Recorridos realizados EcoBici

2020 - 2021

Gastón Barisani, Augusto San Martín y Ángel Rodriguez
Universidad Tecnológica Nacional - Ciencia de Datos - Cluster AI

1 Introducción

El sistema de bicicletas público de la Ciudad Autónoma de Buenos Aires (Argentina) ha tenido un gran crecimiento a partir de la pandemia por el covid-19, ya que es un medio de transporte individual y al aire libre, favoreciendo el cuidado personal. Por este motivo se buscó comprender mejor cómo es utilizado el sistema hoy en día, los efectos de la cuarentena y la posterior vuelta a la "normalidad", recorridos más frecuentes y otros datos interesantes. Además, se buscará predecir la duración de los viajes realizados, de manera que los usuarios puedan tener un estimativo de en cuánto tiempo llegará la próxima bicicleta a una determinada estación, dato de suma importancia cuando una estación no tiene bicicletas disponibles. A su vez, este análisis será útil para poder realizar una mayor segmentación de las estaciones, reduciendo la oferta de bicicletas en las estaciones menos utilizadas y relocalizando las en aquellas estaciones más demandadas.

2 Aplicación

Muchas veces se sufrió en carne propia el estar en una estación de bicicletas y que no haya ninguna disponible, sin saber si continuar con otro medio de transporte o esperar a que arribe una bicicleta a la estación. De esta manera, si el usuario tiene un rumbo fijo hacia una estación, se podría predecir el tiempo que tardará éste en llegar. El usuario sólo deberá, al momento de retirar la bicicleta por la aplicación, indicar la estación a la cual se dirigirá para permitir realizar la predicción tanto para los usuarios que podrían estar esperando que llegue una bicicleta, como para el ciclista que debe contar con un lugar de anclaje disponible para dejar la bicicleta una vez finalizado el viaje.

3 Datasets

Se utilizaron tres datasets para el desarrollo del trabajo:

- Recorridos realizados 2020
- Recorridos realizados 2021 (disponible hasta el mes de junio)
- Ubicación de las estaciones

3.1 Variables utilizadas

Para los recorridos realizados se utilizaron las siguientes 5 variables principales:

- Duración del viaje
- Fecha de inicio de viaje
- Fecha de fin de viaje
- Nombre de estación de inicio
- Nombre de estación de fin

Para las ubicaciones de las estaciones se utilizaron las siguientes variables:

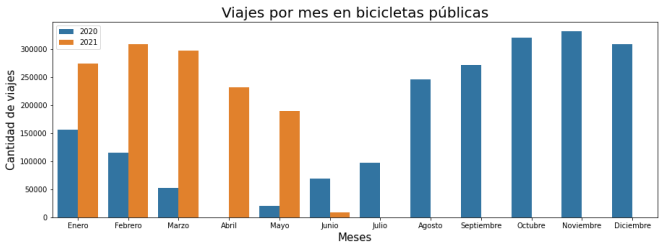
- Ubicación geográfica (longitud y latitud)
- Nombre de estación

3.2 Variables fundamentales

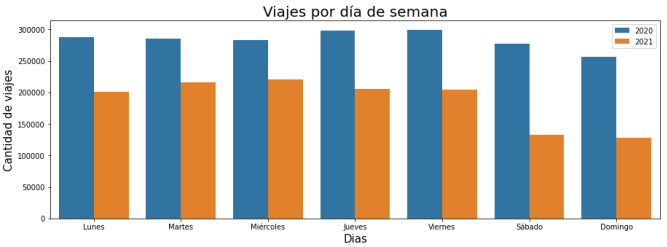
A partir de los datos mencionados, se unieron los datasets y se calculó la distancia recorrida por viaje gracias a las coordenadas de inicio y fin. Esto será de gran utilidad para poder estimar la duración de los viajes. Cabe destacar que se busca predecir los viajes de un punto hacia otro sin paradas, ya que por ejemplo una persona puede tomar una bicicleta y pedalear hasta un parque para reposarse a descansar en el mismo, realizando una distancia recorrida que poco tiene que ver con el tiempo utilizado. Por este motivo, para realizar el aprendizaje supervisado, se segmentaron los datos para velocidades (calculadas con la duración del recorrido y la distancia) mayores a 0,15 km/min, que sería aproximadamente una cuadra y media por minuto (un paseo muy relajado y teniendo en cuenta también tráfico o semáforos).

3.3 Análisis exploratorio de datos

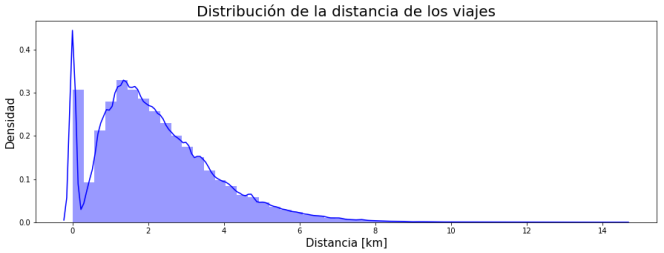
Observamos la cantidad de viajes por mes distinguido por el año en cuestión:



También podemos ver el uso según el día de la semana, decreciendo hacia el final de la misma:

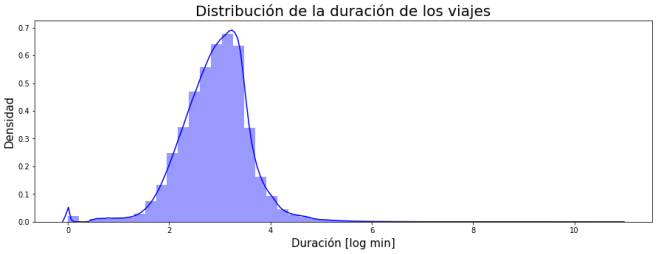


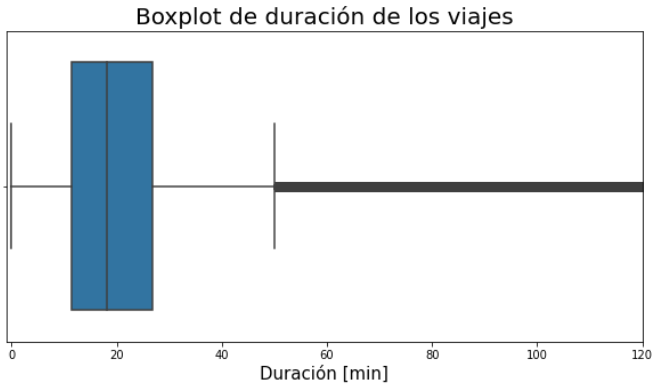
Podemos observar la distribución de los viajes:



Hay un detalle a tener en cuenta en este último gráfico y es la gran cantidad de viajes con duración igual a cero, que se debe filtrar para realizar un correcto análisis.

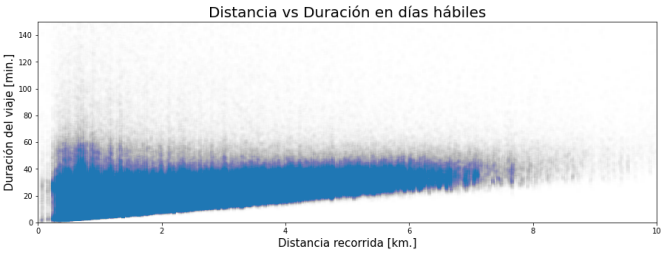
Podemos observar la distribución y un boxplot de la duración de los viajes:

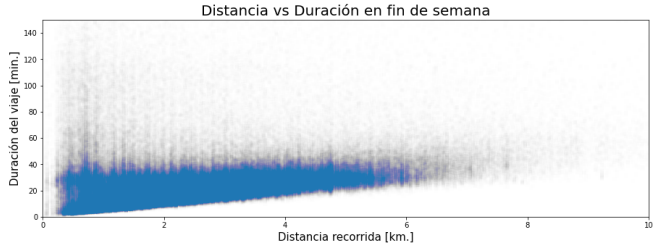




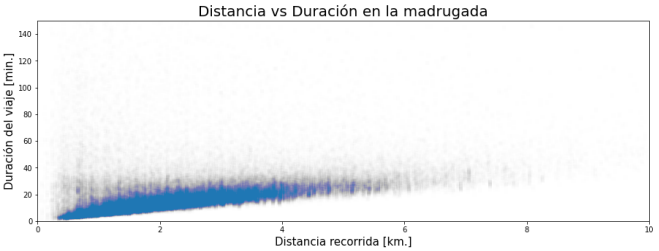
La duración media de los viajes es de 22 minutos.

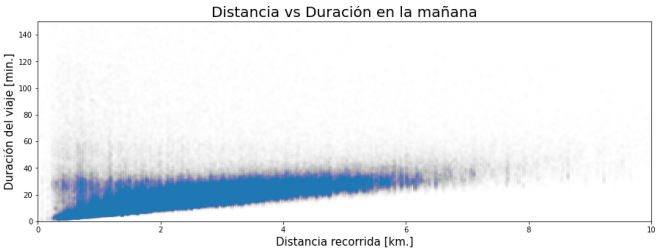
A diferencia de lo que se creía inicialmente, en los fines de semana se puede observar una tendencia más lineal y con una menor cantidad de viajes cortos de larga duración:

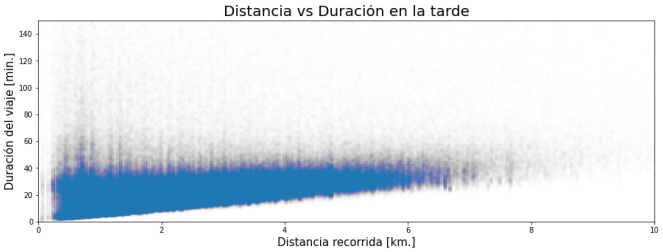


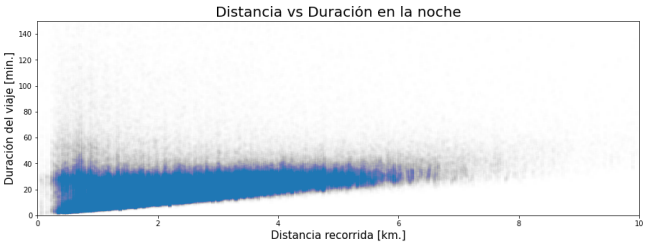


También se pueden observar distintos usos según la hora:

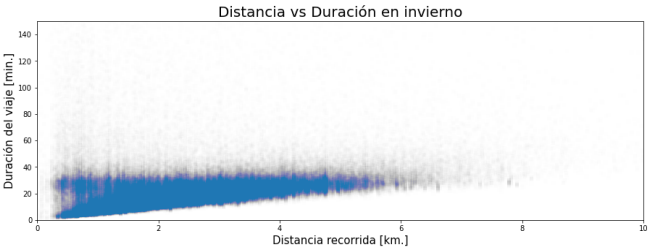


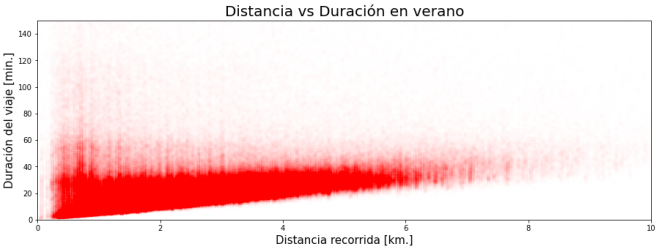




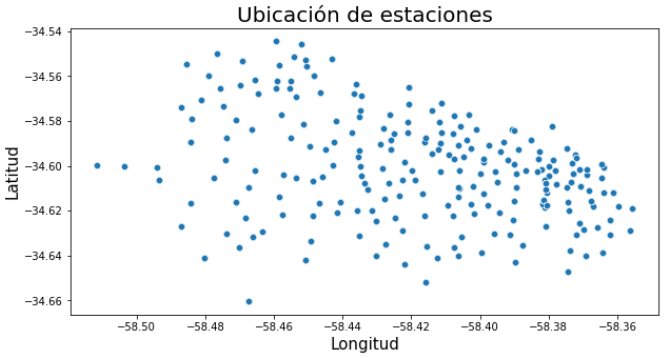


Se realizó la comparación del uso en invierno y verano:





Se imprimieron las ubicaciones de las estaciones:



Se puede observar una mayor concentración de estaciones en el centro de la Capital Federal.

Las estaciones más utilizadas, tanto para iniciar como para terminar los viajes, son:

- `GODOY CRUZ Y LIBERTADOR`
- `PACIFICO`
- `PARQUE CENTENARIO`
- `PLAZA BOLIVIA`
- `BARRANCAS DE BELGRANO`

3.4 Procesamiento de datos

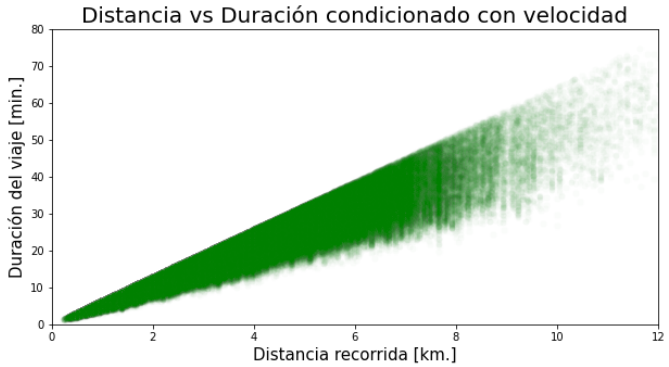
Se utilizaron para la predicción de la duración del viaje tres variables:

- Distancia recorrida
- Mes
- Hora

La variable más significativa y que tiene una gran correlación con la duración del viaje es la distancia recorrida. Por otro lado se añadieron los datos del mes y hora, ya que se pudieron observar distintos comportamientos de los usuarios según la hora del día y el clima (evaluado a través del mes).

Se utilizó el 80% de los datos para el entrenamiento y el 20% restante para validación y test.

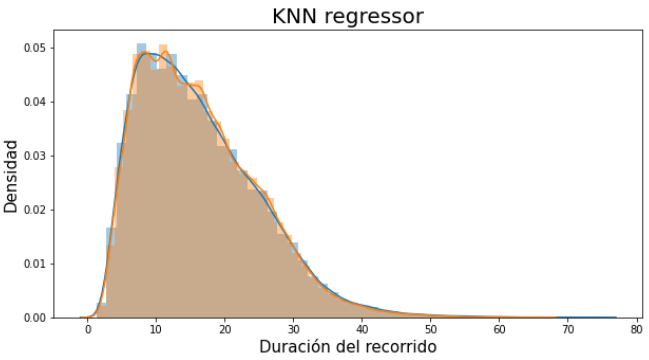
En el siguiente gráfico se puede observar la duración del viaje dependiendo de la distancia recorrida, ajustado por una velocidad mínima para poder quedarnos solamente con los recorridos de una estación hacia otra sin interrupciones.



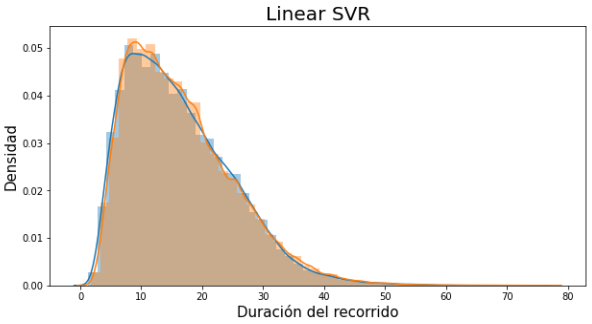
4 Modelos

Se utilizaron dos modelos para realizar la predicción:

- **KNN regressor**: se obtuvo un coeficiente de determinación de `0.9227` y un error cuadrático medio de `2.3926`.



- **Linear SVR (super vector regresor):** se obtuvo un coeficiente de determinación de `0.9246` y un error cuadrático medio de `2.4025`, además de un mejor rendimiento en cuanto al tiempo de entrenamiento.



5 Regulación de hiperparámetros

Se buscó la mejor combinación de hiperparámetros para poder llegar al mejor modelo posible, y este fue gracias al Linear SVR con una tolerancia de `1e-5` (explicado y graficado en el punto 4). A partir de ese valor, si se continuaba aumentando el exponente (-6, -7, etc.) el modelo prácticamente no sufría modificaciones, mientras que si se disminuía (-4, -3, etc.) el modelo empeoraba.

También se ajustaron los parámetros del KNN regressor, sin obtener mejoras en el resultado, solo un mayor tiempo de entrenamiento.

6 Conclusiones

Se pudo comprender mejor el uso de las bicicletas, pudiendo dividirlo en dos grandes grupos: los paseos recreativos (poco recorrido para el tiempo de utilización, o inicio y fin en una misma estación) y los viajes de rutina diarios como por ejemplo para ir al trabajo o a la facultad.

Se pudo notar una clara estacionalidad en el uso de las bicicletas, con un claro descenso en época invernal y diferentes utilizaciones dependiendo de la hora del día.

7 Referencias

https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVR.html

https://www.analyticslane.com/2018/07/20/visualizacion-de-datos-con-seaborn/

Python for Data Analysis by Wes McKinney (O'Reilly)