Universidad Tecnológica de Panamá

Facultad de Sistemas Computacionales

Maestría en Analítica de Datos

Modelos Predictivos

1AN214

Modelo predictivo de la demanda de bicicletas según horario en Capital Bikeshares DC

Presentado por:

Andy Sanjur

Facilitador:

Juan Marcos Castillo, PhD

Jornada documentada al 9 de abril del 2025

Introducción

Capital Bikeshares es un sistema de renta de bicicletas en la ciudad de Washingtong D.C. inaugurada el 20 de septiembre del 2010 cuenta con al menos 700 estaciones y 5400 unidades disponibles para renta, es el segundo sistema más grande de renta de bicicletas de Estados Unidos.

Este sistema está orientado en ser una alternativa que reduzca la huella de carbono, interesado principalmente en reinventar la movilidad de los usuarios en grandes ciudades; es una realidad que el congestionamiento tráfico diario dificulta la movilidad en horas pico bajo ciertas condiciones, sin embargo, Capital Bikeshares DC en respuesta recurre a las bicicletas como la principal herramienta de transporte.

La tecnología actual permite facilitar la logística de un sistema con cientos de estaciones y miles de unidades con el fin de crear una red que abarque una gran área metropolitana; la principal característica de este sistema radica en su practicidad, el usuario puede llegar de un punto a otro sin necesidad de retornar la unidad a un punto de origen, esta flexibilidad lo hace altamente atractivo para uso casual ya que comparado a otros sistemas de transporte colectivo la transacción se hace prácticamente inmediata sin tiempo de espera u horarios.

Cada usuario puede comprar una membresía de 24 horas o 72 horas, esto le permite retirar una bicicleta de cualquier estación realizar un viaje de ida y devolverla a cualquier otra estación, en un periodo máximo de 30 minutos, al excederse de este tiempo se cobra la diferencia.

Los sistemas GPS permiten la trazabilidad de las unidades a lo largo de la ciudad, proporcionan la ubicación, hora de inicio del viaje, hora de finalización y junto con información del día de la semana e información histórica de las condiciones climáticas se procede con un estudio de las principales razones que influyen en el uso del sistema, al igual que proyecta información suficiente para pronosticar la demanda horaria del sistema.

Definición del problema

De acuerdo con el funcionamiento de la red de bicicletas compartidas, un usuario puede iniciar su viaje en una estación particular y finalizar en otra en un radio cercano y permanecerá ahí hasta que otro usuario la utilice y finalice su viaje en otro nodo o regrese al punto de inicio, esto implica que durante horas pico del día algunas estaciones se verán saturadas en bicicletas mientras que otras pueden presentar escasez.

Es necesario poder predecir el volumen del uso de las bicicletas en horas pico para evitar que usuarios con membresías experimenten cuellos de botella frente el auge de la popularidad del sistema de bicicletas compartidas ante el público como medio de transporte diario alternativo.

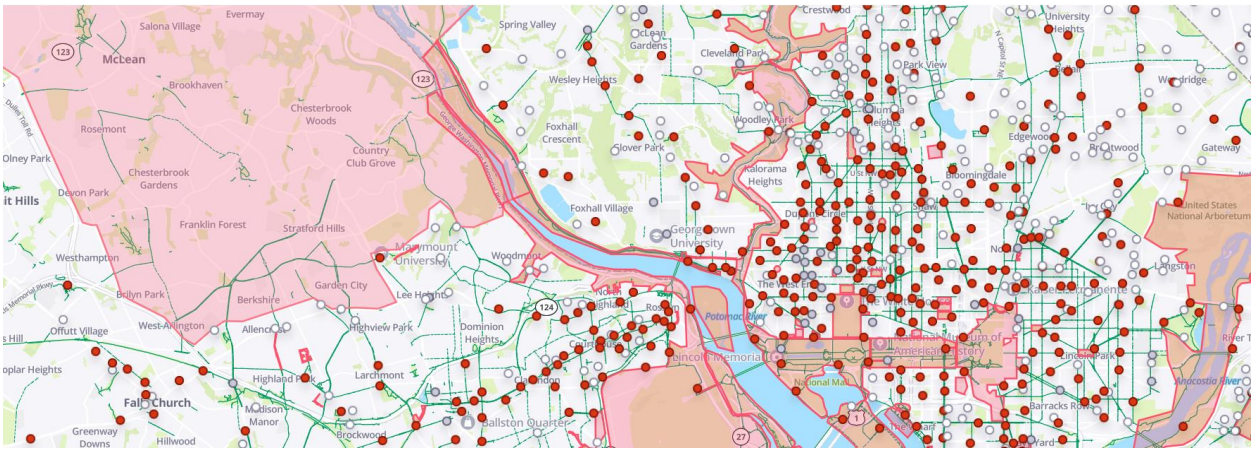

Figura 1. Mapa de estaciones de Capital Bikeshares DC. Disponible en https://capitalbikeshare.com

La disponibilidad de las bicicletas está fuertemente ligada a su nivel de batería, es decir en hora pico la bicicleta puede cumplir un solo ciclo de 30 minutos y no estará disponible hasta que su carga esté nuevamente al 100% esto impacta directamente el inventario y el mantenimiento de los equipos con el fin de rotar las unidades en lotes que estén con suficiente carga.

Antecedentes

El sistema de bicicletas compartidas no es una novedad reciente, inicialmente la ciudad de Ámsterdam en el año 1964 contaba con bicicletas gratuitas que se encontraban distribuidas por la ciudad, este programa llevó a que en un mes las 10 bicicletas de la red fueran robadas o dañadas.

En agosto del 2008 el distrito de Columbia se convirtió en la primera ciudad en lanzar un sistema de bicicletas compartidas, Smart Bike DC contaba con 10 estaciones y 120 bicicletas, en los primeros 2 años de operación al menos 1600 personas se inscribieron en el programa (capital bikeshares, 2025).

Para el año 2022 otros países han implementado este modelo de negocio, al menos 1590 ciudades en 92 países ampliamente popular en los continentes de Europa y Asia (The Meddin Bike-sharing World Map Report, 2022).

Tomando como referencia el artículo científico “Un enfoque de análisis predictivo para pronosticar la demanda de alquiler de bicicletas”, se analiza la correlación de las variables entre las cuales destacan una correlación moderada de la cantidad de bicicletas alquiladas con la temperatura y seguido con una correlación moderada baja la hora del día, estudiando a detalle las variables también el autor destaca que existe una disminución en el conteo para días de vacaciones, feriados y domingos (Karunanithi, Chatasawapreeda, & Ali Khan, 2024).

Los datos utilizados en este estudio fueron encontrados en el repositorio de datos de Kaggle como Rental Bike Sharing Dataset, corresponden a un total de 17379 registros

representativos a 2 años (2011-2012) desde la implementación del proyecto en la ciudad de Washington (Fanaee-T & Gama, 2013).

Los datos muestran la cantidad de rentas de totales, por usuarios casuales y por usuarios que son miembros y se sabe tienen correlación con las condiciones climáticas como humedad, temperatura y velocidad del viento a diferentes horas por día.

Justificación

Es necesario conocer los patrones de comportamiento del sistema a lo largo del día y a lo largo de las semanas, se busca identificar los factores más influyentes en el incremento del volumen de rotación del inventario, es decir describir si el uso de las bicicletas se orienta a la recreación casual o al transporte diario, de igual manera agrega valor poder agrupar las principales horas de movimiento en bloques diarios que puedan hacer más practico el estudio, con esto se espera seleccionar un modelo predictivo que arrojen proyecciones del incremento del volumen de unidades que pueden estar simultáneamente viajando de una estación a otra para evitar escasez de inventario en momentos de alta demanda, la unidades de la flota cuentan con un factor de seguridad que contemple el tiempo que la unidad esté en uso y el tiempo que tarda en recargar.

Cabe destacar que la base de datos utilizada no cuenta con un nivel de granularidad que identifique los niveles de inventario de estaciones específicas en instantes durante el día por lo tanto el presente estudio se limitará principalmente en analizar el volumen bruto de la red y la influencia de los días de la semana, temporadas y diferentes horas del día.

Análisis predictivo

El análisis del conjunto de datos se puede subdividir en 2 principales etapas, análisis descriptivo y análisis predictivo, a continuación, se describen brevemente los resultados encontrados en cada sección.

a. Determinación de la base de datos

Se selecciona el conjunto de datos de Capital BikeShares que cuenta con 2 años de registros 2011 y 2012 para las 24 horas diarias incluyendo datos de temperatura, humedad, velocidad del viento, categorización de los días de la semana y se selecciona como principales salidas inicialmente alquiler por usuarios registrados, alquiler por usuarios casuales y el conjunto.

	instant	dteday	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
0	1	1/1/2011	1	0	1	0	0	6	0	1	0.24	0.2879	0.81	0.0000	3	13	16
1	2	1/1/2011	1	0	1	1	0	6	0	1	0.22	0.2727	0.80	0.0000	8	32	40
2	3	1/1/2011	1	0	1	2	0	6	0	1	0.22	0.2727	0.80	0.0000	5	27	32
3	4	1/1/2011	1	0	1	3	0	6	0	1	0.24	0.2879	0.75	0.0000	3	10	13
4	5	1/1/2011	1	0	1	4	0	6	0	1	0.24	0.2879	0.75	0.0000	0	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
17374	17375	12/31/2012	1	1	12	19	0	1	1	2	0.26	0.2576	0.60	0.1642	11	108	119
17375	17376	12/31/2012	1	1	12	20	0	1	1	2	0.26	0.2576	0.60	0.1642	8	81	89
17376	17377	12/31/2012	1	1	12	21	0	1	1	1	0.26	0.2576	0.60	0.1642	7	83	90
17377	17378	12/31/2012	1	1	12	22	0	1	1	1	0.26	0.2727	0.56	0.1343	13	48	61
17378	17379	12/31/2012	1	1	12	23	0	1	1	1	0.26	0.2727	0.65	0.1343	12	37	49

Figura 2. Resumen del conjunto de datos Capital Bikeshares DC.

Este fue seleccionado debido a que muestra un comportamiento de estacionalidad y cuenta con suficientes datos para realizar un estudio detallado de las principales preferencias horarias de los usuarios, al estudiar un sistema de transporte este se ve orientado por picos de uso a lo largo del día.

b. Preprocesamiento y limpieza

Las siguientes secciones se desarrollan utilizando Visual Studio Code para la limpieza de datos, el conjunto de datos Capital Bikeshares DC se carga como un archivo “.csv” y se define el dataframe como Bikesdf.

Se inicia con un análisis exploratorio de los datos, se aplica la función “.info()” de la librería pandas para explorar los tipos de datos y el conteo de datos nulos en las columnas y se genera el siguiente listado.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17379 entries, 0 to 17378
Data columns (total 17 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   instant     17379 non-null  int64
 1   dteday      17379 non-null  object
 2   season      17379 non-null  int64
 3   yr          17379 non-null  int64
 4   mnth        17379 non-null  int64
 5   hr          17379 non-null  int64
 6   holiday     17379 non-null  int64
 7   weekday     17379 non-null  int64
 8   workingday  17379 non-null  int64
 9   weathersit  17379 non-null  int64
 10  temp        17379 non-null  float64
 11  atemp       17379 non-null  float64
 12  hum         17379 non-null  float64
 13  windspeed   17379 non-null  float64
 14  casual      17379 non-null  int64
 15  registered  17379 non-null  int64
 16  cnt         17379 non-null  int64
dtypes: float64(4), int64(12), object(1)
memory usage: 2.3+ MB
```

Figura 3. Nombre y tipo de columnas, conteo de registros y detección de valores nulos.

El conjunto de datos cuenta con 17379 registros y 17 columnas de las cuales 12 columnas con variables tipo int64, 4 columnas con datos tipo float64 y 1 columna tipo object.

Este conjunto de datos no requiere limpieza de datos ya que no se encuentran datos nulos en ninguna columna, sin embargo, se ejecuta la función (Bikesdf_C.drop_duplicates(subset="instant", inplace=True)

Para la eliminación de duplicados en la columna de id, no se detectaron duplicados, el conjunto de datos se mantiene en 17379 registros.

c. Análisis descriptivo

Para iniciar el análisis descriptivo del conjunto de datos se revisa la estadística básica de los valores utilizando la función ".describe()" de la librería pandas, solo se seleccionan las columnas numéricas que no representan una categoría con el fin de entender sus promedios y dispersiones, se calcula la media, valores máximos, valores mínimos, desviación estándar, quantiles y conteo como se muestra a continuación.

	temp	hum	windspeed	casual	registered	cnt
count	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000
mean	0.496987	0.627229	0.190098	35.676218	153.786869	189.463088
std	0.192556	0.192930	0.122340	49.305030	151.357286	181.387599
min	0.020000	0.000000	0.000000	0.000000	0.000000	1.000000
25%	0.340000	0.480000	0.104500	4.000000	34.000000	40.000000
50%	0.500000	0.630000	0.194000	17.000000	115.000000	142.000000
75%	0.660000	0.780000	0.253700	48.000000	220.000000	281.000000
max	1.000000	1.000000	0.850700	367.000000	886.000000	977.000000

Figura 4. Estadística descriptiva.

Se debe destacar que la temperatura está normalizada en 41°C y la velocidad del viento normalizada en 67 km/h.

Se encuentra una media en los datos para cnt de 189.46, registered con 153.78 y casual de 35.67, por otra parte, al revisar la desviación estándar de los datos se encuentra una alta dispersión entre los datos, en las siguientes secciones se realiza un análisis más detallado de las variables para definir patrones en el comportamiento.

Histogramas

Se generan los histogramas para las variables cnt, registered, casual, temp y windspeed para explorar el comportamiento de las principales variables de salida.

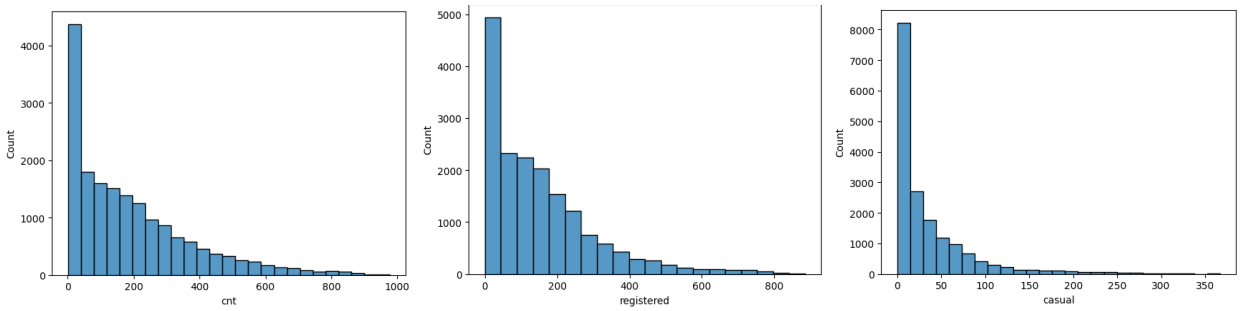

Figura 5. De izquierda a derecha se cuenta con el conteo total de bicicletas, conteo de alquiler por usuarios registrados y alquiler por usuarios casuales.

En los 3 casos de usuarios registrados, casuales y el conjunto se puede identificar un consistente sesgo a la izquierda por lo que la mayoría de los datos se mantienen cercanos a cero.

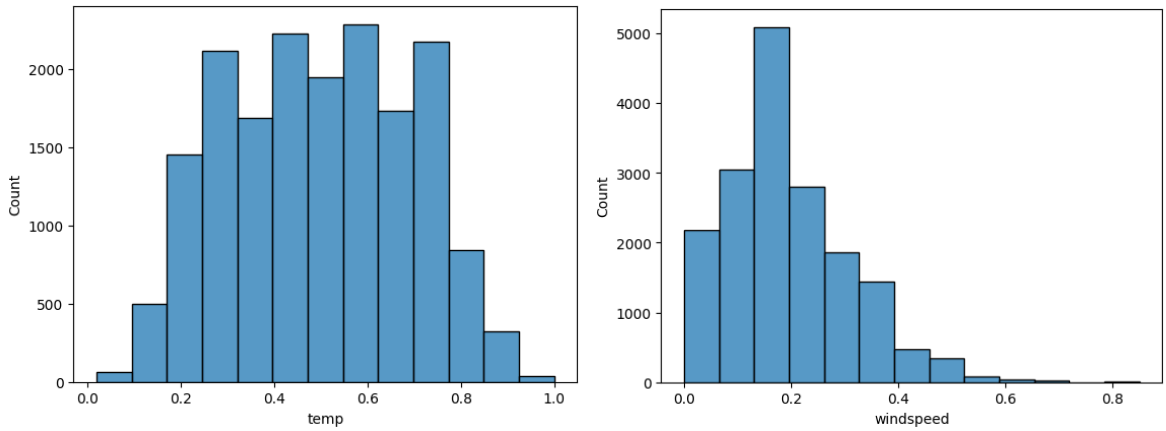

Figura 6. De izquierda a derecha, histograma según la temperatura e histograma según la velocidad del viento

Una vez se determine el histograma de estas 2 variables asemejan simetría en sus distribuciones, para el caso de la temperatura podemos encontrar la mayoría de los datos alrededor de 12.3 °C y 28.7 °C de igual manera la velocidad del viento se encuentra sesgada a la izquierda, con una moda muy marcada alrededor de 0.2 o 13.4 km/h.

Análisis de caja y bigotes

Se generan los diagramas de caja y bigotes para encontrar patrones y valores atípicos en las jerarquías temporales con las que cuenta el conjunto de datos en las variables día de la semana, hora del día y tipo de clima.

Utilizando la función de Boxplot de la librería seaborn se grafican los diagramas de caja y bigotes para la variable de cnt, registered y casual según la hora del día.

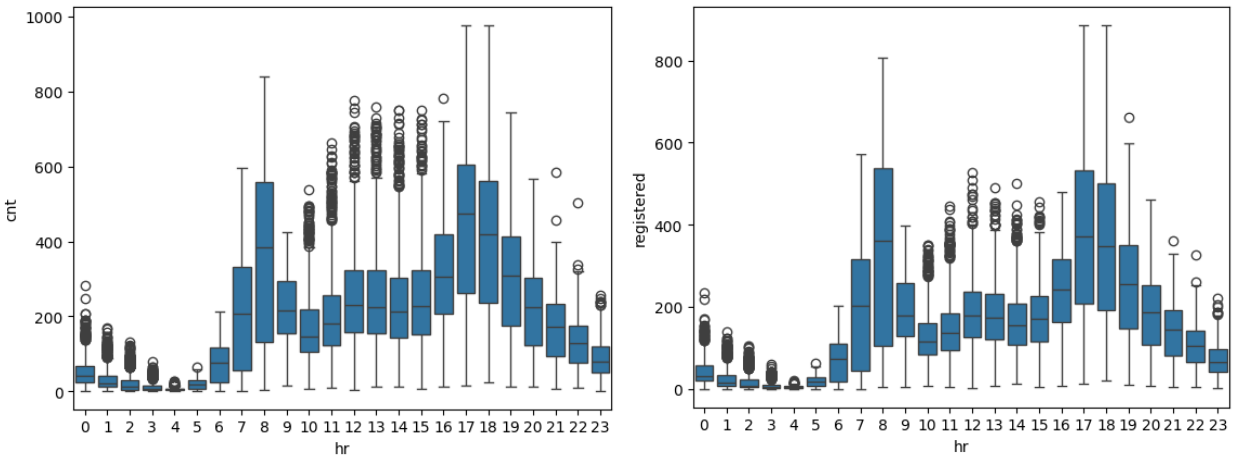

Figura 7. Gráficos de caja y bigotes para las variables cnt vs hr (izquierda) y registered vs hr (derecha)

Se encuentra que para ambos casos el comportamiento es muy similar, pero en diferentes proporciones, sin embargo, se encuentra una gran cantidad de valores atípicos entre las 10:00 y las 16:00, posteriormente se procede con una evaluación de los valores atípicos y se tratarán estos valores.

Se genera el diagrama de caja y bigotes para las variables de salida de usuarios casuales se muestra a continuación.

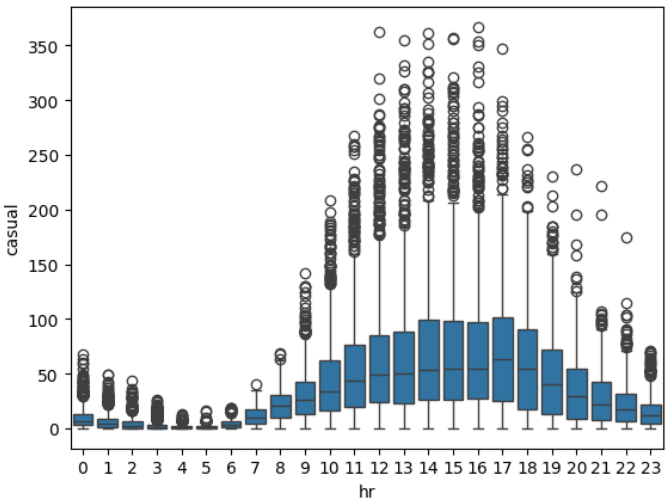

Figura 8. Diagrama caja y bigotes para casual vs hr.

Al comparar los gráficos anteriores se encuentra que estos se concentran en los horarios de menor demanda respecto al ser evaluados con la variable cnt, inconvenientemente también se encuentran múltiples valores atípicos en todas las horas del día a partir de este punto para simplicidad del análisis y nivel representativo de los datos, se procede a continuar los demás análisis en base a la cantidad (cnt) de alquiler total.

Para las variables de salida tomando en cuenta el día de la semana se generan los cuadros de caja y bigotes

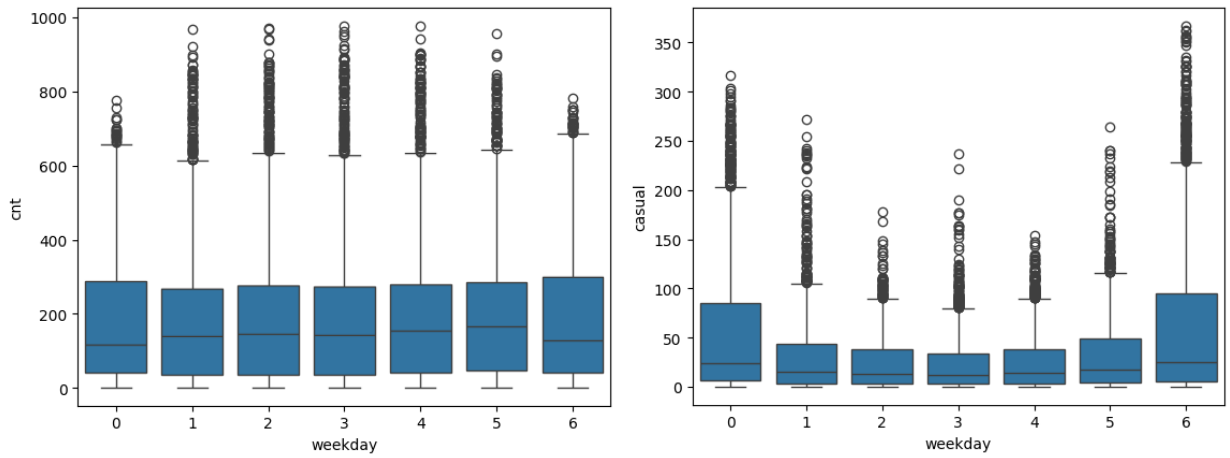

Figura 9. Diagrama de caja y bigotes para cnt vs weekday (izquierda) y casual vs weekday (derecha)

Utilizando las variables registered y cnt se encuentra que la distribución es prácticamente equivalente, en el caso de casual se encuentra que la media de los días de lunes a viernes es menor que para los fines de semana, mientras que para la variable combinada la media es prácticamente igual para cualquier día, en ambos casos la variabilidad es alta y se presentan muchos valores atípicos.

Finalmente, el último análisis de caja y bigote que se realiza corresponde a las condiciones climáticas.

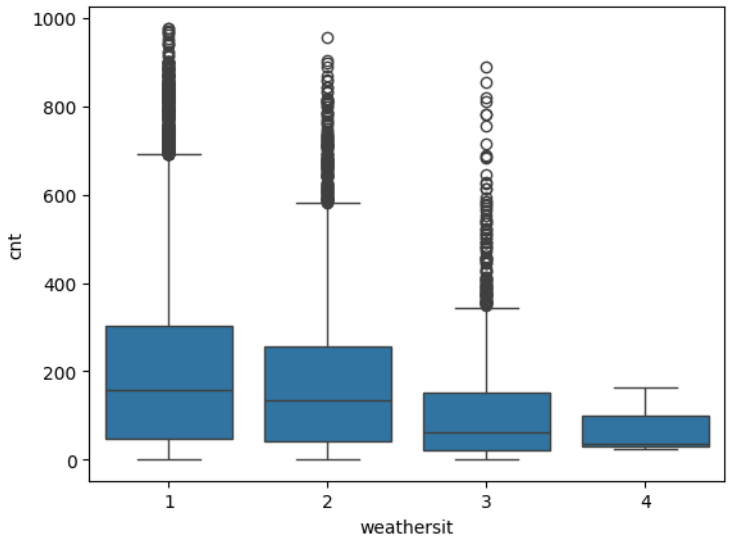

Figura 10. Diagrama de caja y bigote para cnt vs weathersit

Nuevamente, se encuentra alta variabilidad para las 3 primeras condiciones con una gran cantidad de valores atípicos.

Matriz de correlación

Con el fin de validar que las variables seleccionadas tengan una influencia considerable en las variables de salida se realiza una matriz de correlación para evaluar la influencia de las variables del conjunto de datos, a partir de la librería Seaborn se genera la siguiente matriz de correlación.

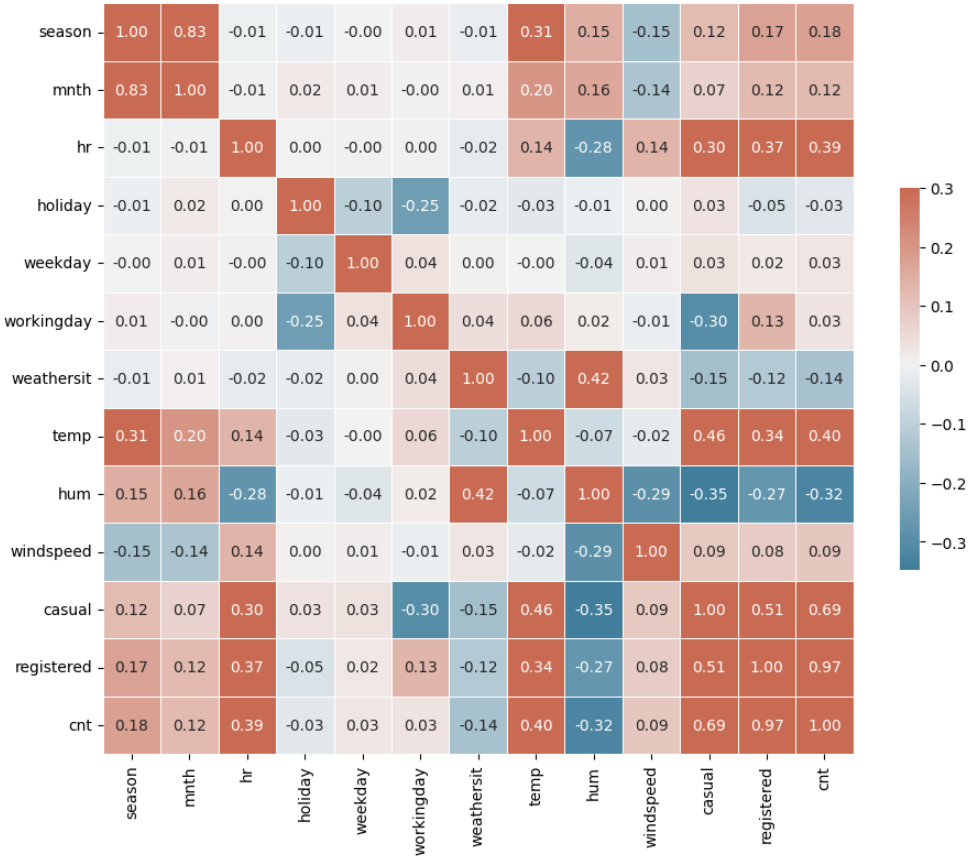

Figura 11. Matriz de correlación Capital Bikeshares.

- El conteo de usuarios combinados (cnt) se ve influenciado principalmente por la temperatura (temp = 0.4) y la hora (hr = 0.39) siendo una correlación moderada baja, además, muestra una correlación negativa con humedad (hum = -0.32).
- Para usuarios registrados que utilizaron el servicio podemos destacar que en primer lugar la hora (hr = 0.37) tiene una correlación prácticamente igual que la temperatura (temp = 0.34), por otra parte, la correlación es negativa para la humedad (hum = -0.27)
- Para usuarios casuales predomina la temperatura (temp = 0.46) y hora (hr = 0.30), sin embargo, se muestra una correlación negativa con la humedad (hum = -0.35) y en relación con el día de trabajo (workingday = -0.30).

Para este punto se encuentra que principalmente los factores con mayor influencia son la hora del día y la temperatura, sin embargo, para un mejor enfoque de la orientación del proyecto se tomará en cuenta la influencia de la hora como principal variable en el modelo predictivo, para poder determinar a corto plazo el incremento de los usuarios por día.

d. Selección de variables

Una vez se cuenta con el conjunto de datos debidamente descrito se procede a seleccionar las principales variables que formarán parte del pronóstico.

Para inicio de la jornada se busca predecir los valores picos que se tendrán a lo largo del tiempo por lo que nuestro estudio se inclina mayormente a un análisis por series de tiempo, más adelante en este documento se detallarán las herramientas y modelos que se utilizaron.

Para este caso particular se utilizará cnt para reducir la cantidad de iteraciones del modelo matemático y al referirse a la sección anterior de las posibles medidas de series de tiempo que utiliza el conjunto de datos se determina que la variable más apropiada es hora, ya que tiene una correlación moderada con los datos de salida.

Para evitar que los valores atípicos influyan en las predicciones se procede a realizar una winsorización agrupando los datos por hora, manteniendo los datos inferiores y aproximando los datos atípicos superiores al valor máximo de su grupo.

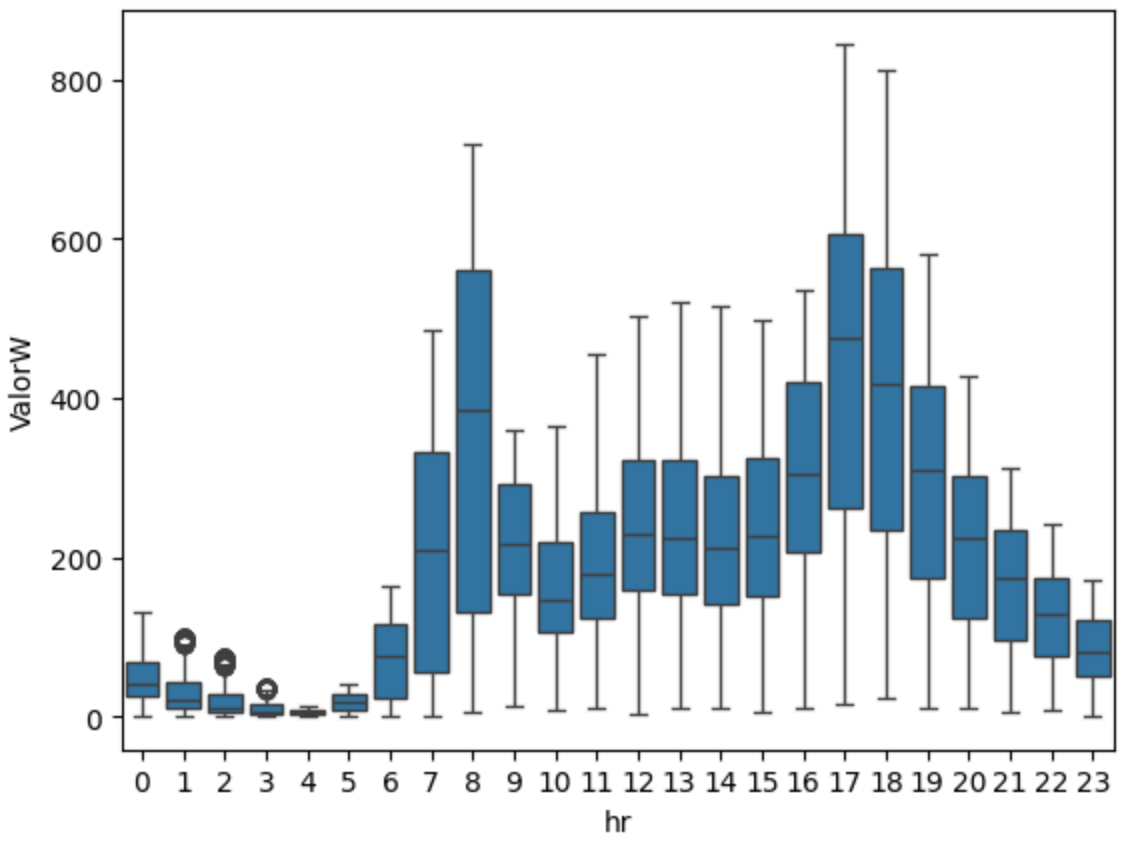

Figura 12. Winsorización de valores atípicos para la variable hr.

De este modo mejora la uniformidad de los datos del conjunto sin alterar considerablemente su comportamiento y medidas estadísticas básicas.

Finalmente, el modelo puede oscilar mucho a lo largo del día por ello se propone realizar un estudio predictivo por bloques, los grupos se dividen en horas de 5:00 a 10:00 am denominado como “Manana”, otro grupo de 11:00 a 14:00 denominado “MedioDia” y el último grupo denominado “Tarde” que es de 15:00 a 20:00, los demás datos se denominan “Noche” y para este estudio no se tomaran en cuenta debido a la baja demanda, por lo que la mayoría de los equipos serán cargados o se les dará mantenimiento durante estas horas.

e. Selección de modelos

- Holt-Winters

Para la selección de modelos se inicia por el modelo Holt Winter disponible en la herramienta de Machine Learning Weka, ya que Weka solo acepta datos temporales en el siguiente formato “yyyy-MM-dd HH:mm:ss” se realizó la transformación del conjunto de datos a exportar y se generaró 1 archivo .csv correspondientes al bloque de la tarde, de mayor interés de estudio por el volumen que usualmente maneja suelen ser los picos principales de alquiler.

Se estudia diariamente de manera aislada el bloque de la tarde que comprende desde las 15:00 hasta las 20:00 para un espacio de 3 meses desde agosto hasta octubre del año 2012.

Al ejecutar con los parámetros HoltWinter con alpha=0.2, beta=0.2 y gamma=0.2 utilizando la fecha como timestamp y periodicidad horaria se obtienen los siguientes resultados.

```
=== Evaluation on training data ===
Target                            1-step-ahead  2-steps-ahead  3-steps-ahead  4-steps-ahead  5-steps-ahead  6-steps-ahead
==========================================================================================================================
ValorW
  N                                       1492           1491           1490           1489           1488           1487
  Mean absolute error                 112.5726       128.1351       140.5043       149.6942       156.3393       160.0428
  Mean absolute percentage error       29.9377        33.8229        36.9709        39.2702        41.1278        42.4491
  Root mean squared error             154.7783        178.629       197.5096       211.5819       217.0526       212.8686
  Mean squared error                23956.3205     31908.3175     39010.0448     44766.8914     47111.8322     45313.0534

Total number of instances: 1516

=== Evaluation on test data ===
Target                            1-step-ahead  2-steps-ahead  3-steps-ahead  4-steps-ahead  5-steps-ahead  6-steps-ahead
==========================================================================================================================
ValorW
  N                                        674            673            672            671            670            669
  Mean absolute error                  104.452       117.4787       128.0105       136.0775        142.123        147.956
  Mean absolute percentage error       34.2293        38.4256        41.6741        43.9204        45.3331        47.3889
  Root mean squared error             142.4854       159.0316       170.7631       181.7431       190.7961       196.4681
  Mean squared error                20302.0852     25291.0593     29160.0297     33030.5414     36403.1623     38599.7168

Total number of instances: 674
```

Figura 13. Resultados del ensayo Holt-Winters ejecutado en Weka (alpha=0.2, beta=0.2 y gamma=0.2).

Sin embargo, al revisar la gráfica de predicción al utilizar la fecha como id de tiempo, se encuentran valores que el software interpola para llenar los espacios vacíos, esto es debido a que a pesar solo se estén estudiando 6 horas de una porción del día, Weka debe completar la información para 24 horas del día.

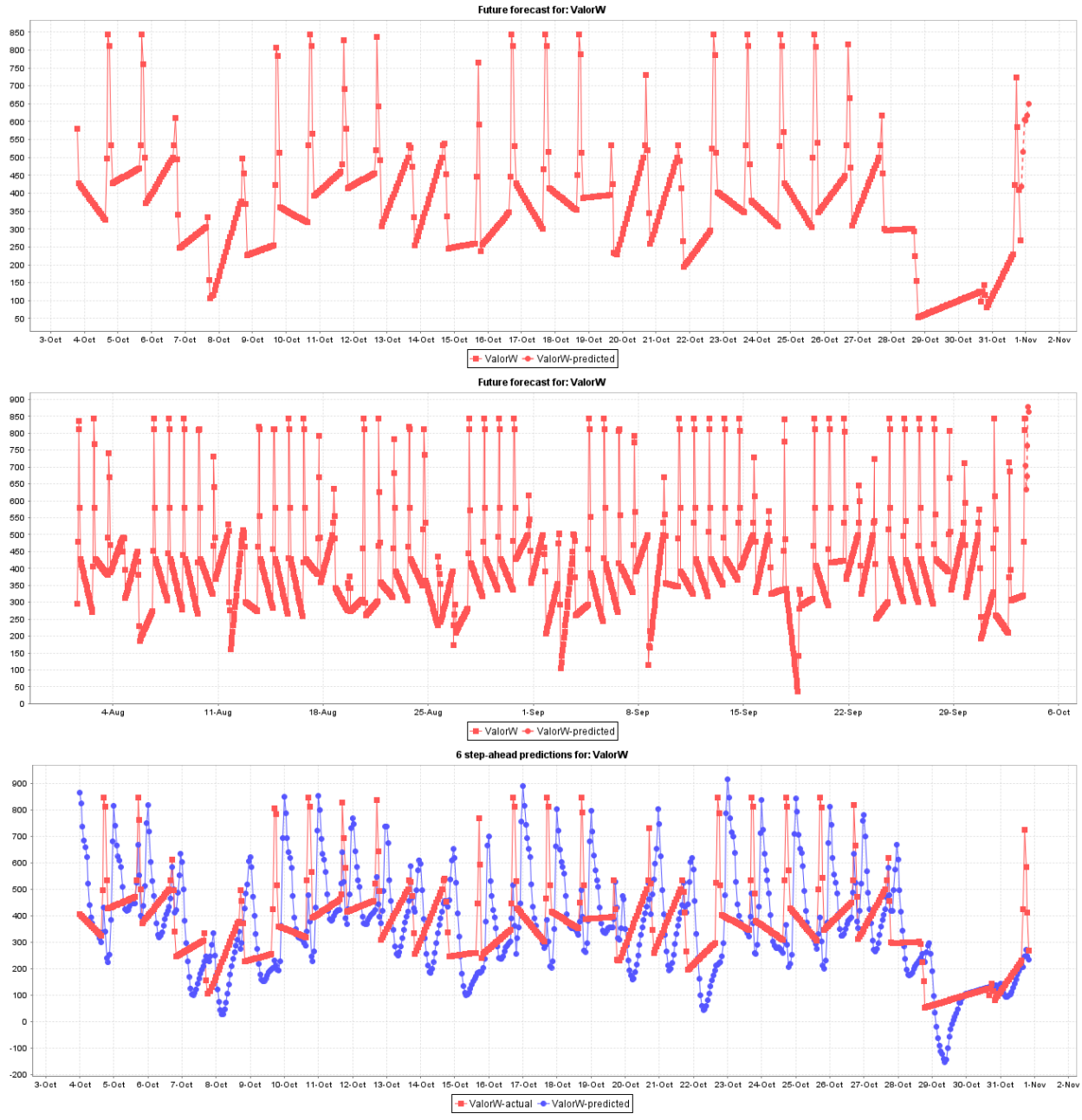

Figura 14. Test predictivo (Superior), Train predictivo (Medio), Train predictivo con la muestra original (Inferior).

Esto redirige el enfoque a asumir instancias de tiempo consecutivas para evitar la interpolación de datos.

Se modifica el id de tiempo a un índice artificial con distribución horaria, y nuevamente se somete el modelo a ejecución.

Inicialmente se utiliza la configuración alpha=0.2, beta=0.2 y gamma=0.2

```
=== Evaluation on training data ===
Target                              1-step-ahead  2-steps-ahead  3-steps-ahead  4-steps-ahead  5-steps-ahead  6-steps-ahead
==========================================================================================================================
ValorW
  N                                          358            357            356            355            354            353
  Mean absolute error                   232.1747       240.2971       215.6134       186.4873       145.2315       159.8867
  Mean absolute percentage error         47.5561        50.7561        50.7562        48.4151        37.6535        39.1337
  Root mean squared error              1268.8098      1248.0924       853.2623       626.4923        241.595        276.515
  Mean squared error                1609878.2744   1557734.7626    728056.5546    392492.6414     58368.1653     76460.5246

Total number of instances: 382

=== Evaluation on test data ===
Target                              1-step-ahead  2-steps-ahead  3-steps-ahead  4-steps-ahead  5-steps-ahead  6-steps-ahead
==========================================================================================================================
ValorW
  N                                          164            163            162            161            160            159
  Mean absolute error                   154.0275       174.1024        181.693       178.8215       172.8642       193.2161
  Mean absolute percentage error         41.4043        47.6386        50.9986        51.9091         52.416        58.6379
  Root mean squared error               199.2646       224.0506       234.8851       233.6917       231.3303       256.8501
  Mean squared error                 39706.3809     50198.6803     55171.0135     54611.7965     53513.7232     65971.9831

Total number of instances: 164
```

Figura 15. Primera iteración del modelo HoltWinters (alpha=0.2, beta=0.2 y gamma=0.2)

Con el fin de reducir el MAPE, se itera el modelo ajustando las constantes quedando con alpha=0.1, beta=0.1 y gamma=0.5.

```
=== Evaluation on training data ===
Target                              1-step-ahead  2-steps-ahead  3-steps-ahead  4-steps-ahead  5-steps-ahead  6-steps-ahead
==========================================================================================================================
ValorW
  N                                          358            357            356            355            354            353
  Mean absolute error                   142.9293       154.5541       153.3413       148.6319       135.9192       135.2831
  Mean absolute percentage error          34.301        36.4995        37.4824        37.7466        34.6899        34.0633
  Root mean squared error               396.7952       394.5088       304.2633        252.638       183.1207       178.7445
  Mean squared error                 157446.3998    155637.1731     92576.1294     63825.9462     33533.2051     31949.586

Total number of instances: 382

=== Evaluation on test data ===
Target                              1-step-ahead  2-steps-ahead  3-steps-ahead  4-steps-ahead  5-steps-ahead  6-steps-ahead
==========================================================================================================================
ValorW
  N                                          164            163            162            161            160            159
  Mean absolute error                   117.3671       125.4544       127.8357       128.4371       129.5534       134.2201
  Mean absolute percentage error         36.7463        39.1203        40.7924        42.3779         43.856        45.8884
  Root mean squared error               148.4085       158.4659       162.3638       164.7826       171.1391       179.1097
  Mean squared error                 22025.0865     25111.4269     26362.0174      27153.321     29288.5971      32080.277

Total number of instances: 164
```

Figura 16.Modelo HoltWinters optimizado alpha=0.1, beta=0.1 y gamma=0.5.

Con esta configuración se encuentra que el MAPE se reduce de 58% a 34%.

Se revisan las respuestas del modelo, al utilizar índices artificiales el gráfico por bloques deja de interpolar valores y se obtiene un modelo de entrenamiento mucho más coherente.

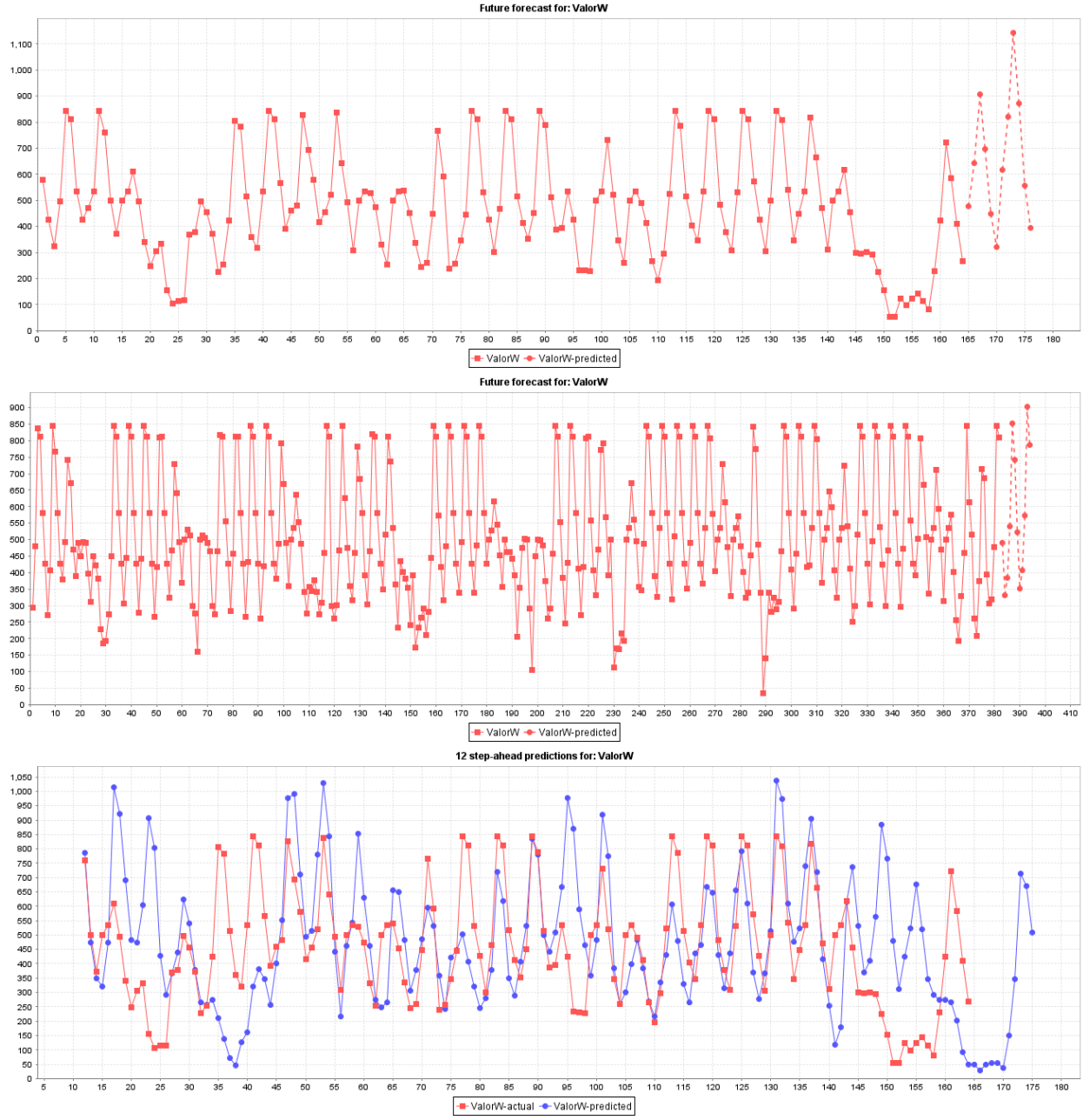

Figura 17. Modelo HoltWinters con constantes optimizadas (alpha=0.1, beta=0.1 y gamma=0.5)

Se puede notar una mejoría en el ajuste del modelo respecto a los datos de entrenamiento, paralelo se puede apreciar una predicción del próximo ciclo que asemeja más el comportamiento del conjunto de datos.

A partir de este análisis, el modelo matemático de Holt Winters se compara con otros modelos que fueron ejecutados en Excel.

- Promedio Móvil

Inicialmente el comportamiento se modela utilizando promedio móvil.

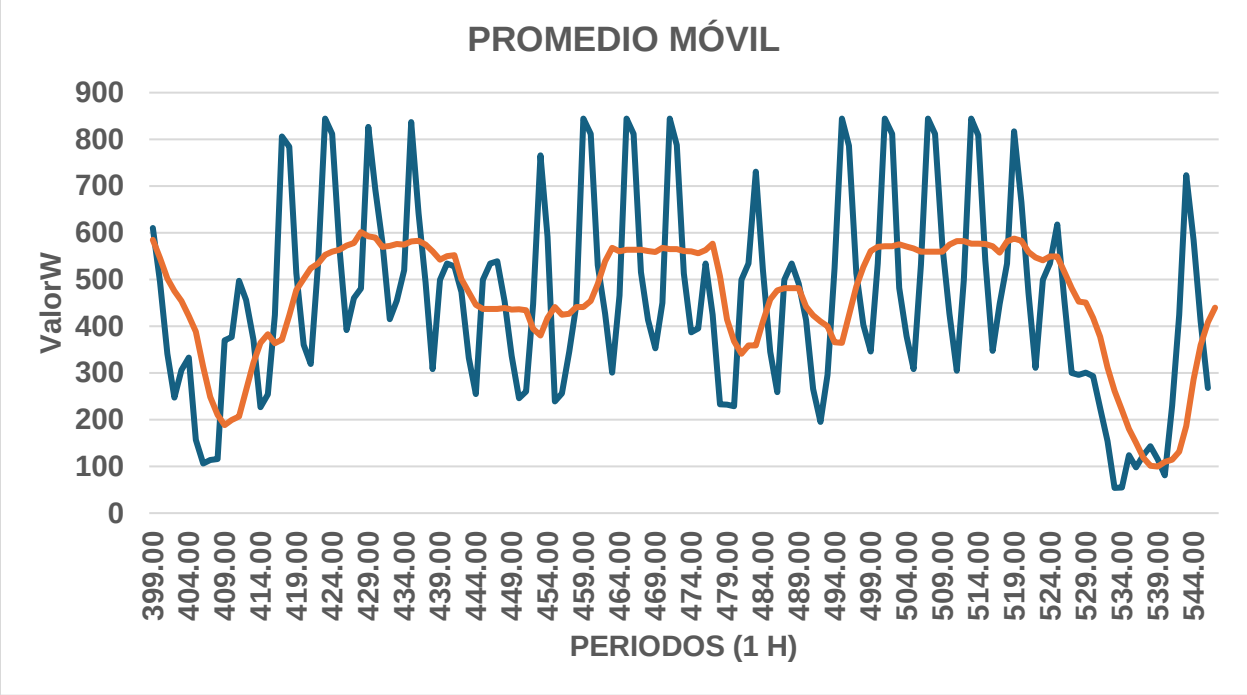

Figura 18. Promedio Móvil

El promedio móvil muestra la tendencia que describe el conjunto de datos sin embargo su principal limitante es el de pronóstico, para los siguientes valores se asume un valor constante a partir del último valor calculado.

- Suavización exponencial simple

Evaluando un tercer modelo se cuenta con la suavización exponencial simple, utilizando una constante de Alpha=0.2

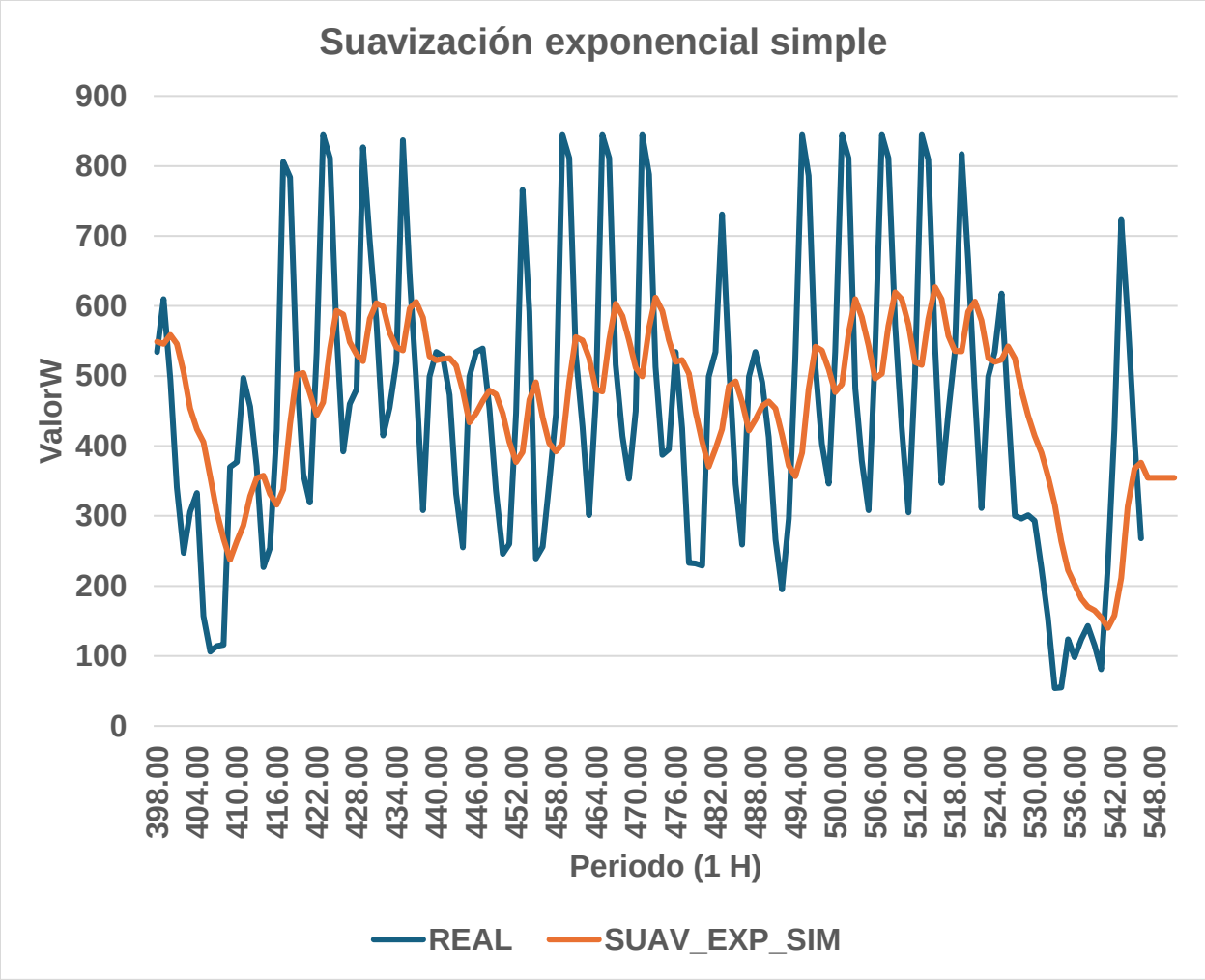

Figura 19. Suavización exponencial simple.

Para la suavización exponencial simple se cuenta con las mismas limitaciones que el promedio móvil, el último dato pronosticado se prolonga a lo largo del tiempo de predicción.

Para los 2 casos anteriores el pronóstico es realizado a corto plazo, siendo el instante siguiente, al comparar el MAD y MAPE de ambos modelos se encuentra que son muy cercanos, este resumen se explorará a detalle más adelante en esta sección.

- Modelo Holt

Para la ejecución del modelo de Holt se utilizaron las constantes Alpha=0.3 y beta=0.3.

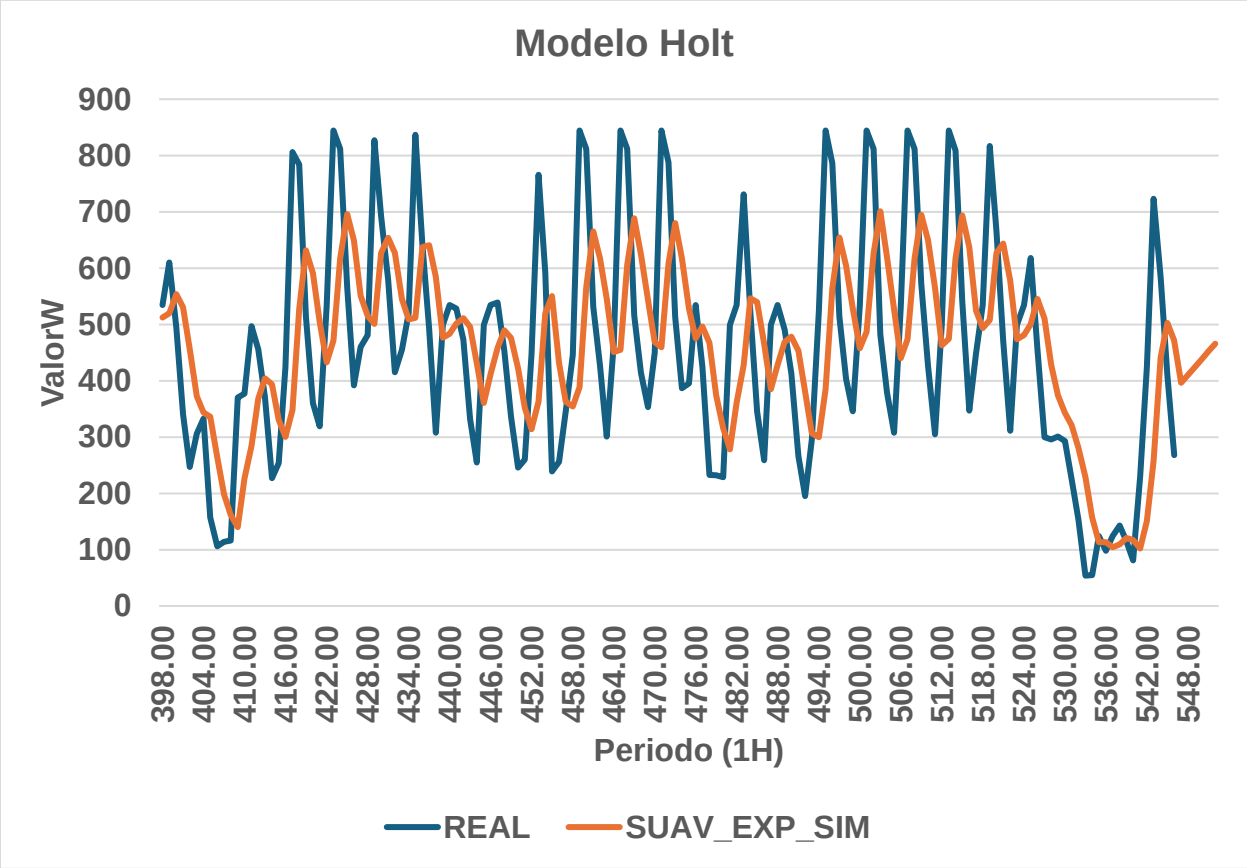

Figura 20. Modelo de Holt

Al obtener los resultados del modelo de Holt se encuentra que la curva de pronóstico se ajusta mejor que las 2 anteriores, presenta una notable mejora en el desempeño del MAPE, se revisará al final de la sección.

- Modelo de Winter

Finalmente, se ejecuta el modelo de Winter para revisar su ajuste y error a comparación de los modelos anteriormente propuestos, para este se ejecuta la regresión lineal de los

valores obtenidos en el modelo desestacionalizado donde se obtiene un nivel inicial de Lt_0=520.1 y una tendenciaT_0=(-0.09) asumiendo alpha=beta=0.1, S1=0.7210, S2=0.9358, S3=1.4281, S4=1.2990, S5=0.9339 y S6=0.6812.

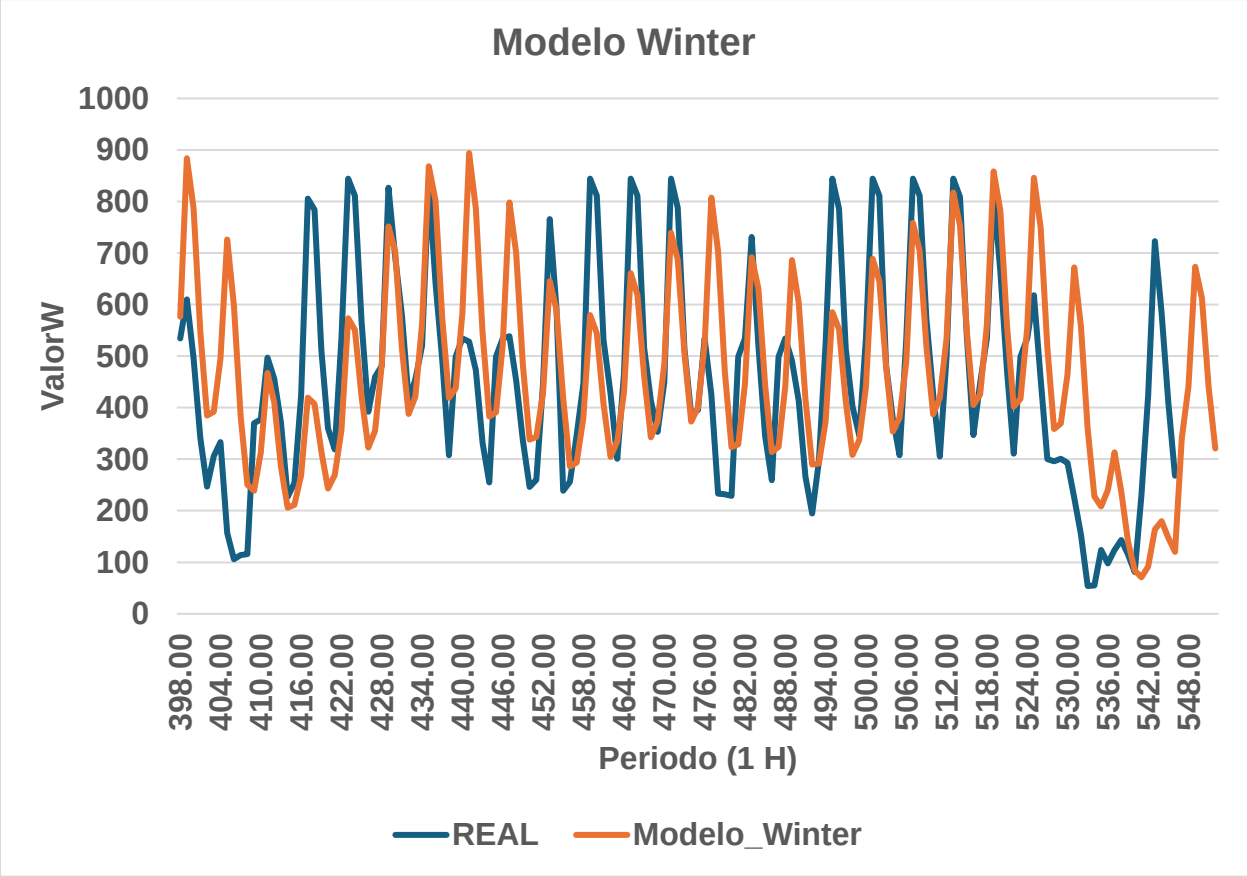

Figura 21. Modelo de Winter.

Se obtiene este modelo bastante similar al comportamiento de los datos de entrenamiento, al comparar los 5 modelos estudiados se encuentra lo siguiente.

Tabla 1. Comparación de los errores para los diferentes modelos.

MÉTODO	MAD	MAPE (%)	RANGO TS INF	RANGO TS SUP
PROMEDIO MÓVIL (EXCEL)	151	40	-208.71	5353.03
SUAVISACIÓN EXPONENCIAL (EXCEL)	156	41	-4.63	11.42
MODELO DE HOLT (EXCEL)	161	39	-2.91	6.20
MODELO DE WINTER (EXCEL)	114	33	-12.38	22.64
MODELO HOLT-WINTER (WEKA)	135	34		

Al probar el conjunto de datos en los diferentes modelos predictivos se encuentra que el modelo de Winter arrojó el MAD más bajo con 114, mientras que el modelo de Holt arroja 161, por otra parte, al comparar MAPE (%) más alto es para la suavización exponencial simple con 41% y un 33% para el modelo de Winter siendo el error más bajo de los modelos explorados.

Conclusiones

El promedio móvil y la suavización exponencial simple ayudan a modelar el comportamiento de un conjunto de datos sin embargo su capacidad de predicción se encuentran limitadas, el último valor es el correspondiente a la predicción en ambos casos.

Los modelos predictivos de Holt son más robustos que los anteriormente mencionados pero muestra un pronóstico poco preciso para modelos con mucha variabilidad, es sensible a valores atípicos.

Al ejecutar los modelos de winters y Holt-Winters se encontró un modelo que muestra mejor ajuste y menor error, para ambos modelos el intervalo de predicción arrojó un comportamiento más acorde al comportamiento histórico.

El software Weka requiere de base de datos que cuenten con las fechas con el formato específico “yyyy-MM-dd HH:mm:ss”, de cierta forma dependiendo de la orientación del análisis se deben definir claramente las configuraciones de la periocidad, para modelos que no cumplen con un intervalo se deben utilizar índices artificiales.

Los datos atípicos no siempre son mediciones erróneas, en ocasiones son valores influidos por otros factores externos y según su origen se deben manejar adecuadamente.

Recomendaciones y futuros estudios

- Para un análisis más detallado, segregar los conjuntos de datos en diferentes estaciones y estudiar el comportamiento horario para días con condiciones similares.
- Optimizar los modelos matemáticos con el solver de Excel.
- Utilizar otros modelos predictivos como ARIMA, SARIMA y probar modelos predictivos multivariados.
- Incluir la probabilidad de ocurrencia según la temperatura, hora del día y estación.

Estudios futuros incluye contar con las cantidades de alquiler de bicicletas por estación y analizar el flujo según horario, cuáles son las estaciones que tienen mayor inventario y cuales tienen inventarios bajos.

Cantidad de tiempo promedio que las bicicletas están fuera de las estaciones y contabilizar la cantidad de unidades que no están disponibles durante un intervalo de tiempo.

Análisis de comportamiento en el transportan cuales son las más concurridas y estudiar la probabilidad de que un usuario en la estación A termine su recorrido en las estaciones B, C, D… etc.

Bibliografía

[1] *capital bikeshares*. (2025, 04 02). Récupéré sur capital bikeshares: https://ride.capitalbikeshare.com/about

[2] Chopra, S., & Meindl, P. (2013). *Administración de la cadena de suministro estrat*. México: PEARSON.

[3] Fanaee-T, H., & Gama, J. (2013). Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, pp. 1-15. doi:doi:10.1007/s13748-013-0040-3

[4] Karunanithi, M., Chatasawapreeda, P., & Ali Khan, T. (2024). Un enfoque de análisis predictivo para pronosticar la demanda de alquiler de bicicletas. *ELSEVIER*.

[5] (2022). *The Meddin Bike-sharing World Map Report*. PBSC Urban Solutions.

Anexos

[1] Repositorio en GitHub, Modelo predictivo de la demanda de bicicletas según horario en Capital Bikeshares DC, enlace: https://github.com/ASANJUR/Modelo-predictivo-de-la-demanda-de-bicicletas-seg-n-horario-en-Capital-Bikeshares-DC

Base de datos completa, extracto de la base de datos, base de datos en formato ARFF, código en Jupyter Notebook.