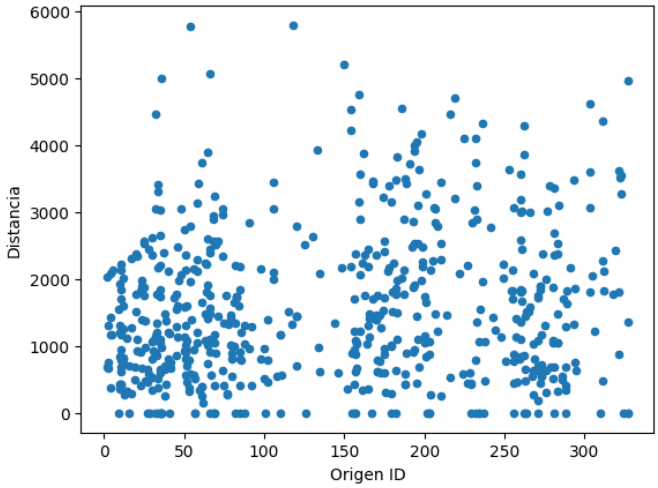


Esta gráfica deja evidente que, si bien hay un par de datos donde los hombres tienen una mayor duración del viaje que las mujeres, en la gran mayoría de los datos no se observa una diferencia significativa entre hombres y mujeres, por lo cual podríamos inferir (Véase Sección [4]) que quizás el género no tiene mucha influencia en la duración del viaje.

Algo que si podemos ver en otra gráfica es que al parecer los hombres son los únicos que realizan viajes en bici después de la medianoche, lo cual puede ser debido a la situación de inseguridad para el sexo femenino en Guadalajara, como se muestra en la siguiente imagen.



Finalmente, otra gráfica que podría resultar de interés es la siguiente



Más que para inferir una relación entre variables, nos puede dar una idea de que estaciones (por su ID) suelen ser más concurridas y desde cuales por lo general se realizan viajes más largos.

En las siguientes secciones, tomaremos algunas variables de interés y a través del uso de herramientas vistas a lo largo del curso, determinaremos cómo se relacionan con el resto de variables en conjunto, además de determinar si existe alguna influencia, relación significativa y/o poder predictivo de ciertas variables con respecto a la de interés.

4. Variable de interés: Duración del Viaje

En esta sección consideraremos como variable de interés a la **Duración del viaje** con el fin de analizar la existencia de relacionas multivariadas y examinar el poder predictivo de las demás variables con respecto a la dicha variable de interés. Para ello, proseguimos dicho análisis exploratorio haciendo PCA a nuestro conjunto de datos. Notemos que existen variables categóricas y nominales que no necesariamente tienen una distribución normal, por lo tanto, solo estaremos considerando el siguiente *dataframe* para la reducción de dimensiones.

	Edad	Dia	Hora	Distancia
0	36	1	18	791.957123
1	28	5	19	1802.525253
2	24	2	11	1997.349364
3	53	4	19	1430.854395
4	26	7	21	1316.957751

Figura 4.1: Head of Dataframe PCA

En este sentido, se realizó una prueba de componentes principales (sin la variable de interés) con los datos estandarizados. Consecuentemente, analicemos la siguiente gráfica de la varianza explicada al realizar PCA.

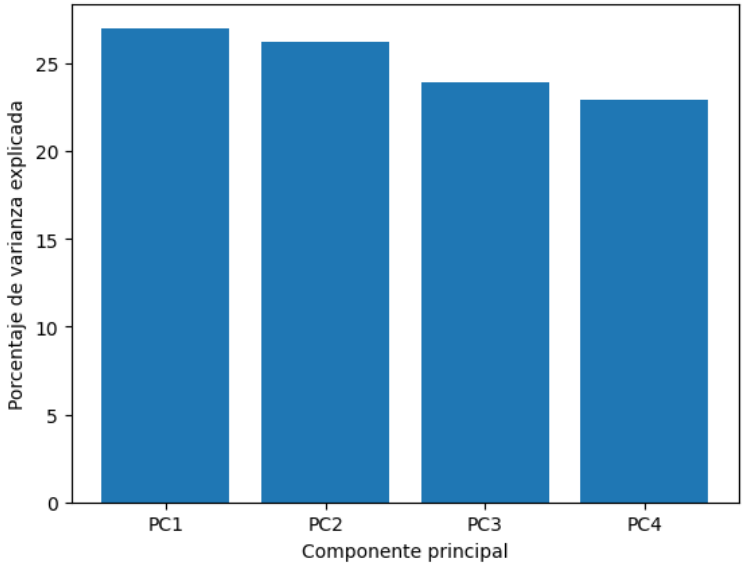


Figura 4.2: Grafica de Varianza Explicada

Viendo la Figura 4.2 podemos aplicar la regla del codo y es posible observar que en teoría se necesitan los primeros dos componentes para obtener una representación adecuada de los datos. Ahora, también notemos que para los 4 componentes obtenemos un porcentaje relativamente bajo de varianza explicada entre los componentes principales y como tal no existe un componente con mayor desviación estándar que describa correctamente los datos. Como el porcentaje en general es muy bajo, ya que el primer componente apenas supera el 25 % de varianza explicada, la visualización por PCA quizás no capture completamente la estructura de los datos, pero al menos nos puede dar una noción acerca de las posibles relaciones que existan entre variables. Consideremos entonces la siguiente gráfica de los datos representados a través de los primeros dos componentes principales.

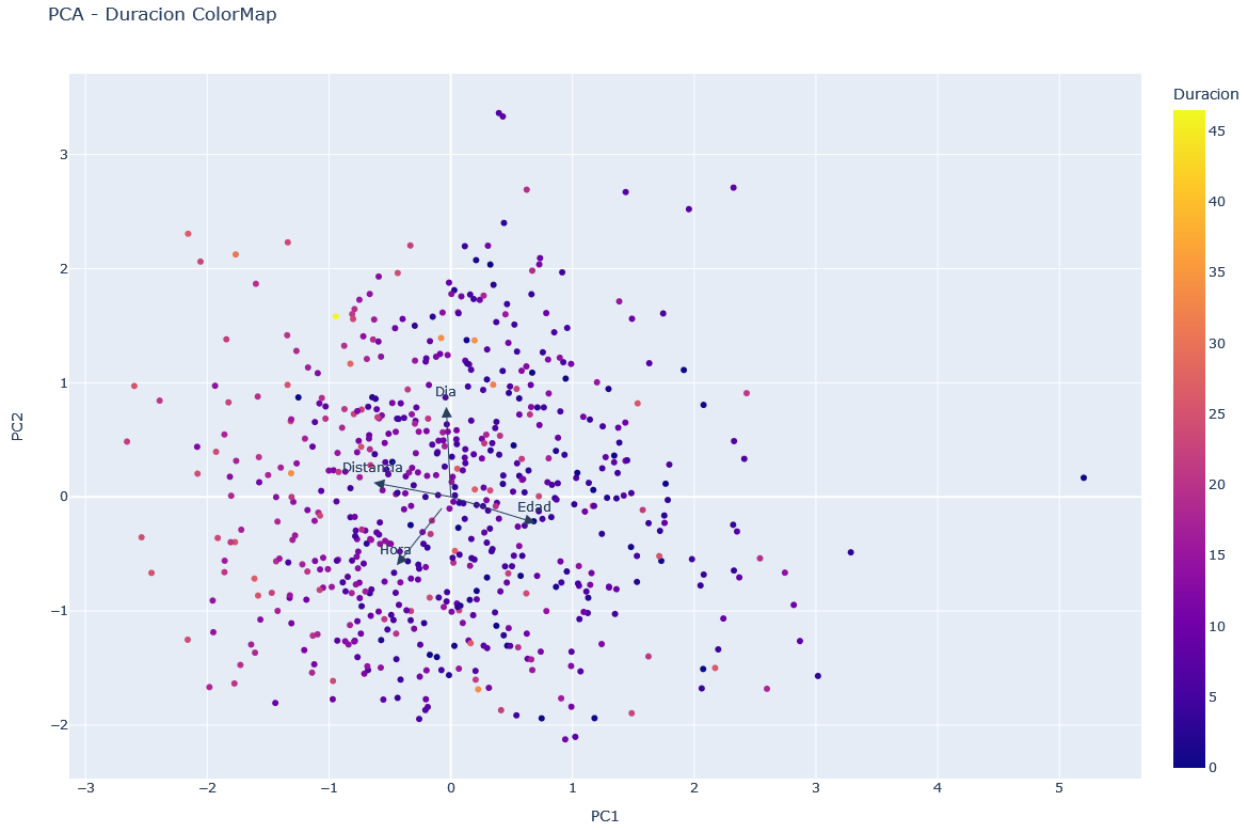


Figura 4.3: PCA - Duración de viajes

Cada punto en el gráfico representa un viaje y está coloreado según la duración de dicho viaje. Notemos que además de los puntos se graficaron las direcciones de varianza de cada característica. Es importante observar que en dicha gráfica la **Distancia** influye de gran manera ya que en la dirección de la característica se sitúan las muestras con la mayor duración el los viajes (esquina superior izquierda). Recordemos por la sección [3] que existe una fuerte correlación entre la distancia y la duración.

Un dato interesante que podemos observar es que mientras la distancia aumenta, la edad disminuye y vice versa, mientras la edad aumenta, la distancia disminuye. Lo anterior nos indica que la mayoría de los viajes con duraciones más extensas es realizada por usuarios jóvenes o de mediana edad. Ahora, una pregunta que nos podemos plantear es si dichas variables (edad, día, hora y distancia) influyen de alguna manera para realizar una distinción del genero, por lo tanto, obtuvimos el siguiente grafo con un colormap con respecto a el género.

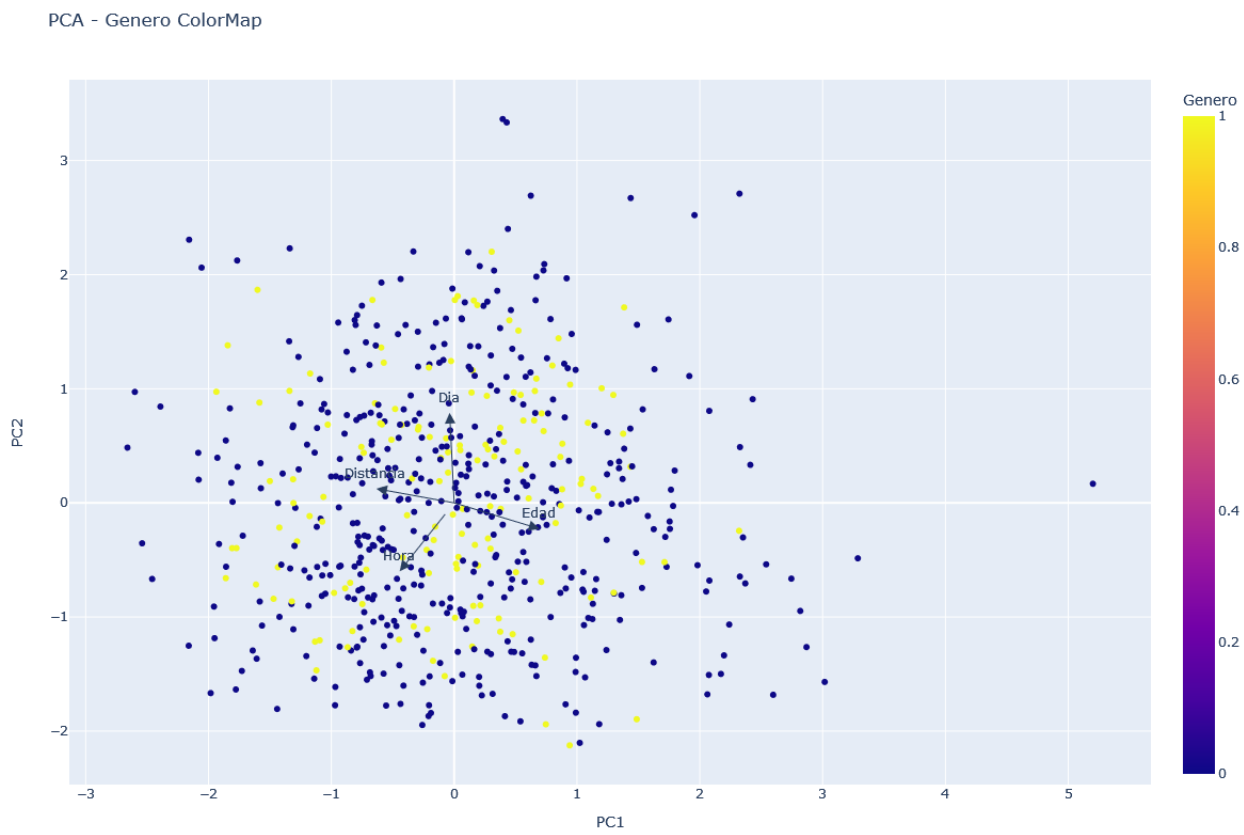


Figura 4.4: PCA - Genero

Cada punto está graficado según su genero (0 - Masculino, 1 - Femenino) y por lo visto no existe ninguna correlación entre los datos que haga distinción del sexo del usuario. Con todo lo anterior establecido, es posible formular otra pregunta, **¿La distancia tiene una gran influencia para diferenciar los datos de mayor duración?** Para poder contestar la pregunta, debemos de considerar un *dataframe* sin la variable distancia, es decir,

	Edad	Dia	Hora
0	36	1	18
1	28	5	19
2	24	2	11
3	53	4	19
4	26	7	21

Figura 4.5: Head of Dataframe without Distance

Con este nuevo *dataframe* realizaremos PCA y observaremos la influencia de la distancia para diferenciar los datos concuerdo a su duración. Para ello, primero consideremos los porcentajes de varianza explicada.

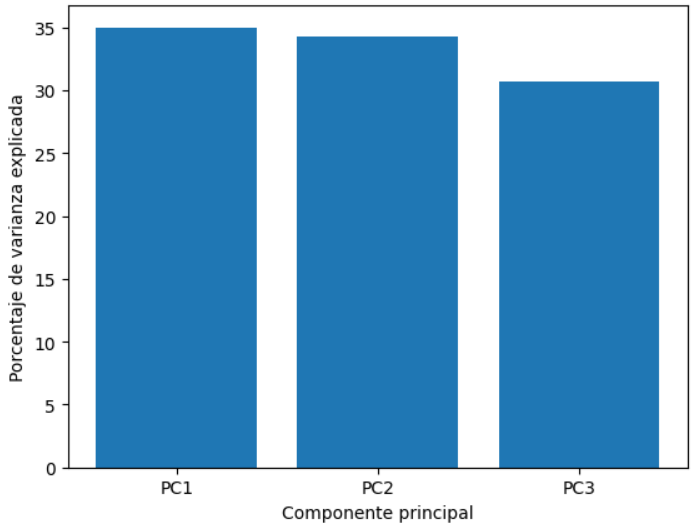


Figura 4.6: Grafica de Varianza Explicada

A comparación de la gráfica 4.2, los 3 componentes presentan un mayor porcentaje de varianza para describir los datos aunque en general sigue siendo muy bajo. Con esto en mente, veamos la siguiente gráfica de los primeros dos componentes.

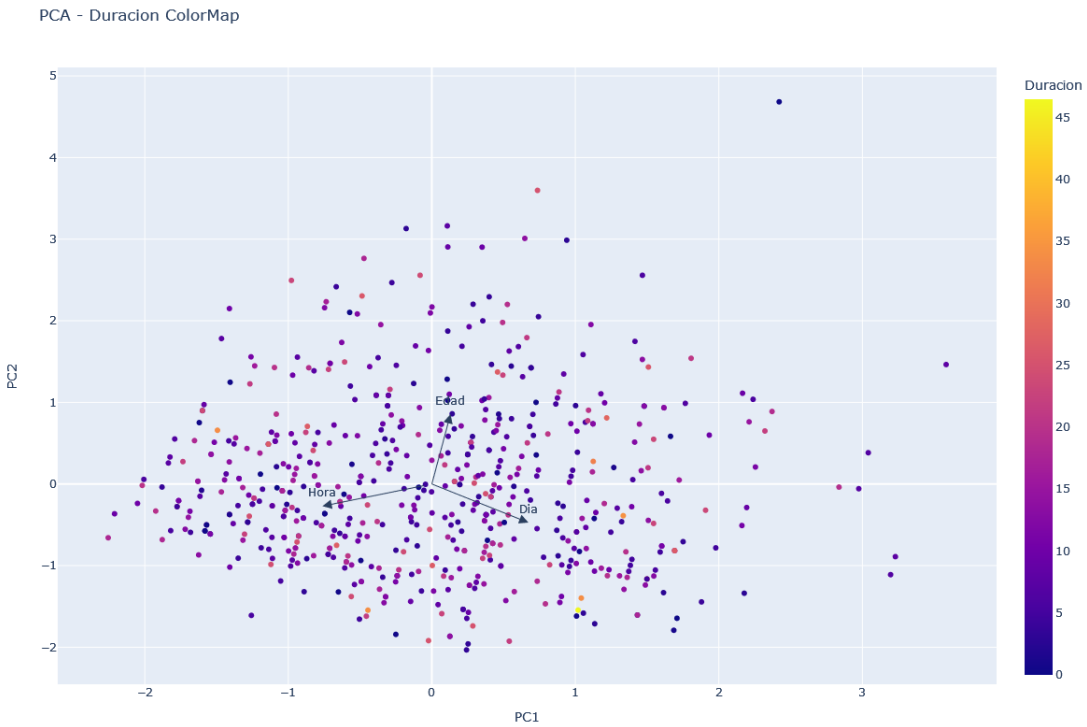


Figura 4.7: PCA sin Distancia

Notemos que obtuvimos datos dispersos para visualizar una representación de la duración de los viajes con respecto a la hora, edad y día de las muestras. Por lo tanto, al no considerar la distancia observamos que no existe una separación de las observaciones dependiendo de la duración, por lo que, no hay un patrón en

la que se distribuyen los datos. Veamos si existe alguna agrupación de los datos en función de una variable con el propósito de una posible predicción.

Primero, consideremos el uso de un *dendograma* para obtener una idea primitiva de la cantidad de clusters que se pueden generar con los datos.

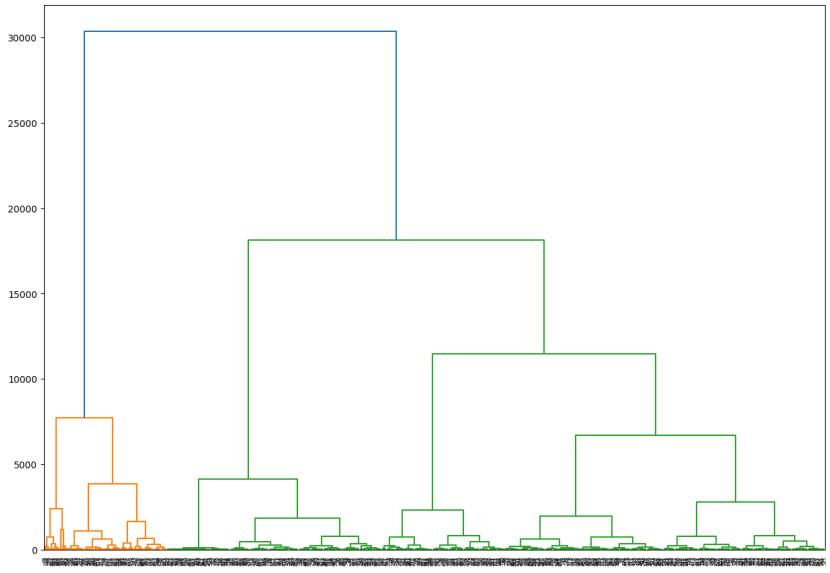


Figura 4.8: Dendograma

Notemos que se generan alrededor de 2 y 3 clusters. Para confirmar un numero exacto de clusters, procederemos a realizar k-means, porque dicho algoritmo divide nuestros datos en K grupos de tal manera que la suma de las distancias cuadradas de cada punto del conjunto de datos a su centroide más cercano sea la mínima. En este sentido, determinaremos primeramente el numero adecuado de clusters usando la regla del codo.

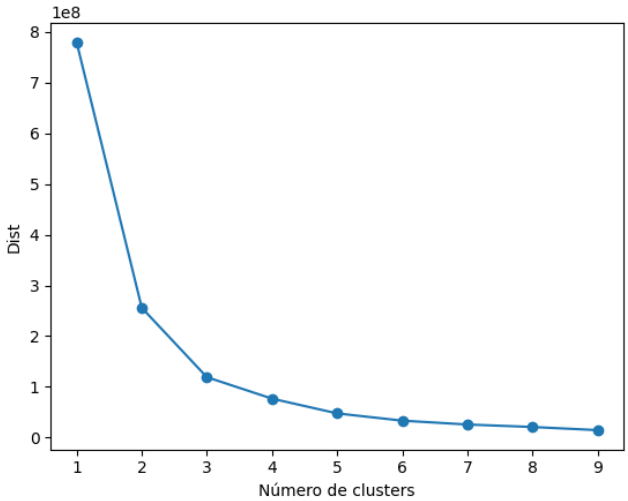


Figura 4.9: Variabilidad promedio por número de clusters en K-means

Es posible notar que a partir de 3 clusters no se produce ningún cambio significativo, por lo que procedemos a realizar k-*means* con 3 centroides.

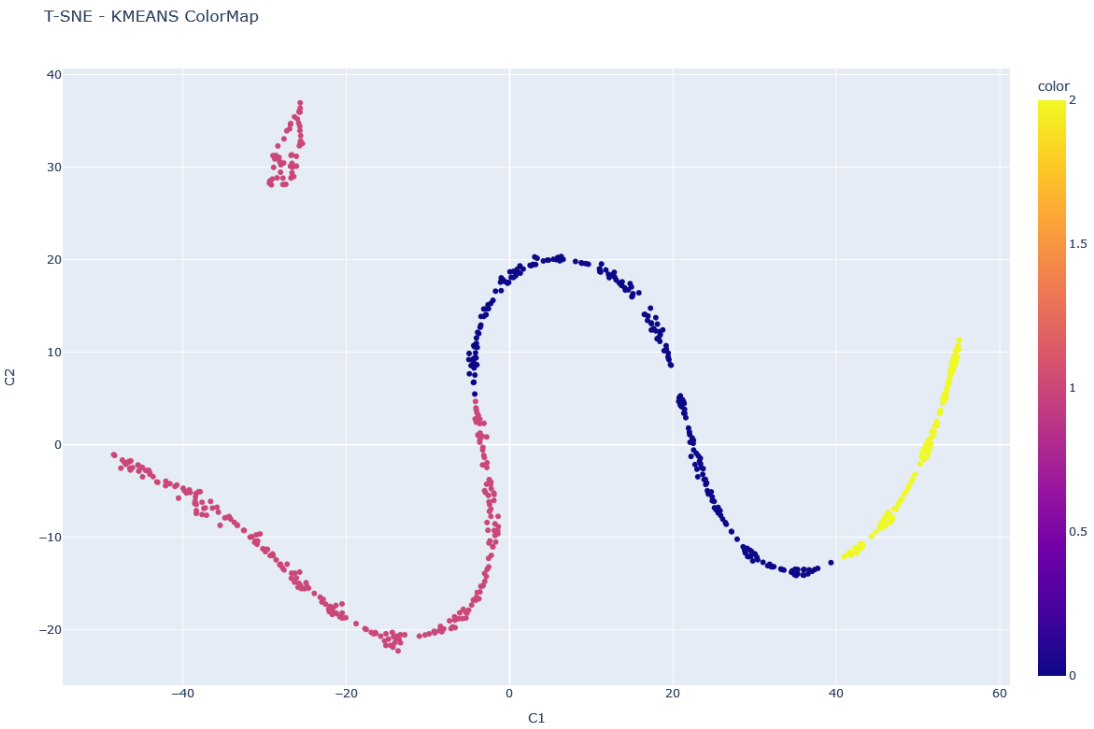
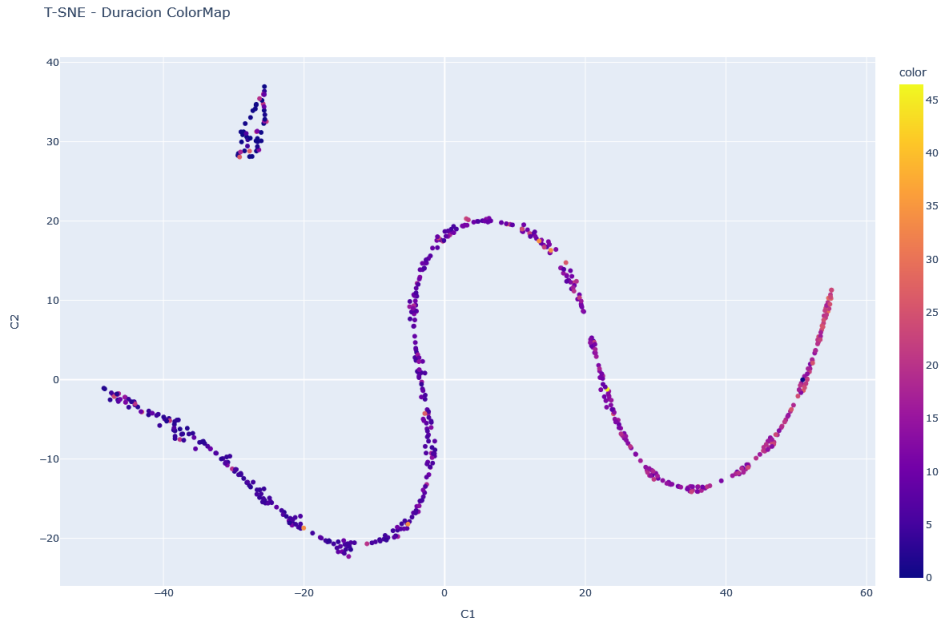


Figura 4.10: K-means con T-SNE

En esta gráfica usamos T-SNE para graficar los puntos y coloreamos con respecto a los clusters generados por k-*means*. Ahora, usando de nuevo T-SNE pero con un diferente colormap hay que verificar si las clases de k-*means* coinciden con la variable de interés, ésta siendo la duración.



En efecto, obtuvimos que existe una agrupación dependiendo de la duración del viaje. En el lado izquierdo se concentran los viajes de duración más corta, mientras que en el medio obtenemos los datos de duraciones intermedias y finalmente en el lado derecho llegamos a tener los viajes con las mayores duraciones. Cabe mencionar que para dicho agrupamiento nos consideramos a la **distancia**. Veamos cómo es que esa variable influye en la agrupación anterior.

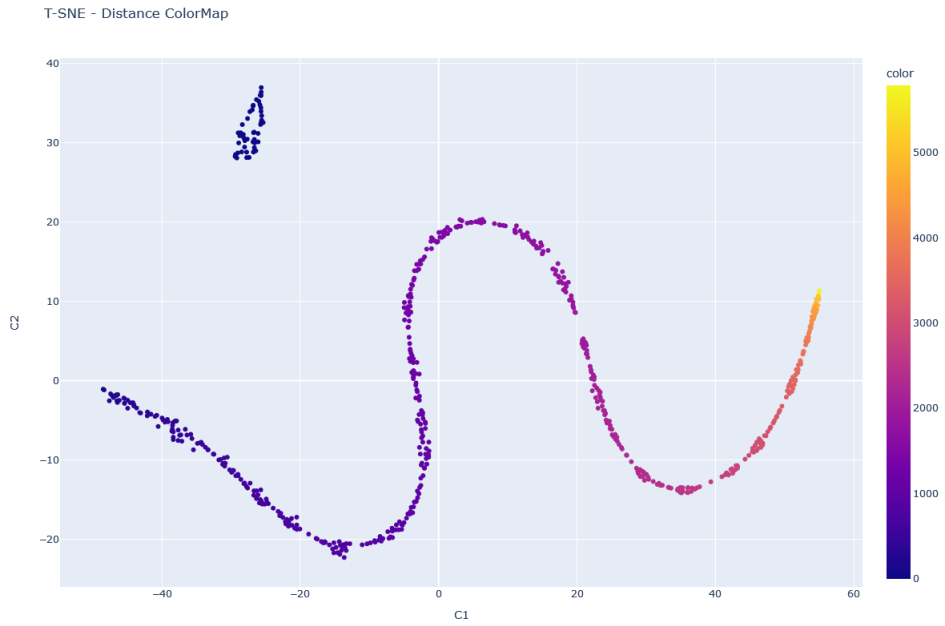


Figura 4.11: T-SNE con Distancia Colormap

Notemos que realizamos el colormap con respecto a la distancia recorrida en cada viaje y es claro que existe una relación entre la agrupación por k-*means* y la **distancia**. Entonces, es posible que dicha variable influya de gran de manera en los resultados obtenidos y esté generando un sesgo a la hora de analizar los datos. Con esto en mente, nos podemos preguntar, **¿cómo cambia la agrupación si quitamos la distancia y sigue existiendo una relación con la duración?**

Entonces, realizaremos el procedimiento anterior pero con un *dataframe* sin la variable **distancia**.

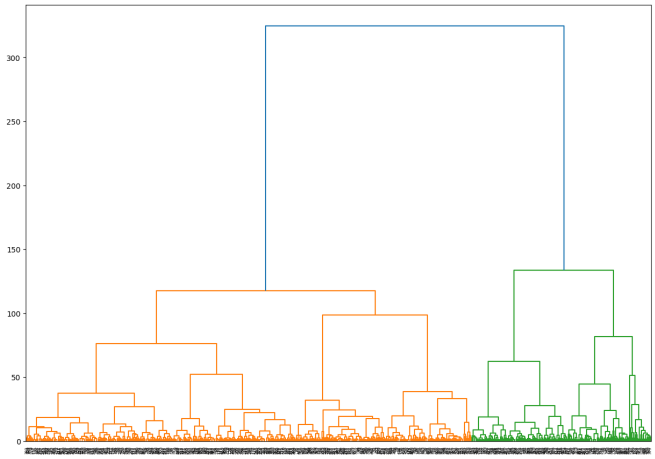


Figura 4.12: Dendograma sin distancia

Con la figura anterior podemos observar que se generan dos clusters. Para confirmar lo anterior, de nuevo haremos uso de la regla del codo. Veamos que

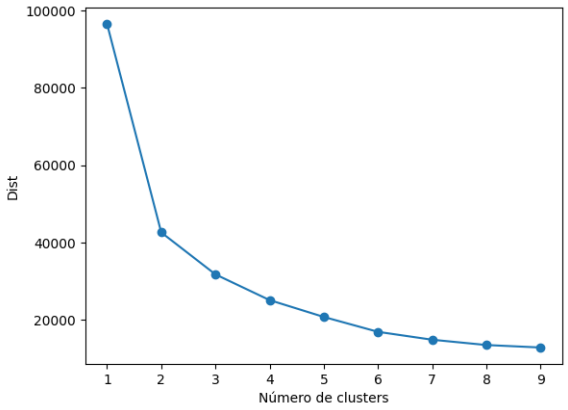


Figura 4.13: Regla codo sin distancia

Notemos que con la regla del codo podemos inferir que los datos se dividen claramente en 2 clusters. Con esto en mente, de nuevo realicemos k-*means* y T-SNE para verificar si existe una relación de agrupación entre las demás variables y la variable de interés.

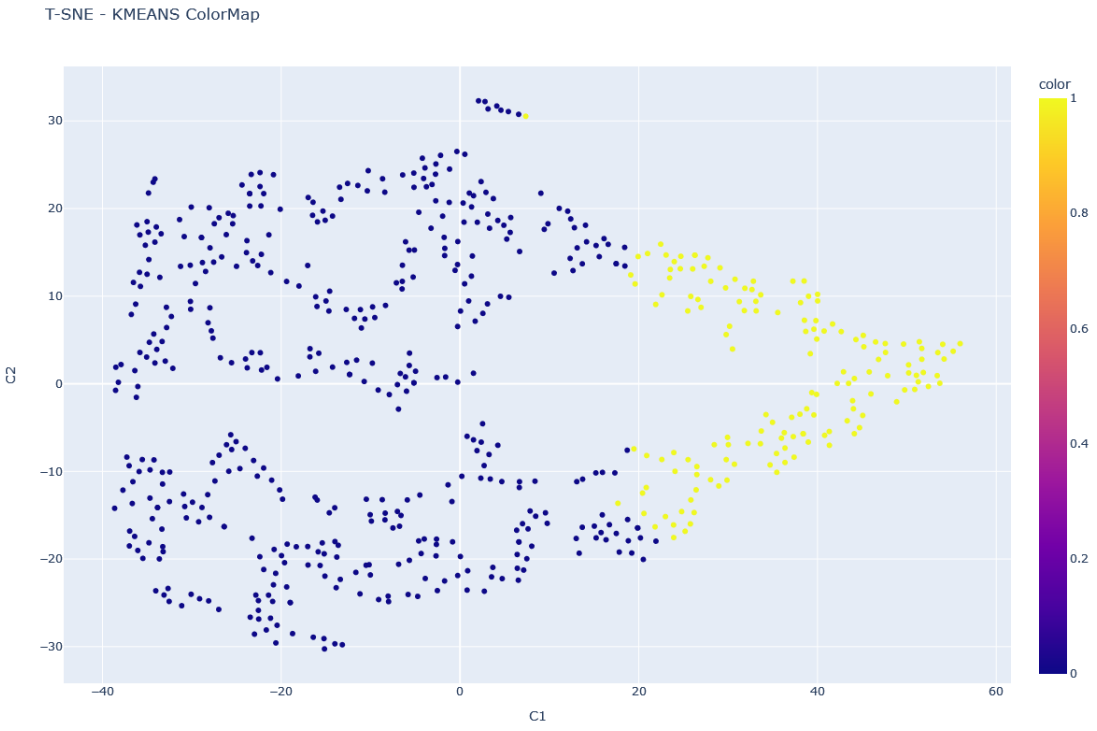


Figura 4.14: T-SNE con k-means colormap

Con la figura anterior, consideremos un colormap en función de la duración y veamos si existen indicios de una agrupación con respecto a dicha variable.

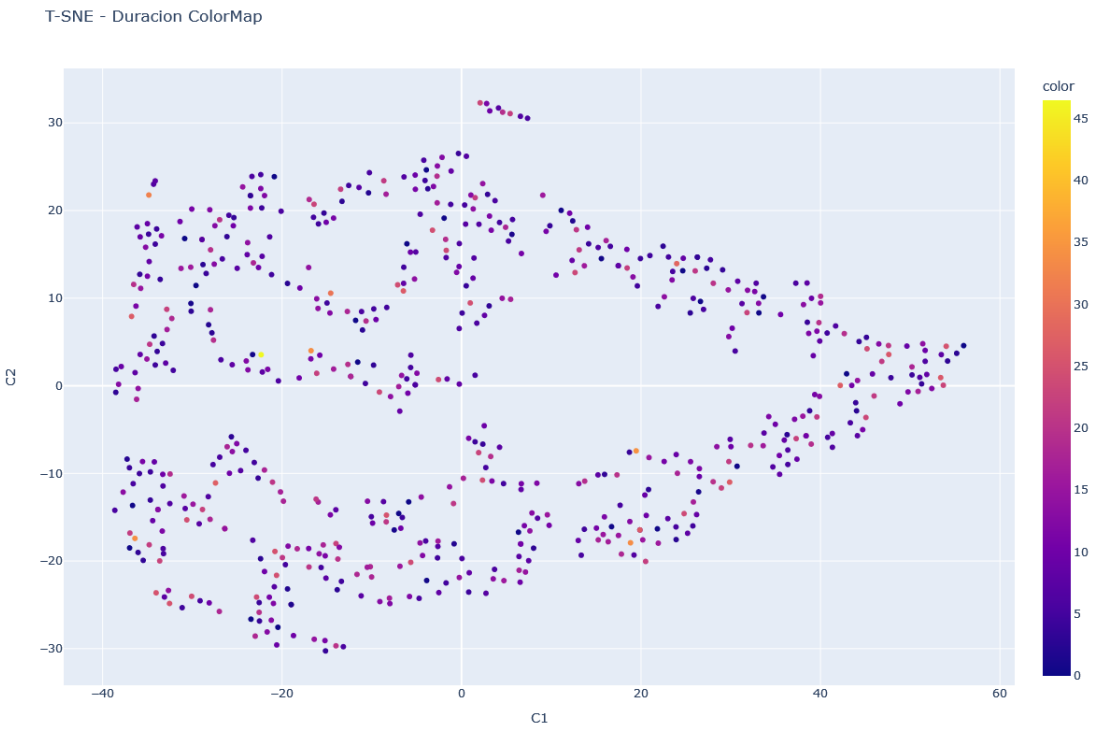


Figura 4.15: T-SNE con duración colormap

Comparando ambos figuras es posible concluir que los grupos determinados por k-*means* no coinciden con el colormap en función de la duración, por lo que no existe una agrupación de la muestra con respecto a la duración. Esto es un posible indicador de que el poder predictivo de las demás variables es muy débil para predecir la duración del viaje al no considerar la **distancia** en el *dataframe*.

Al haber realizado PCA e implementado los métodos de agrupación procederemos a verificar el poder predictivo usando modelos de clasificación vistos en el curso. Consideraremos dos predictores, **Árboles de Decisión** y **Clasificador Bayesiano Óptimo**.

- **Árboles de Decisión**

  En este clasificador se usó el siguiente *dataframe*

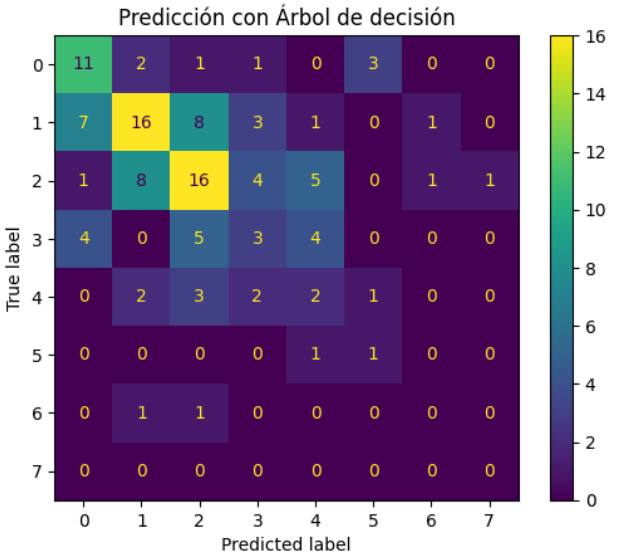
	Edad	Genero	Dia	Hora	Distancia	Zona de origen	Zona de destino
0	36	0	1	18	791.957123	1	1
1	28	0	5	19	1802.525253	1	1
2	24	0	2	11	1997.349364	2	2
3	53	1	4	19	1430.854395	1	1
4	26	0	7	21	1316.957751	2	2

Figura 4.16

Notemos que no se incluyen los IDs de las estaciones de origen y destino ya que son simplemente etiquetas y no tiene influencia predictiva, son meramente nominales. Ahora, también quitamos la variable **duración** e **Intervalos de Duración**. Vamos a tratar de predecir los intervalos de duración. A continuación se presentan los resultados obtenidos

	precision	recall	f1-score	support
1	0.48	0.61	0.54	18
2	0.55	0.44	0.49	36
3	0.47	0.44	0.46	36
4	0.23	0.19	0.21	16
5	0.15	0.20	0.17	10
6	0.20	0.50	0.29	2
7	0.00	0.00	0.00	2
8	0.00	0.00	0.00	0
accuracy			0.41	120
macro avg	0.26	0.30	0.27	120
weighted avg	0.43	0.41	0.41	120

(a) Reporte



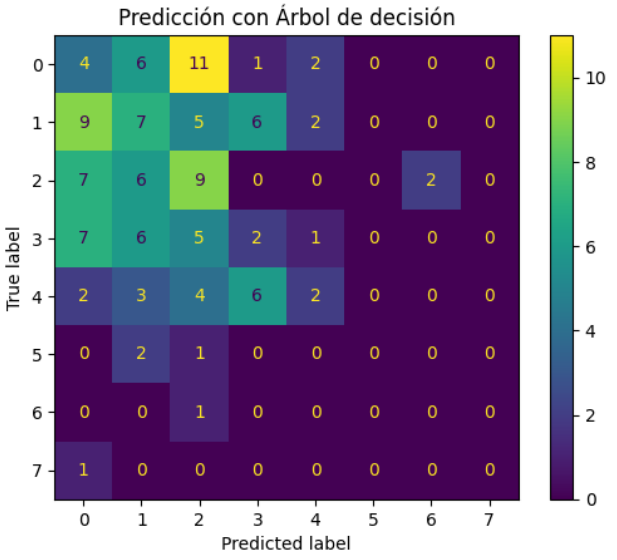
(b) Matriz

Figura 4.17: Resultados de Predicción

Obtuvimos una exactitud (accuracy) del 41 % lo cual nos indica que nuestro modelo de predicción tiene errores grandes. Si nos enfocamos en la matriz de confusión es posible notar que el modelo predijo mejor las clases 1 y 2, las cuales corresponden a (5 - 10) min y (10-15) min respectivamente. Esto nos indica que la mayoría de los usuarios tienen una duración promedio entre 5 y 15 min. Así, podemos concluir que el poder predictivo de las variables edad, genero, distancia, hora, origen y destino es muy bajo para predecir la duración del usuario. Con esto en mente, veamos que si no consideramos la **distancia** obtendremos una menor exactitud de predicción. Por ende, llegamos a los siguientes resultados

	precision	recall	f1-score	support
1	0.13	0.17	0.15	24
2	0.23	0.24	0.24	29
3	0.25	0.38	0.30	24
4	0.13	0.10	0.11	21
5	0.29	0.12	0.17	17
6	0.00	0.00	0.00	3
7	0.00	0.00	0.00	1
8	0.00	0.00	0.00	1
accuracy			0.20	120
macro avg	0.13	0.12	0.12	120
weighted avg	0.20	0.20	0.19	120

(a) Reporte



(b) Matriz

Figura 4.18: Resultados de Predicción

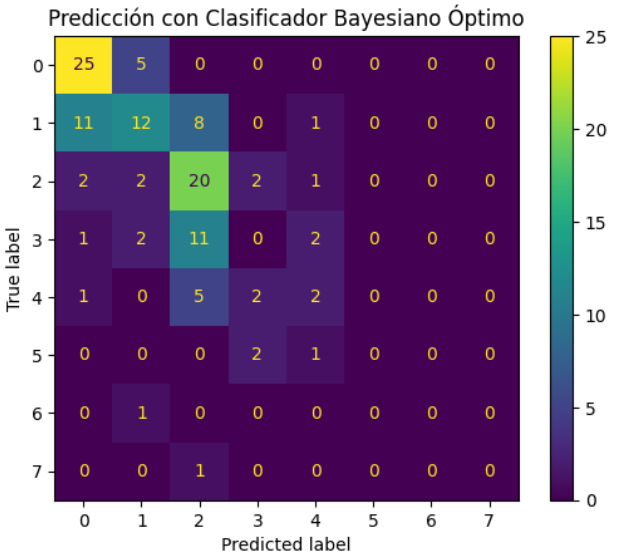
Observando la Figura 4.18 vemos que al no considerar la distancia obtenemos una exactitud del 20 %, es decir, se redujo a la mitad nuestro poder predicitvo. Lo anterior nos indica que en realidad la única variable que tiene influencia a la hora de predecir la duración del viaje es la distancia recorrida por el usuario. Veamos qué ocurre cuando consideramos otro clasificador.

- **Clasificador Bayesiano Óptimo**

  Considerando el mismo *dataframe* de la Figura 4.16 obtuvimos los siguientes resultados para predecir los intervalos de duración con el modelo Clasificador Bayesiano Óptimo.

	precision	recall	f1-score	support
1	0.62	0.83	0.71	30
2	0.55	0.38	0.44	32
3	0.44	0.74	0.56	27
4	0.00	0.00	0.00	16
5	0.29	0.20	0.24	10
6	0.00	0.00	0.00	3
7	0.00	0.00	0.00	1
8	0.00	0.00	0.00	1
accuracy			0.49	120
macro avg	0.24	0.27	0.24	120
weighted avg	0.43	0.49	0.44	120

(a) Reporte



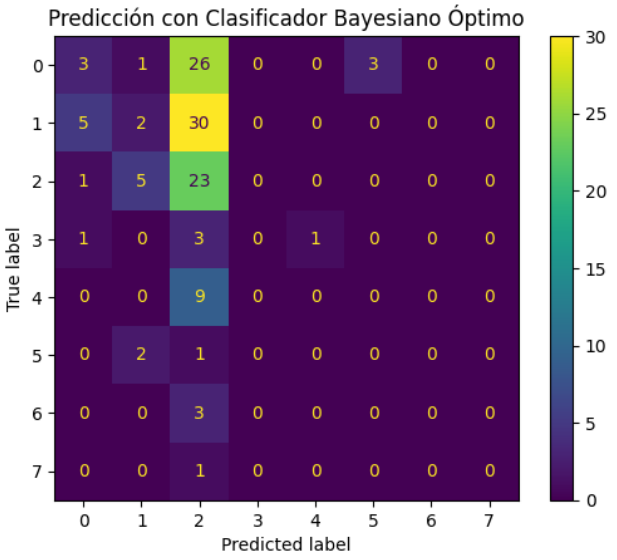
(b) Matriz

Figura 4.19: Resultados de Predicción

Comparando éstos resultados con los de la Figura 4.17 sí obtuvimos una mejoría pero fue mínima. En este caso llegamos a una exactitud del 49 % y los intervalos que tuvieron la mayor cantidad de predicciones correctas fueron el 0 y el 2. Esto confirma de nuevo que la mayoría de los usuarios duran alrededor de 1 a 15 minutos en la bicicleta. De manera análoga, consideremos el *dataframe* sin la **distancia** y veamos los cambios en el modelo predictivo.

	precision	recall	f1-score	support
1	0.30	0.09	0.14	33
2	0.20	0.05	0.09	37
3	0.24	0.79	0.37	29
4	0.00	0.00	0.00	5
5	0.00	0.00	0.00	9
6	0.00	0.00	0.00	3
7	0.00	0.00	0.00	3
8	0.00	0.00	0.00	1
accuracy			0.23	120
macro avg	0.09	0.12	0.07	120
weighted avg	0.20	0.23	0.15	120

(a) Reporte



(b) Matriz

Figura 4.20: Resultados de Predicción

Notemos que de nuevo la exactitud se redujo a un 23 %, con muy mínima diferencia de los resultados de la Figura 4.18. Un efecto interesante es que si no consideramos a la distancia el modelo predice en su mayoría por la etiqueta 2. Con esto establecido, se cumple que en efecto el poder predictivo de las variables es muy bajo para predecir la duración y todavía disminuye si no consideramos a la distancia.

5. Variable de interés: Zona de origen

Para esta sección y la siguiente, tomamos como variables de interés las zonas involucradas en el viaje. Desde luego, dada la distribución de los datos y la gran cantidad de IDs de estaciones que hay (más de 200), resulta prácticamente imposible determinar con exactitud el ID de la estación de origen y/o de destino si tenemos disponibles el resto de información en las otras variables, sin embargo, una forma mucho más práctica de resolver este problema y que, como veremos, arroja buenos resultados, es enfocarnos en las zonas donde se encuentran las estaciones del viaje. Por dicha razón, es que se categorizaron las estaciones en las zonas a partir de la nomenclatura provista en el sitio de miBici, como se explicó en la Sección [2]

De esta manera, en esta sección exploraremos los datos para ver si existe alguna relación multivariada, poder predictivo o influencia de otras variables con respecto a la **Zona de Origen**. De forma similar a la sección anterior, comenzaremos el análisis exploratorio con una visualización de los datos a través de PCA. Para ello, las variables consideradas para este caso, considerando que se omiten aquellas con las que definitivamente no podemos suponer una distribución normal, son las siguientes.

	Edad	Dia	Hora	Duracion	Distancia
0	36	1	18	6.916667	791.957123
1	28	5	19	11.383333	1802.525253
2	24	2	11	12.333333	1997.349364
3	53	4	19	8.866667	1430.854395
4	26	7	21	7.533333	1316.957751

Figura 5.1: Head of dataframe for PCA

Con esto establecido, veamos los porcentajes de varianza explicada que obtuvimos para cada uno de los componentes, los cuales se despliegan a continuación.

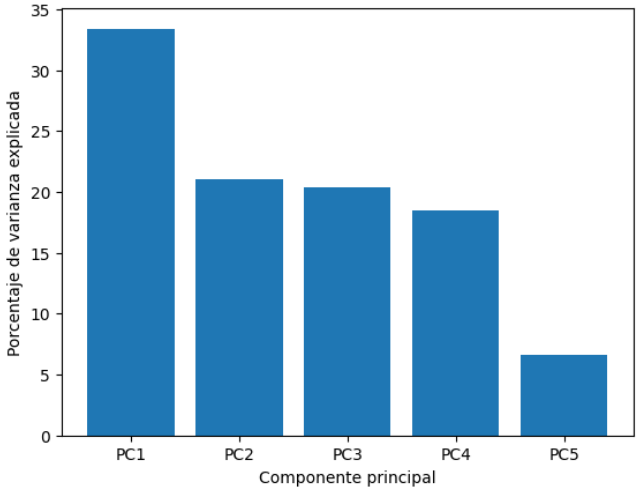


Figura 5.2: Gráfica de varianza explicada

Para este caso, se decidió utilizar los primeros dos componentes. Cabe destacar que aún así el porcentaje de varianza total explicada apenas supera el 50 %, por lo que tampoco se espera que la visualización revele completamente la estructura de los datos, sin embargo, nuevamente nos puede dar una noción de las relaciones entre las variables consideradas para la visualización y la de interés, como se puede observar en la siguiente imagen al notar las direcciones de los vectores correspondientes a las variables.

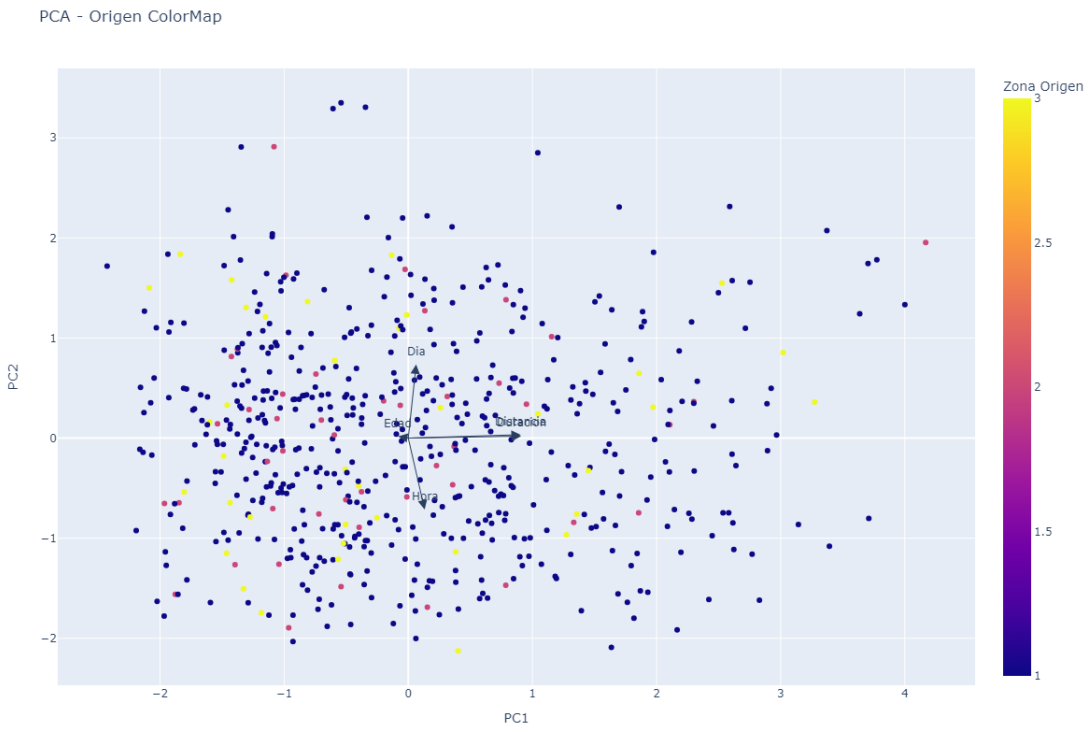


Figura 5.3: PCA - Colormap por zona de Origen

A través del colormap por la zona de origen podemos identificar varias cosas relevantes. La primera y probablemente la más evidente es que la zona con más viajes, es decir, la más concurrida es la Zona 1 - POLÍGONO CENTRAL. Otro aspecto relevante que podemos notar es que gran parte de los puntos rojos correspondientes a la Zona 2, se encuentran del lado opuesto a donde apuntan las flechas de Duración y de Distancia, lo cual nos indica que la Zona 2 - ZAPOPAN CENTRO es probablemente la zona más compacta o cuyas estaciones no son tan lejanas entre sí, esto ya que la mayoría de los viajes se realizan dentro de la misma zona.

Además de lo anterior, no podemos inferir mucho más respecto a la relación de otras variables con la zona de Origen, pues vemos que los puntos correspondientes a las Zonas 1 y 3 se encuentran bastante dispersos a lo largo de la gráfica.

Continuando con el análisis exploratorio, consideramos utilizar métodos de agrupamiento con el fin de ver si podemos identificar la presencia de clústers en los datos relacionados a la variable de interés. Para este caso, ya podremos considerar más variables, como se muestra a continuación.

	Edad	Genero	Dia	Hora	Duracion	Distancia	Intervalo de duracion	Zona de destino
0	36	0	1	18	6.916667	791.957123	2	1
1	28	0	5	19	11.383333	1802.525253	3	1
2	24	0	2	11	12.333333	1997.349364	3	2
3	53	1	4	19	8.866667	1430.854395	2	1
4	26	0	7	21	7.533333	1316.957751	2	2

Figura 5.4: Head of dataframe for clustering

Al igual que en la sección anterior, primeramente se optó por generar primero un dendograma y así tener una noción de cuántos clústers podrían formarse con los datos actuales.

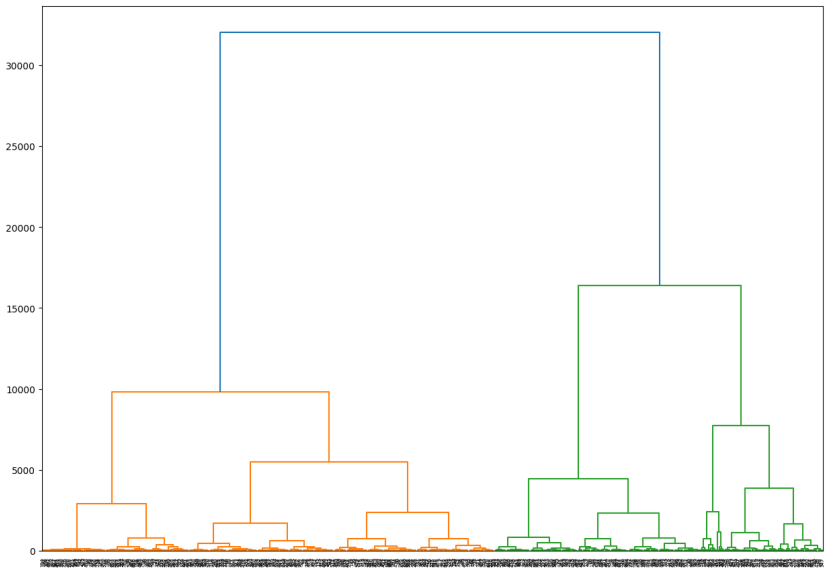


Figura 5.5: Dendograma

También consideramos graficar la variabilidad promedio obtenida por el algoritmo K-means variando sobre distintos números de clústers, y con ello, seguir la *regla del codo*.

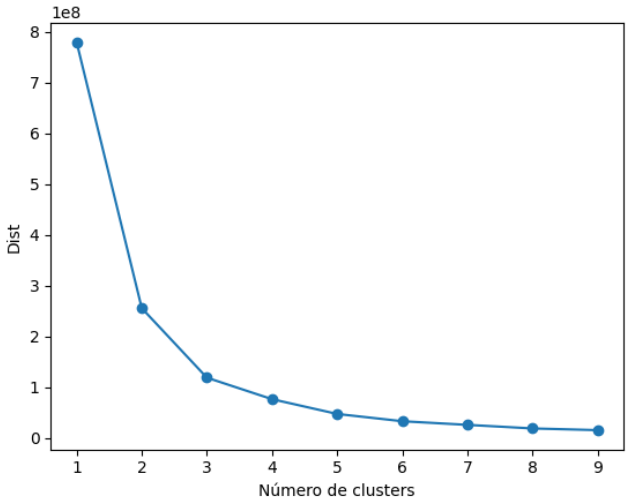


Figura 5.6: Variabilidad promedio por número de clusters en K-means

Notemos que para ambos casos se pueden inferir de 2 a 3 clusters, por lo tanto, se aplicó el algoritmo K-means considerando $k = 3$ y se obtuvo la siguiente visualización utilizando el método T-SNE.

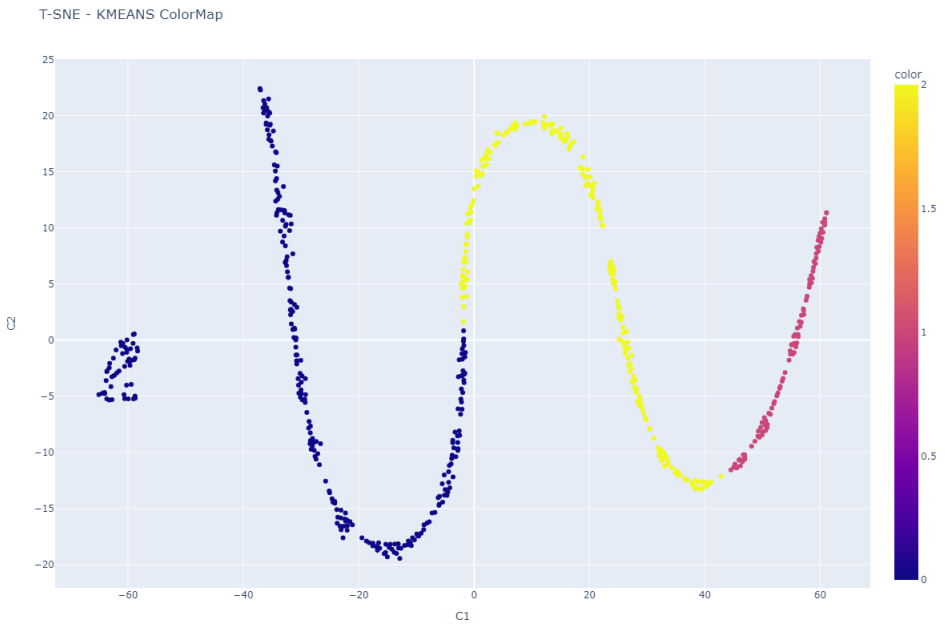


Figura 5.7: K-means con T-SNE

Vemos que efectivamente se visualizan los 3 clústers, sin embargo, cabe preguntarse ¿Estos grupos tendrán alguna relación con nuestra variable de interés? Desafortunadamente, no es el caso, pues al considerar la misma gráfica haciendo un ColorMap por zona de Origen, se obtiene lo siguiente

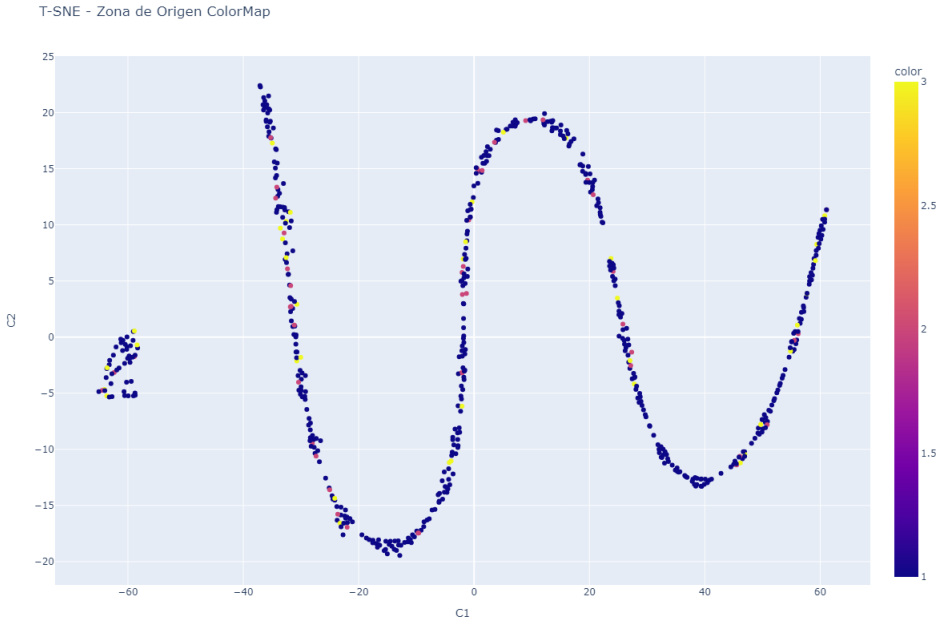


Figura 5.8: T-SNE con colormap de Zona de Origen

Claramente no se observa coincidencia en las gráficas, por lo que esto nos puede indicar que los datos no se agrupan siguiendo la zona de Origen (variable de interés) bajo este contexto y con las variables consideradas

previamente.

Por último, resta ver si con las variables consideradas nos es posible predecir la zona de Origen del viaje, es decir, veremos si dichas variables tienen cierto poder predictivo sobre la variable de interés. El dataframe utilizado para este caso es el siguiente

	Edad	Genero	Dia	Hora	Duracion	Distancia	Intervalo de duracion	Zona de destino
0	36	0	1	18	6.916667	791.957123	2	1
1	28	0	5	19	11.383333	1802.525253	3	1
2	24	0	2	11	12.333333	1997.349364	3	2
3	53	1	4	19	8.866667	1430.854395	2	1
4	26	0	7	21	7.533333	1316.957751	2	2

Figura 5.9: Head of dataframe for prediction

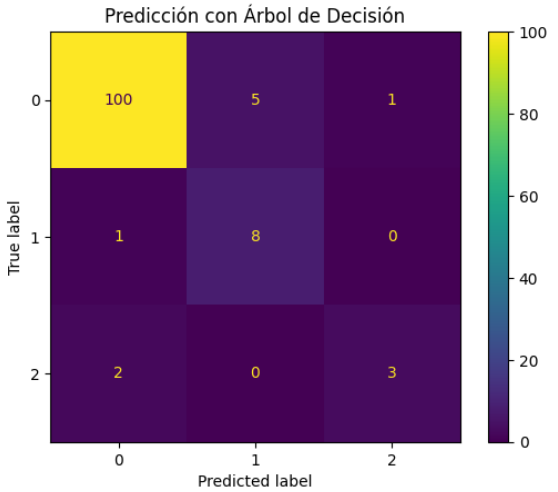
Nuevamente, utilizamos dos métodos de clasificación/predicción.

- **Árboles de decisión**

  Los resultados obtenidos para este caso son los siguientes.

	precision	recall	f1-score	support
1	0.97	0.94	0.96	106
2	0.62	0.89	0.73	9
3	0.75	0.60	0.67	5
accuracy			0.93	120
macro avg	0.78	0.81	0.78	120
weighted avg	0.94	0.93	0.93	120

(a) Reporte



(b) Matriz

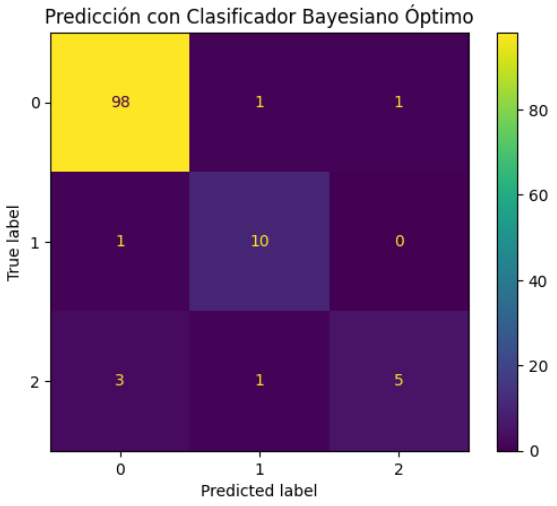
Figura 5.10: Resultados de Predicción

Vemos que los resultados son bastante buenos, pues se obtuvo una precisión del 93 % y muy pocos errores. Nótese que la mayoría de las predicciones fueron sobre la zona 1 (índice 0), nuevamente evidenciando la gran cantidad de viajes realizados en esa zona y además dándonos a entender que quizás la exactitud en la predicción recae en que en la gran mayoría de los casos, la zona de Origen y la zona de Destino coinciden, confirmando el hecho de que la grande correlación entre estas dos variables se debe al hecho de que las zonas son muy grandes y por lo tanto una gran cantidad de viajes se hace dentro de la misma zona. Además, podemos concluir que la variable con más poder predictivo sobre la zona de Origen es justamente la zona de Destino.

- **Clasificador Bayesiano Óptimo** Los resultados para este caso fueron bastante similares al caso anterior.

	precision	recall	f1-score	support
1	0.96	0.98	0.97	100
2	0.83	0.91	0.87	11
3	0.83	0.56	0.67	9
accuracy			0.94	120
macro avg	0.88	0.81	0.84	120
weighted avg	0.94	0.94	0.94	120

(a) Reporte



(b) Matriz

Figura 5.11: Resultados de Predicción

Las conclusiones a las que llegamos son prácticamente las mismas, solo destacando el hecho de que la precisión alcanzada en este caso es del 94 %, ligeramente mayor al anterior método de predicción, aunque la diferencia tampoco es muy significativa.

6. Variable de interés: Zona de destino

La forma de proceder en esta sección fue prácticamente análoga a la de la sección anterior, dada la similitud que existe entre las variables de interés. Por dicha razón es que para visualización con PCA se consideró el mismo dataframe de la Figura 5.1 y por lo tanto, los porcentajes de varianza explicada por el método coinciden en ambos casos. La diferencia en este caso es que el ColorMap en la visualización del PCA se hizo esta vez para la variable zona de Destino, obteniendo la siguiente gráfica

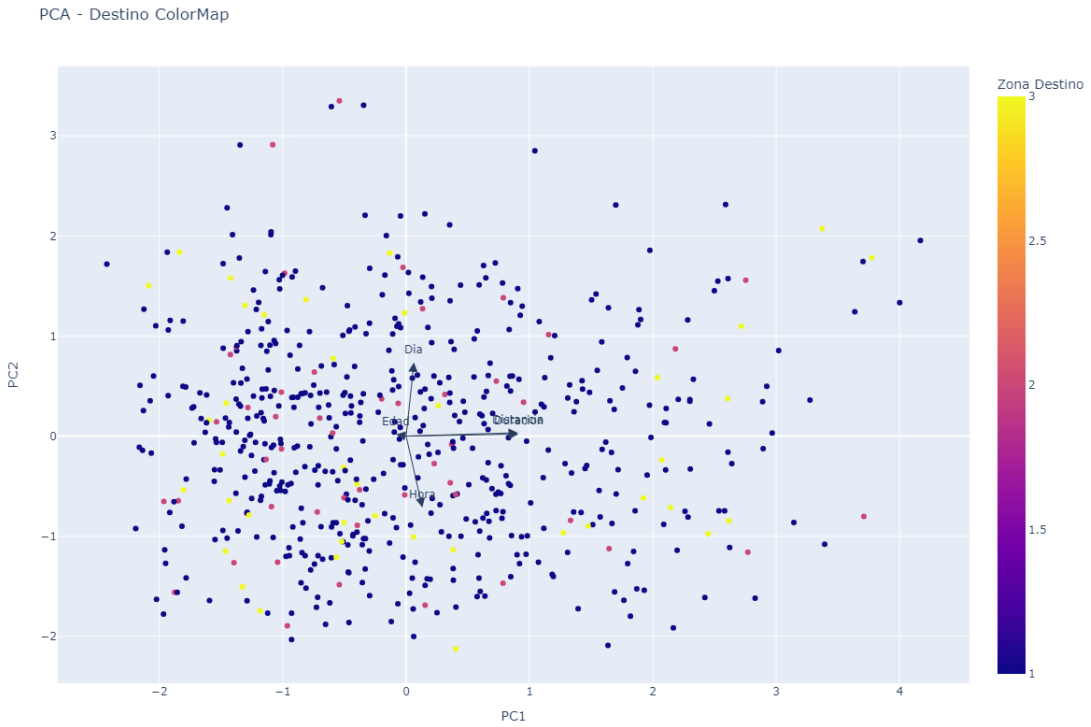


Figura 6.1: PCA - Colormap por zona de Destino

Nuevamente, podemos notar que los puntos rojos (zona 2) se concentran en su mayoría en dirección opuesta a las flechas de Distancia y de Duración, además de que, a pesar de ser variables diferentes, la distribución de los colores es bastante similar a la de la Figura 5.3, nuevamente dejando más que claro el hecho de que la mayoría de los viajes suceden dentro de la misma zona y confirmándonos que en efecto, la zona 2- ZAPOPAN CENTRO es probablemente la zona más compacta de las tres.

De forma similar a la sección anterior, también se consideró complementar el análisis exploratorio con métodos de agrupamiento, con el fin de ver si al menos para esta variable sucede algo interesante en el agrupamiento. Las variables consideradas en este caso son las siguientes.

	Edad	Genero	Dia	Hora	Duracion	Distancia	Intervalo de duracion	Zona de origen
0	36	0	1	18	6.916667	791.957123	2	1
1	28	0	5	19	11.383333	1802.525253	3	1
2	24	0	2	11	12.333333	1997.349364	3	2
3	53	1	4	19	8.866667	1430.854395	2	1
4	26	0	7	21	7.533333	1316.957751	2	2

Figura 6.2: Head of dataframe for clustering

El dendograma y la gráfica de variabilidad promedio para el algoritmo K-means para estos datos se

muestran a continuación.

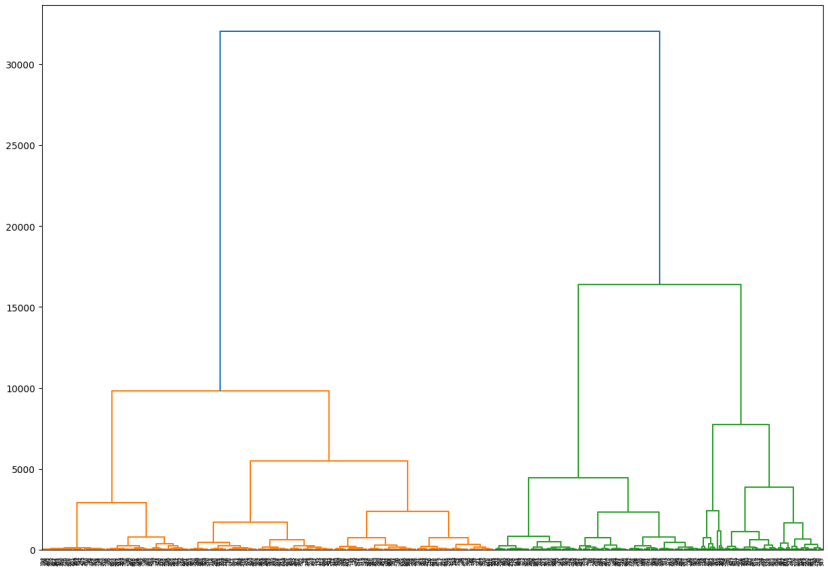


Figura 6.3: Dendograma

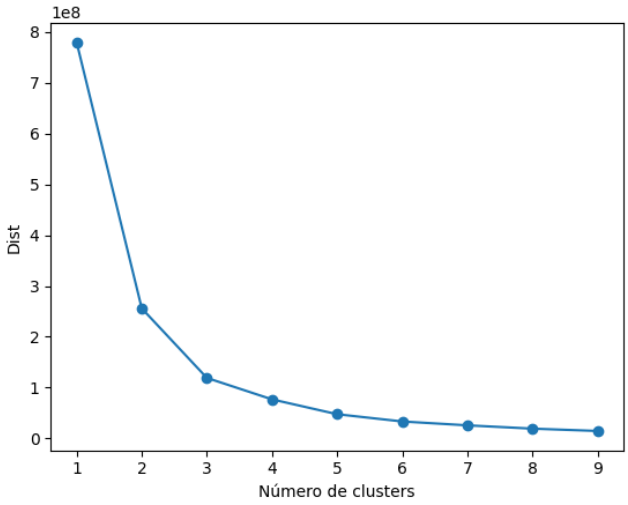


Figura 6.4: Variabilidad promedio por número de clusters en K-means

Con todo lo anterior, se consideró de nueva cuenta $k = 3$ en el algoritmo k-means y para la visualización de los clústers se utilizó el método T-SNE.

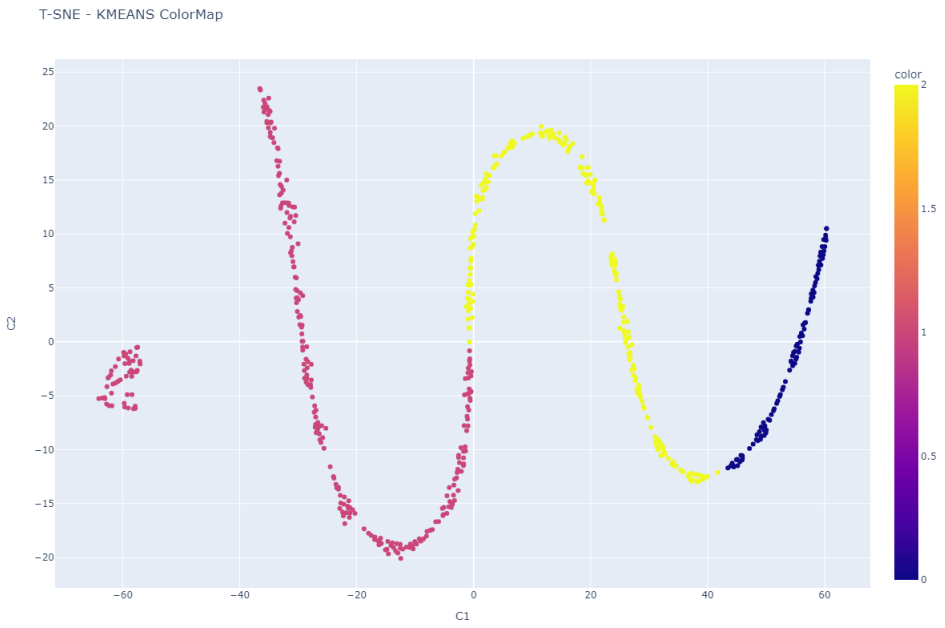


Figura 6.5: K-means con T-SNE

Nuevamente nos hacemos la pregunta ¿Para este caso los clústers pueden llegar a coincidir o tener relación con la zona de Destino? Desafortunadamente otra vez sucede que no hay una relación aparente, al menos en este caso y bajo este contexto.

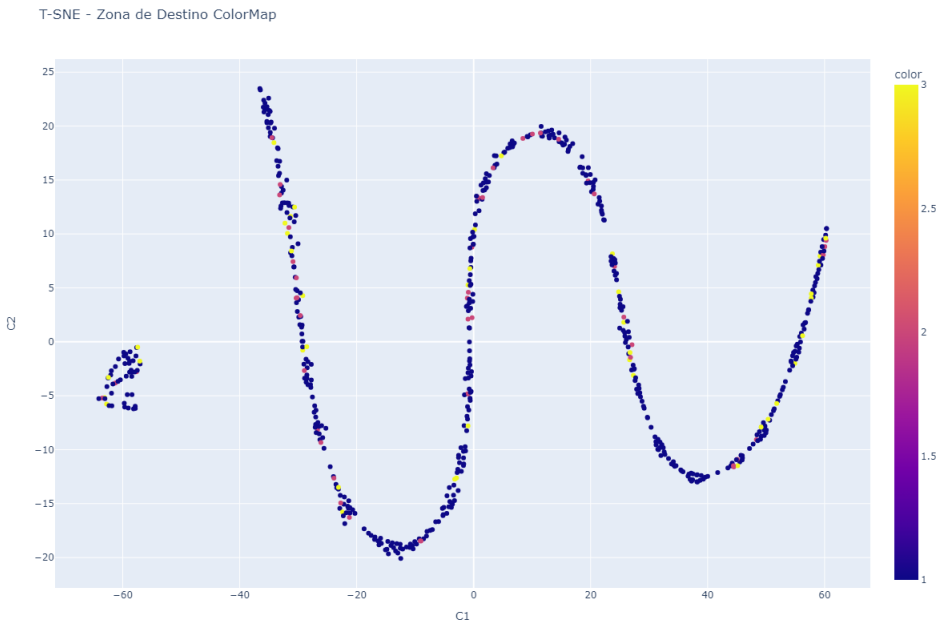


Figura 6.6: T-SNE con colormap de Zona de Origen

Por último, resta ver el poder predictivo del resto de variables, representadas con el siguiente dataframe, con respecto a la zona de Destino. Los métodos usados son los mismos que en las otras secciones.

	Edad	Genero	Dia	Hora	Duracion	Distancia	Zona de origen
0	36	0	1	18	6.916667	791.957123	1
1	28	0	5	19	11.383333	1802.525253	1
2	24	0	2	11	12.333333	1997.349364	2
3	53	1	4	19	8.866667	1430.854395	1
4	26	0	7	21	7.533333	1316.957751	2

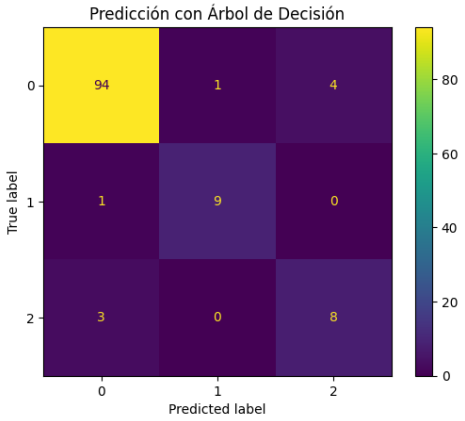
Figura 6.7: Head of dataframe for prediction

- **Árboles de decisión**

  Los resultados obtenidos para este caso son los siguientes.

	precision	recall	f1-score	support
1	0.96	0.95	0.95	99
2	0.90	0.90	0.90	10
3	0.67	0.73	0.70	11
accuracy			0.93	120
macro avg	0.84	0.86	0.85	120
weighted avg	0.93	0.93	0.93	120

(a) Reporte



(b) Matriz

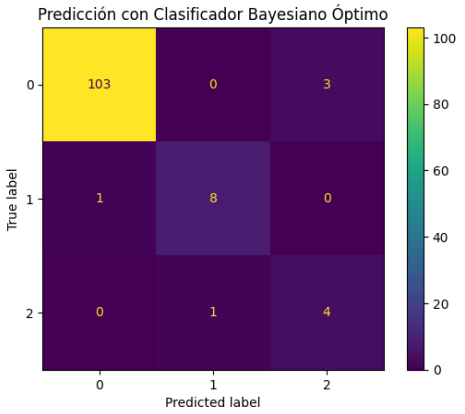
Figura 6.8: Resultados de Predicción

Esto no hace más que confirmar la innegable correlación que hay entre la zona de Origen y la zona de Destino, dado que volvimos a obtener una excelente precisión en las predicciones, del 93 %.

- **Clasificador Bayesiano Óptimo**

	precision	recall	f1-score	support
1	0.99	0.97	0.98	106
2	0.89	0.89	0.89	9
3	0.57	0.80	0.67	5
accuracy			0.96	120
macro avg	0.82	0.89	0.85	120
weighted avg	0.97	0.96	0.96	120

(a) Reporte



(b) Matriz

Figura 6.9: Resultados de Predicción

Algo a destacar con esto es que el Clasificador Bayesiano Óptimo demostró otra vez ser mejor que el método por árboles de decisión, ya que alcanzó una precisión del 96 %.

7. Variable de interés: Hora del viaje

En esta sección consideramos como variable de interés a la **Hora del viaje**. Como se estableció en la sección [2], en el *dataframe* original manejamos las horas de 0 a 23, pero debemos de encontrar una categorización de dichas horas para manejar la variable en términos de modelos predictivos. Por ende, consideremos la siguiente categorización

Clase	Intervalo de Tiempo
0	(0 - 6)
1	(6 - 12)
2	(12 - 18)
3	(18 - 23)

Cuadro 1: Categorización de las horas del viaje

Como en las secciones anteriores, proseguimos el análisis exploratorio realizando Pca con el siguiente dataframe

	Edad	Dia	Duracion	Distancia
0	36	1	6.916667	791.957123
1	28	5	11.383333	1802.525253
2	24	2	12.333333	1997.349364
3	53	4	8.866667	1430.854395
4	26	7	7.533333	1316.957751

Figura 7.1

Entonces, al realizar una prueba de componentes principales (exceptuando la variable de hora) con los datos estandarizados obtuvimos la siguiente gráfica de la varianza explicada.

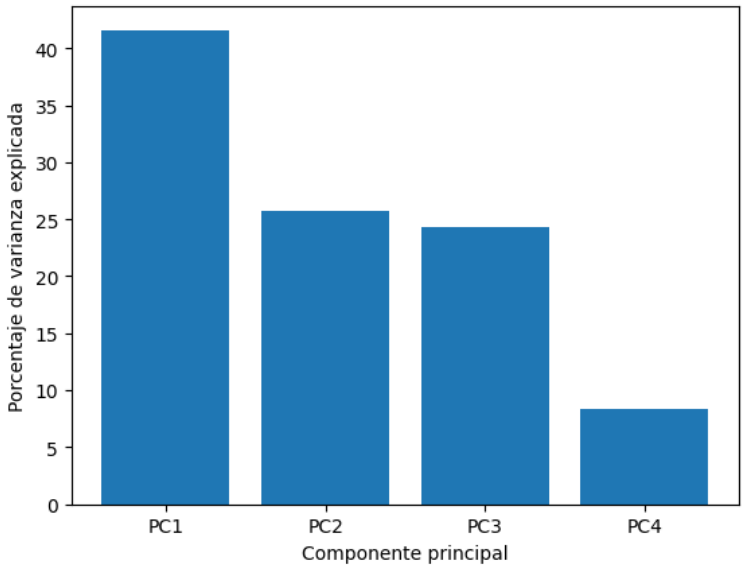


Figura 7.2

Notemos en la figura anterior que el primer componente tiene un porcentaje de varianza explicada alrededor 40 %, lo cual es relativamente alto cuando lo comparamos con lo obtenido en las secciones anteriores. Entonces, para visualizar los datos consideraremos los primeros dos componentes y graficaremos el color de los puntos con respecto a el Cuadro 1. Veamos que

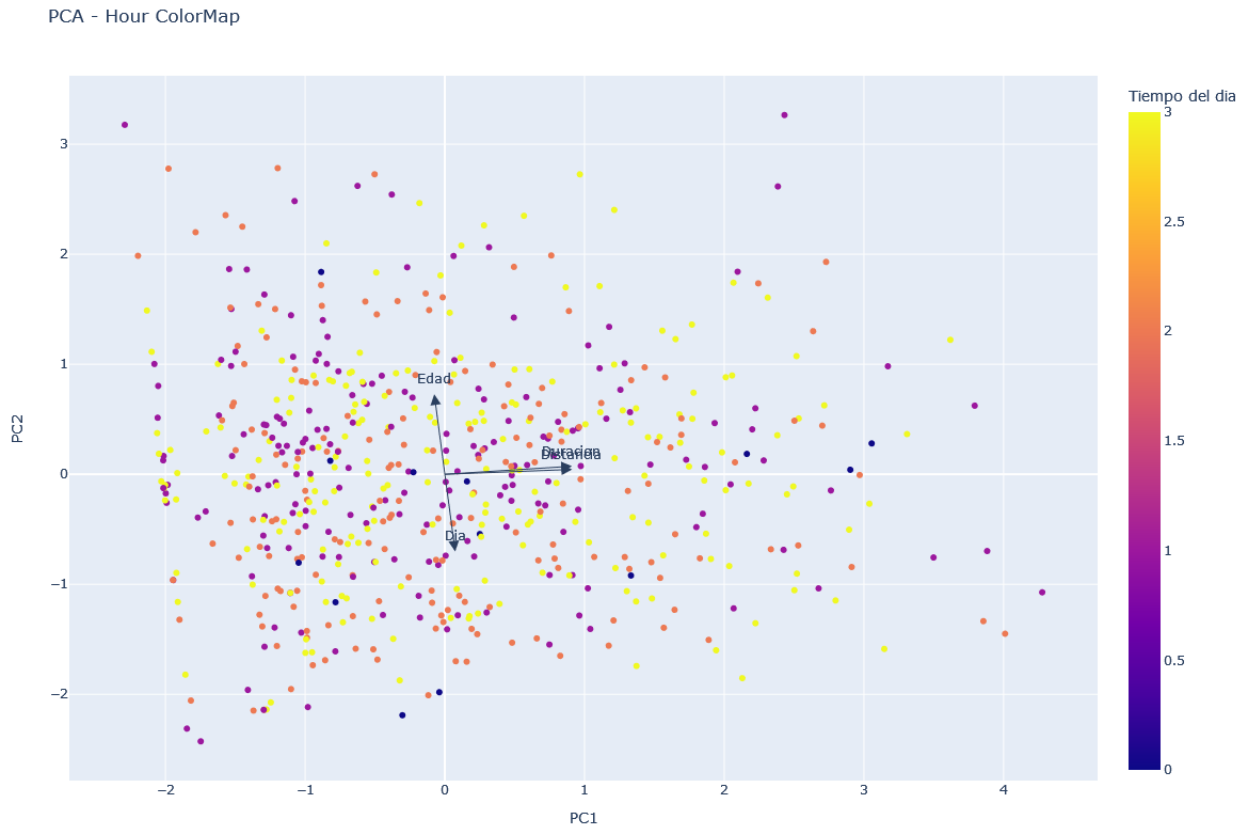


Figura 7.3

Situándonos en la Figura 7.3 podemos observar inmediatamente que no existe una representación o una relación visual entre los datos del dataframe de la Figura 7.1 y la variable del Tiempo del día. Ahora, como era de esperar, en cuanto a las direcciones de varianza de las características, la **duración** y la **distancia** están correlacionadas. Con la gráfica anterior podemos realizar la siguiente suposición, el poder predictivo de las variables Edad, Día, Duración y Distancia para predecir la hora del viaje es bajo. Para confirmar dicha suposición procederemos por métodos de agrupamiento. De nuevo, comenzamos observando el dendograma para darnos una idea del numero posible de clusters que se pueden obtener de la muestra.

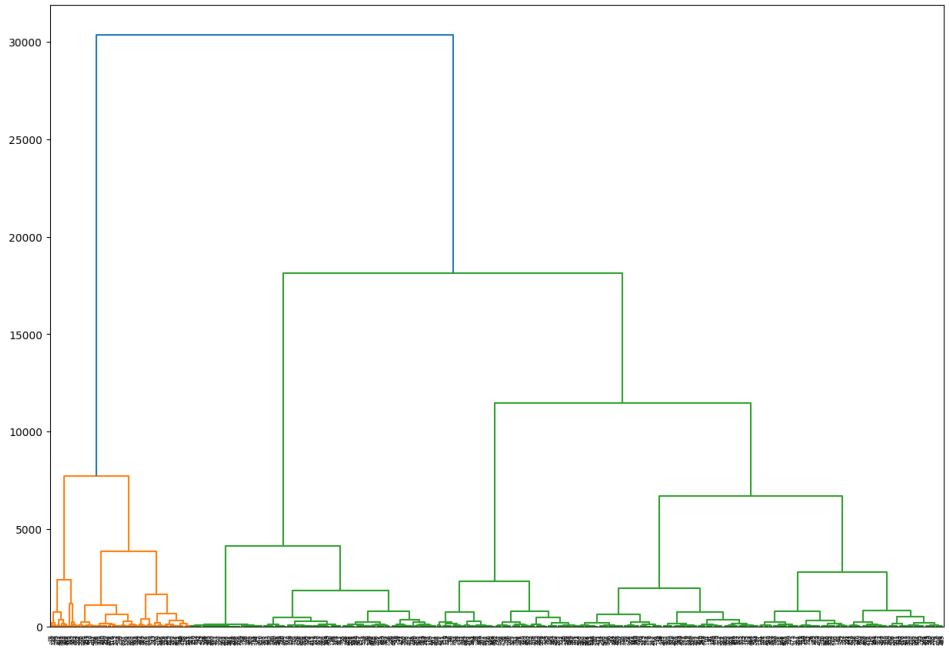


Figura 7.4

Ahora, notemos que el dendograma no es muy claro, pero por lo menos se pueden identificar 2 clusters, por lo que debemos de implementar la regla del codo para realizar k-*means*. Veamos el siguiente grafo

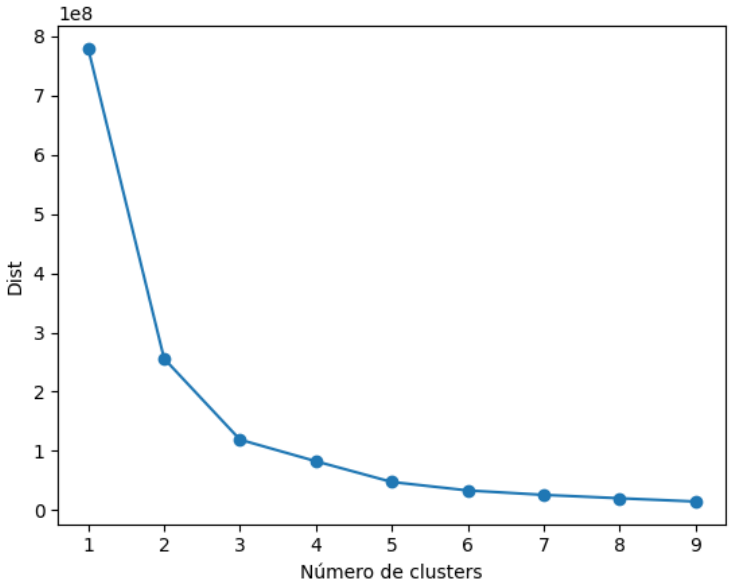


Figura 7.5

Observemos que a partir de 3 cluster no se produce ningún cambio significatio en la función, por lo tanto realizaremos k-*means* considerando 3 centroides. Como se realizó en las demás secciones primero haremos uso de T-SNE para obtener una representación bidimensional de los datos e identificaremos los puntos graficados con el colormap correspondiente a los clusters obtenidos de k-*means*.

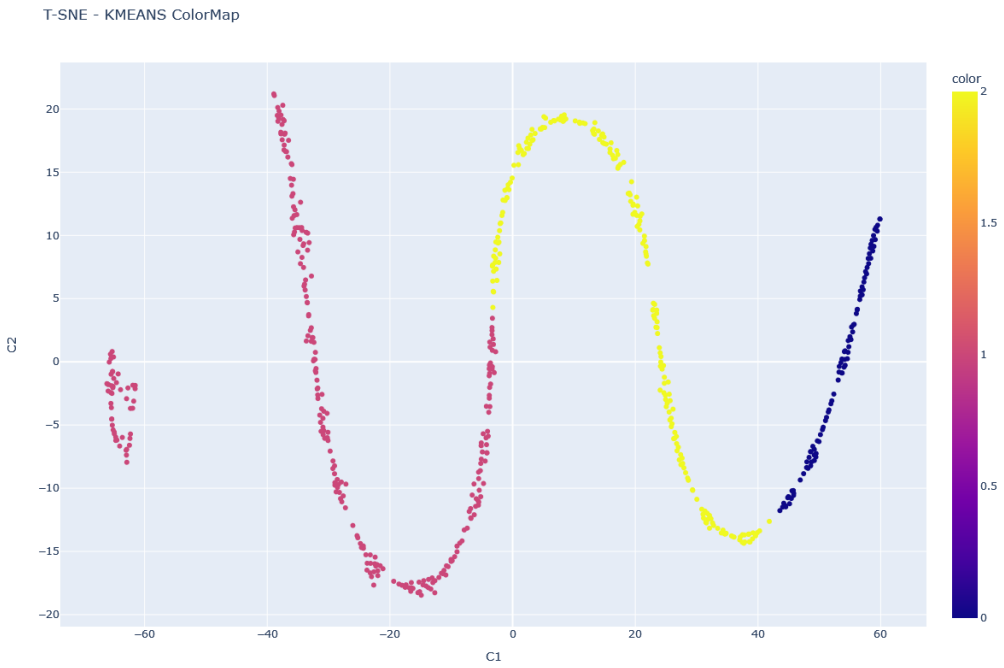


Figura 7.6: Colormap de k-means

Ahora, considerando dicha gráfica procederemos a realizar un colormap diferente, usando tanto las horas continuas como las clases del cuadro 1 para verificar si existe una coincidencia con las clases de k-*means* lo cual nos daría una evidencia de agrupamiento con respecto a la hora del día.

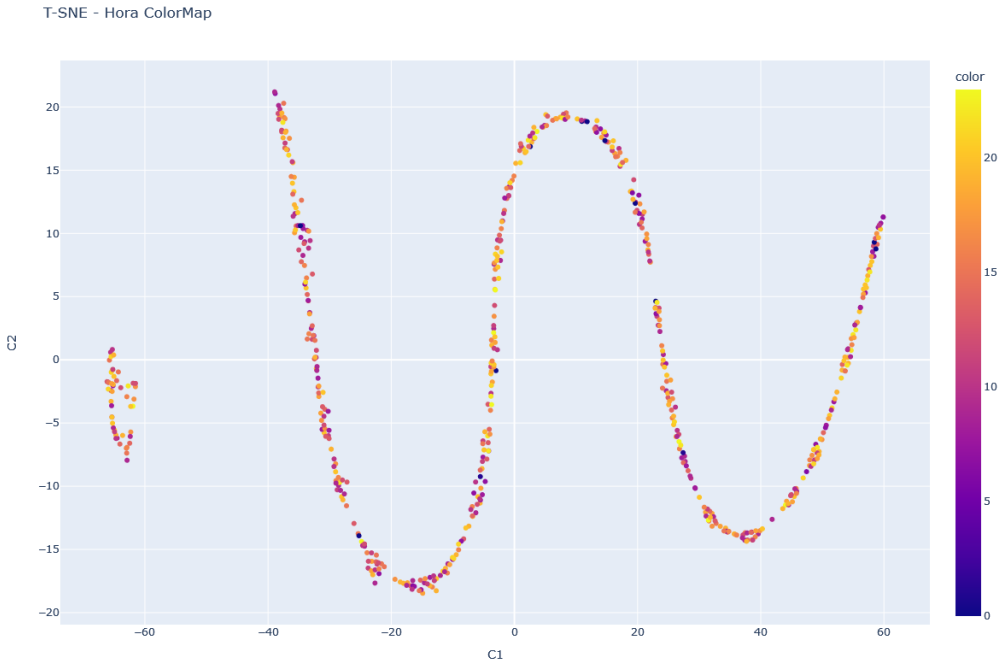


Figura 7.7: Colormap de hora (continua)

Notemos que al comparar tanto la Figura 7.6 como la Figura 7.7 no existe una clara relación entre las clases obtenidas por k-*means* y las el colormap de las horas del día y por lo tanto, no existe una agrupación de los datos con respecto a las horas del día. Ahora, consideremos la misma gráfica T-SNE pero con un colormap del cuadro 1.

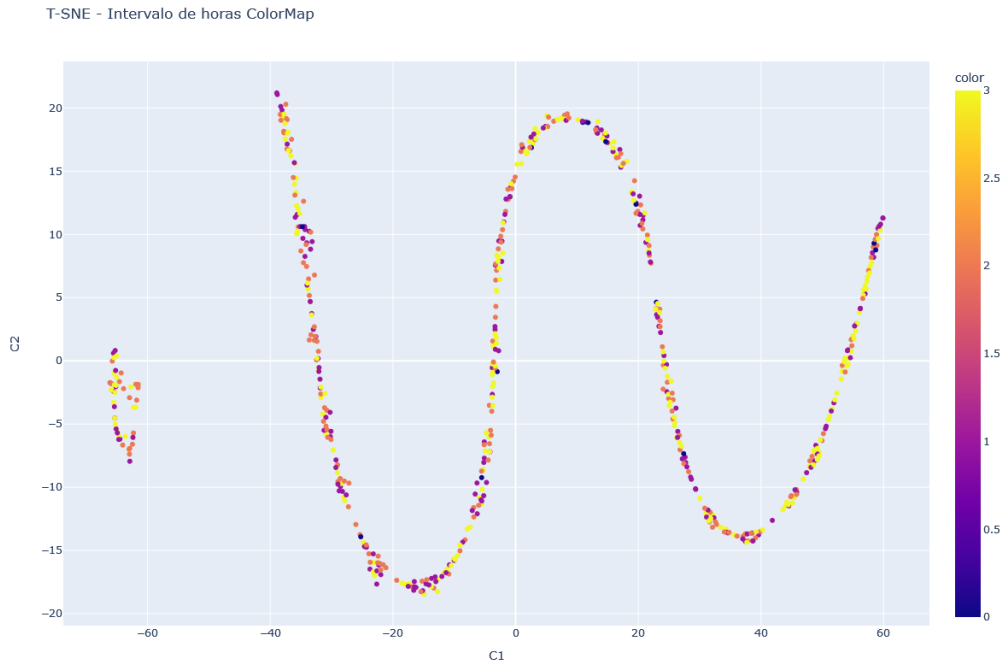


Figura 7.8: Colormap de clase cuadro 1

De nuevo, no existe una relación con el agrupamiento dado por k-*means* al colorear con la clase dada por el cuadro 1. Con esto establecido, sabemos que nuestro poder predictivo va a ser muy bajo para predecir el la hora del día en la que se realiza el viaje. Para confirmarlo, procederemos a implementar modelos de clasificación. En ambos modelos se usa el siguiente *dataframe*

	Edad	Genero	Dia	Duracion	Distancia	Zona de origen	Zona de destino
0	36	0	1	6.916667	791.957123	1	1
1	28	0	5	11.383333	1802.525253	1	1
2	24	0	2	12.333333	1997.349364	2	2
3	53	1	4	8.866667	1430.854395	1	1
4	26	0	7	7.533333	1316.957751	2	2

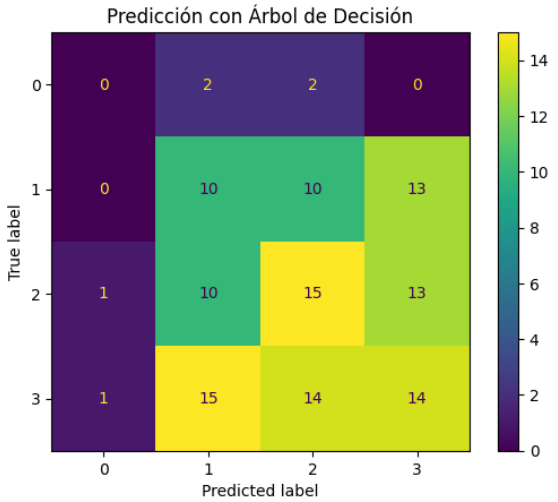
Figura 7.9: Head of Dataframe - Hour

- **Árboles de Decisión**

  Los resultados de un modelo de arboles de decisión son los siguientes

	precision	recall	f1-score	support
0	0.00	0.00	0.00	4
1	0.27	0.30	0.29	33
2	0.37	0.38	0.37	39
3	0.35	0.32	0.33	44
accuracy			0.33	120
macro avg	0.25	0.25	0.25	120
weighted avg	0.32	0.33	0.32	120

(a) Reporte



(b) Matriz

Figura 7.10: Resultados de Predicción

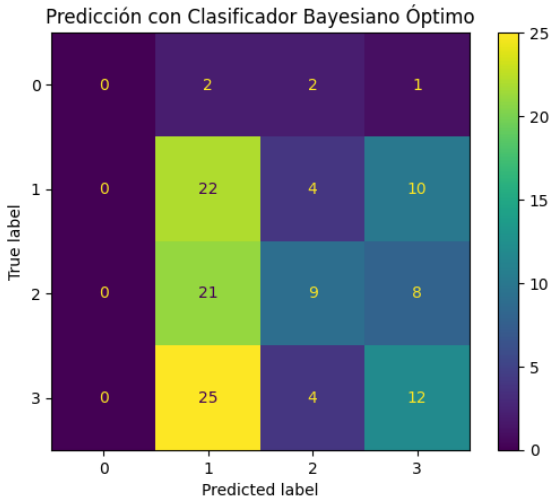
Obtuvimos una exactitud de 33 %, la cual es muy baja pero recordemos que estamos trabajando con datos reales. Ahora, con esto podemos concluir usando árboles de decisión que el poder predictivo de variables como la Edad, el Genero, el Día, la Duración y las Zonas de Origen y Destino son bajas para predecir la hora del viaje. Ahora, sin enfocamos en la Matriz de Confusión podemos observar que usualmente el modelo optaba por las últimas 3 clases, es decir, que el mayor movimiento de viajes se da después de las 6 am.

- **Clasificador Bayesiano Óptimo**

  Con esto modelo predicitvo obtuvimos resultados similares aunque hubo una mejoría muy pequeña.

	precision	recall	f1-score	support
0	0.00	0.00	0.00	5
1	0.31	0.61	0.42	36
2	0.47	0.24	0.32	38
3	0.39	0.29	0.33	41
accuracy			0.36	120
macro avg	0.29	0.29	0.27	120
weighted avg	0.38	0.36	0.34	120

(a) Reporte



(b) Matriz

Figura 7.11: Resultados de Predicción

Notemos que de 33 % pasamos a 36 % y ahora en la Matriz de Confusión la clase por la que el modelo opta es la clase 1 (6 - 12). De todos modos, obtenemos resultados de predicción no favorables.

8. Conclusiones

Como conclusión, los datos no muestran evidencia de aportar un buen poder predictivo para predecir la duración del viaje y la hora en la que se realizó el viaje. A la hora de analizar el comportamiento entre las zonas de origen y destino llegamos a que existe cierto sesgo para predecir esas variables de interés ya que las zonas son muy grandes y por ende una gran mayoría de los viajes se realizan dentro de la misma zona. En cuanto a la duración, se cumple que la variable con mayor poder predictivo es la distancia recorrida y esto se debe a que existe una gran correlación entre ambas variables.

Es importante mencionar que el análisis exploratorio de la muestra es un parte crucial de este proyecto para encontrar modelos estadísticos que posiblemente describen a los datos. Específicamente, analizar variables relevantes, representar gráficamente patrones y identificar valores atípicos y datos faltantes son actividades fundamentales para mejorar la precisión de los modelos de aprendizaje automático.