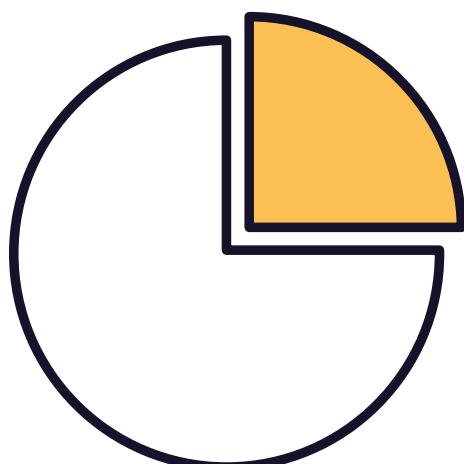


Proyecto Final

DATA SCIENCE - CODERHOUSE
DIEGO BOUZADA - MATÍAS PARIENTE - LUCAS HERNÁNDEZ



CODER HOUSE



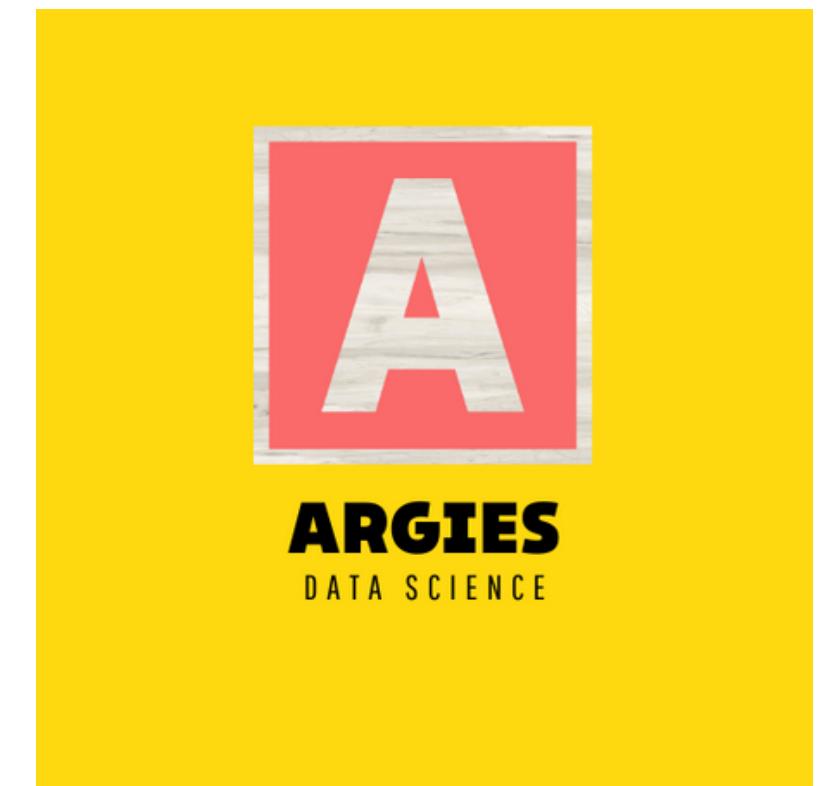
Índice

Páginas - Contenido

- 2 - Índice
- 3 - Presentación Argies
- 4 - Equipo
- 5 - ¿ Que es EcoBici?
- 6 - Mapa de Estaciones
- 7 - Investigación
- 8 - Trabajo sobre el Dataset
- 9 - Variables
- 10 - Análisis de Variables
- 11 - Subset del conjunto de datos
- 12 a 16 - Modelos
- 17 - Conclusiones y futuras líneas



¡Hola!, somos Argies



Somos ARGIES Data Science, nuestra misión es acompañar a nuestros clientes en el aprovechamiento de sus datos para que puedan optimizar su funcionamiento y lograr un mejor desempeño en todas sus áreas.

Equipo



Matías Pariente

Soy Ingeniero en Electronica, Vivo en Ciudad de Buenos Aires. Hincha y socio de Boca. Mediocampista de marca y quite.

Tutor:
Matías Souto



Diego Bouzada

Soy Licenciado en Comunicación Social y Data Analyst. Vivo en Paraná, Entre Ríos, Soy Fana de River Plate y me gusta hacer deportes.

Profesor:
Jaime Fraustro Valdez



Lucas Hernández

Soy Estudiante de Economía, Data Analyst, hincha de Huracán y fanático del futbol manager.



Eco Bici

¿Que es EcoBici?

Ecobici es el sistema de transporte público de bicicletas compartidas que desarrolló el Gobierno de la Ciudad de Buenos Aires, brindando a quienes se mueven en la Ciudad una nueva alternativa de transporte. La operación de BA Ecobici es gestionada por Tembici, una startup brasileña de movilidad urbana.

¿Cómo se utiliza?

Para poder utilizar el sistema es necesario que cuentes con la última versión de la aplicación BA Ecobici por Tembici que está disponible tanto para Android como para iOS en forma gratuita. Este registro es únicamente con tarjeta de crédito internacional. Si no contás con tarjeta de crédito podés inscribirte de forma presencial (sacando turno previo a través de la web BA Ecobici) en las oficinas de Ecobici por Tembici en Capital Federal.

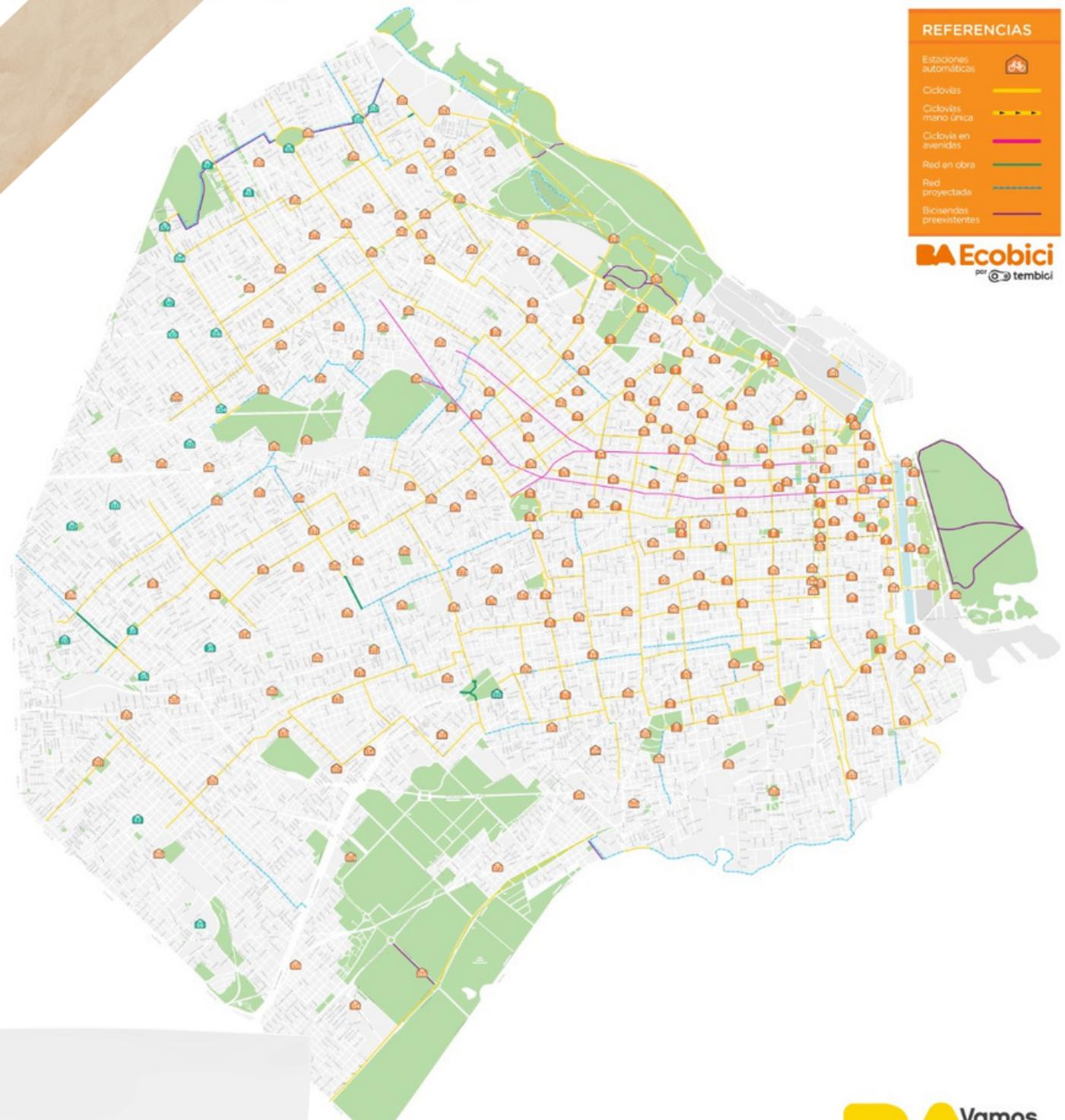
Ecobici es gratuito para todas las personas residentes del país de lunes a viernes (días hábiles) con hasta cuatro viajes de 30 minutos cada uno.

Si querés utilizar las bicis los fines de semana o pedalear por más tiempo, podés conocer los nuevos pases que ya están habilitados. Si sos turista extranjero y querés usar Ecobici tenés que obtener un pase turístico exclusivo.

Eco Bici

Estaciones

En los ultimos días se anexaron 40 estaciones nuevas (y 400 bicicletas) alcanzando los 320 puestos dentro de la ciudad de Buenos Aires



Problema y Objetivos

Nuestro objetivo principal es proponerle al cliente un modelo de predicción sobre el tiempo de recorrido por usuario entre la estación de origen y la de destino.

Además, como objetivos secundarios, nos interesa responder:

- ✓ ¿Cuáles son las estaciones más utilizadas?
- ✓ ¿Qué rasgos tienen los clientes habituales? Edad, Sexo.

Datos

Los datos fueron recolectadas de la pagina del gobierno de la ciudad, el link de dicha fuente es el siguiente:

<https://data.buenosaires.gob.ar/dataset/bicicletas-publicas>

Fecha de publicación 10 de Mayo de 2021 - Fecha de actualización 10 de Diciembre de 2021



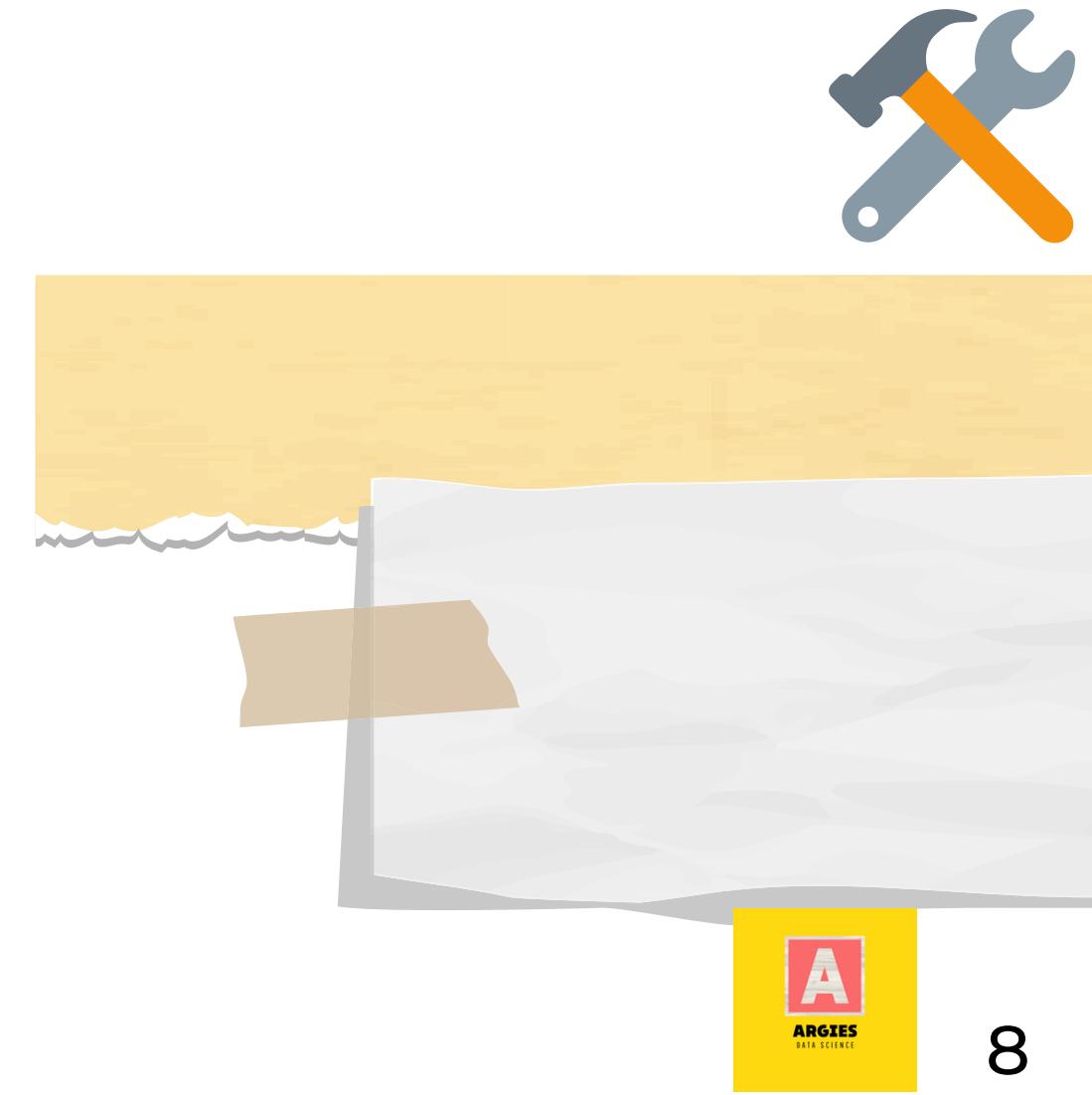
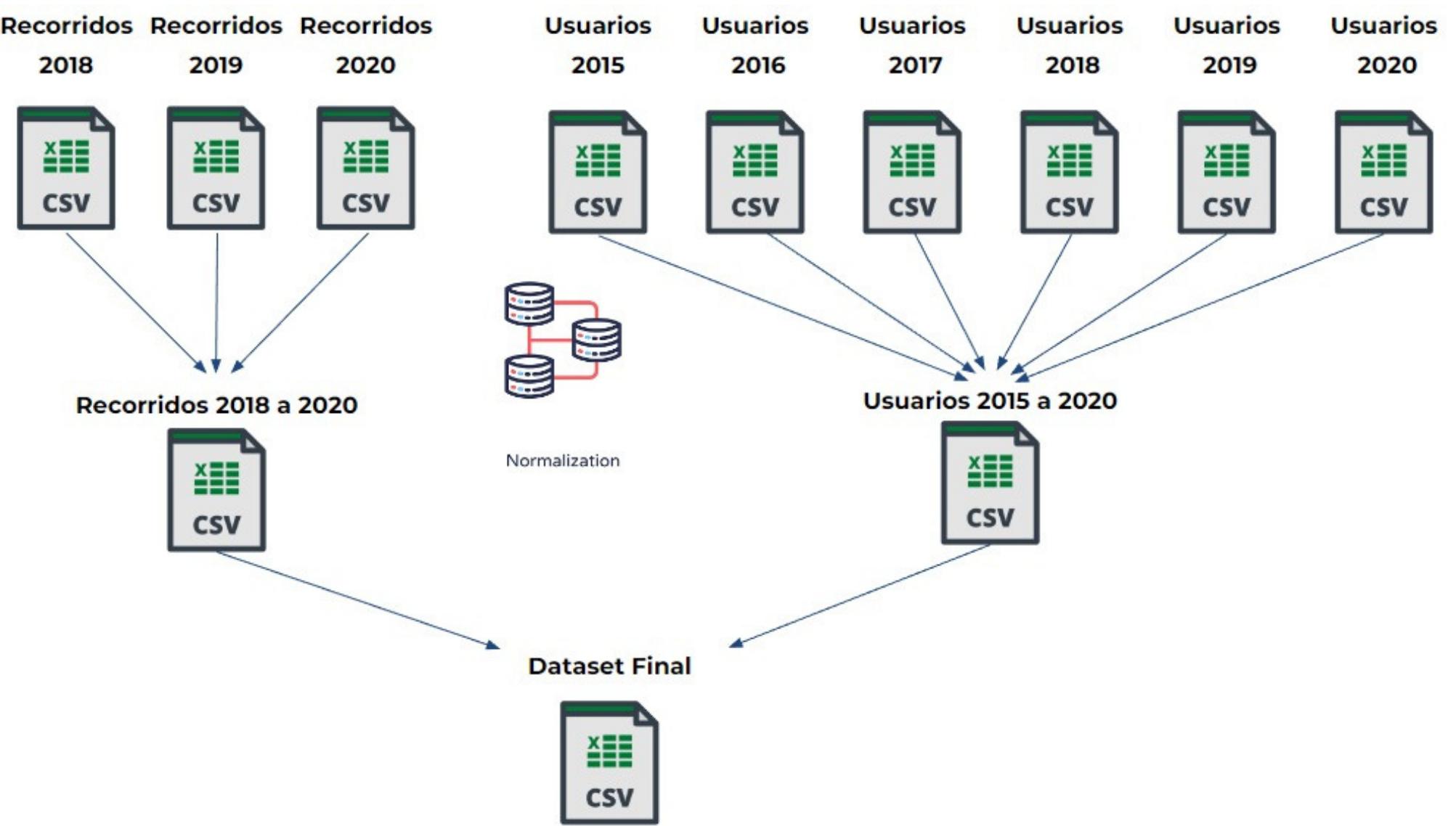
Investigación



Trabajo sobre el Dataset

El proceso para la creación del dataset final, consto de varios pasos realizados en Python, los datos estaban en formatos distintos y separados por año.

- * Se cargaron los datasets de Usuarios desde el 2015 al 2020.
- * Se cargaron los dataset de recorridos desde el 2018 al 2020.
- * Se normalizaron todos los datos de las tablas, para luego unirlos y generar el dataset a utilizar.



Variables

Listado de variables del Dataset

id_usuario: número de cada usuario registrado

fecha_origen_recorrido: fecha y hora en que inicia el recorrido

id_estacion_origen: número de identificación de la estación de origen

nombre_estacion_origen: nombre de la estación de origen

fecha_destino_recorrido: fecha y hora en que finaliza el recorrido

id_estacion_destino: número de identificación de la estación de destino

nombre_estacion_destino: nombre de la estación de destino

genero_usuario: Género del usuario

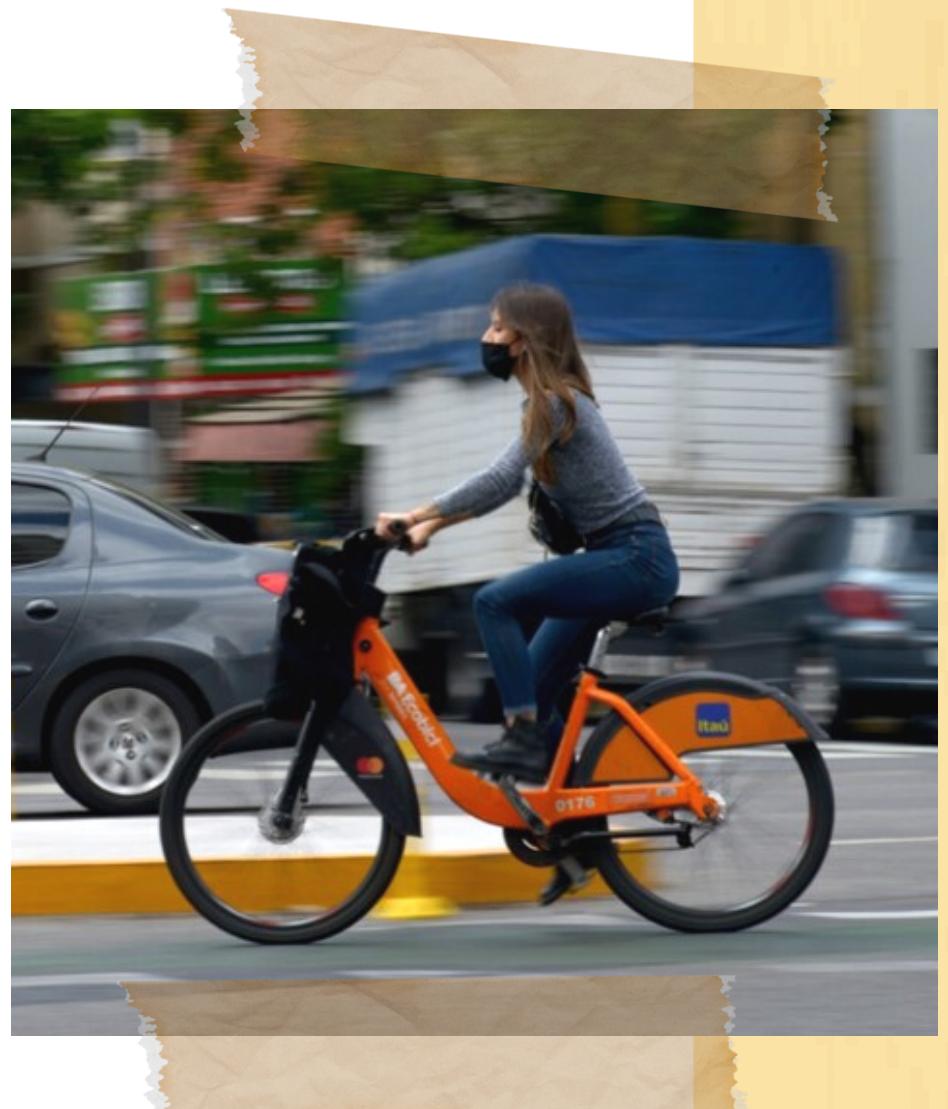
edad_usuario: Edad del usuario

tiempo_recorrido: Duración del recorrido

Este data set tiene 4.176.885 filas, y 10 columnas

Limpieza del Dataset

Luego de revisar la ausencia de valores nulos, detectamos valores atípicos en las variables Edad de Usuario y Tiempo Recorrido, eliminando los mismos ya que en cantidad eran muy pocos en relación al tamaño del dataset



Análisis de Variables



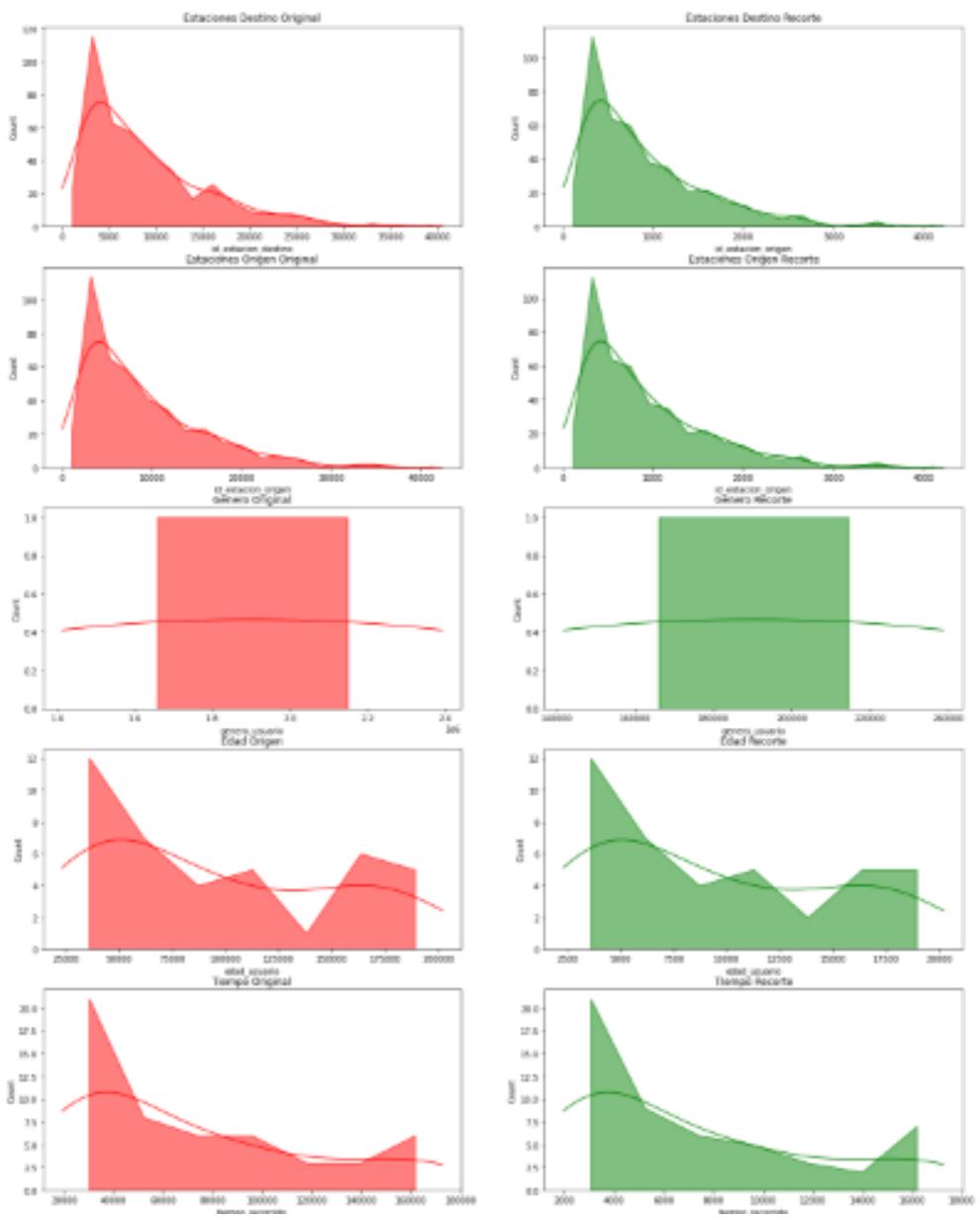
Relación entre variables

En nuestro análisis entre variables no logramos establecer relaciones lineales entre dos o más variables. Nos generó inconvenientes la cantidad de estaciones que tiene EcoBici (mas de 300) por lo que hicimos un recorte de las cuatro estaciones mas utilizadas para investigar relaciones entre variables.

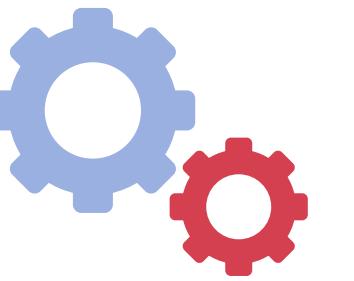
- *El 63 % de los usuarios son varones*
- *Los usuarios son principalmente jóvenes entre 20 y 30 años, los cuales a su vez tienen un tiempo de recorrido mayor.*
- *La estación mas utilizada por los varones es "Pacífico", las mujeres utilizan mas la estación "Parque Centenario".*
- *La relación máxima entre variables es entre estacion de origen y destino (0.61) lo que indica que mas de la mitad de los viajes salen y regresan en la misma estación*

Subset del Conjunto de datos

Nuestro dataset supera los 3 millones de registros, al momento de probar los algoritmos nos encontrábamos con errores de memoria y tiempos que superaban las 4 horas para ver los resultados. Entonces realizamos un subset del dataset original al 10%. En el siguiente gráfico se puede ver la como cada una de las variables tiene un comportamiento similar en el set original, como en el subset, esto lo realizamos para medir la representatividad del mismo



Modelos



Nuestro objetivo es predecir el tiempo de recorrido (la duración de cada viaje) teniendo en cuenta las variables de Estación de Origen, Estación de Destino, Edad y Género. Al ser una variable cuantitativa numérica, pensamos establecer un modelo de regresión, pero no encontramos relaciones entre las variables que nos permitan hacer ese modelo.

Por consiguiente categorizamos la variable en dos, tomando la mediana de "tiempo recorrido" (17 minutos,) y separando en dos categorías, viaje corto y viaje largo.

Nuestro Target (Y)

Para los modelos de predicción usamos la variable target de tiempo de recorrido, dividimos la duración de los viajes en 2 categorías:

- Viaje Corto (todos los viajes menores a 17 minutos)
- Viaje Largo (todos los viajes mayores a 17 minutos)

Variables para formar (X)

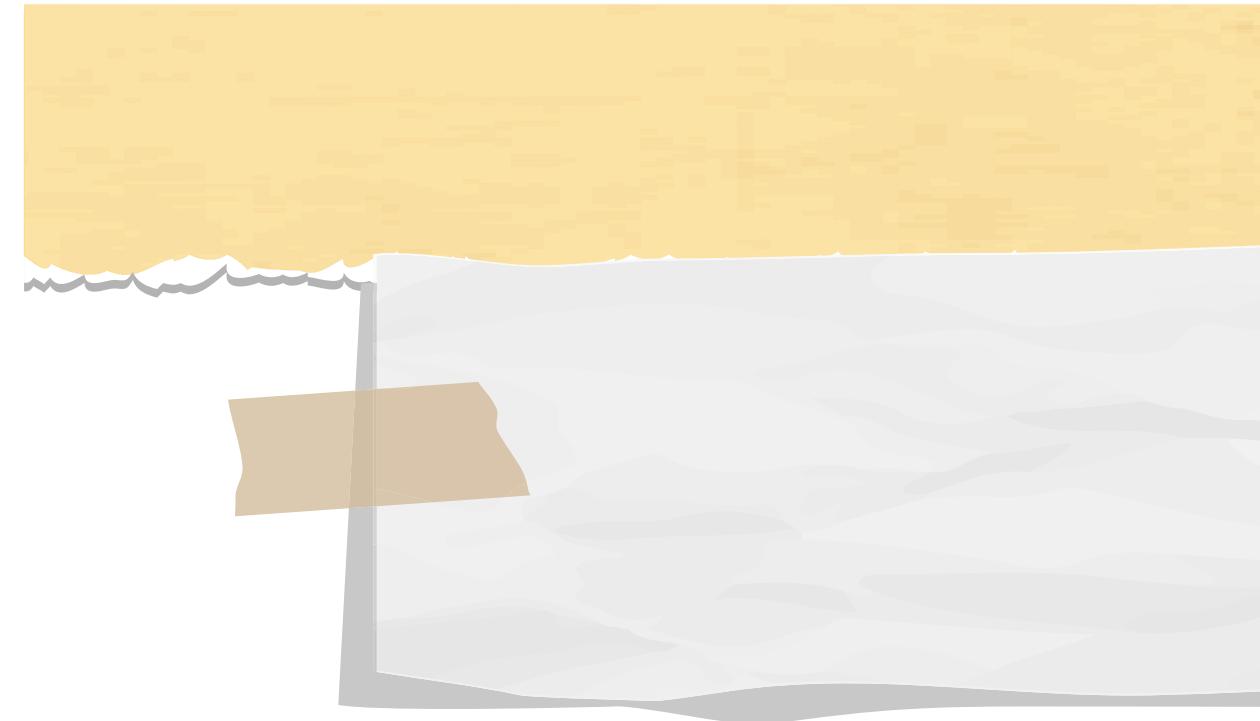
Usaremos para predecir el tiempo de recorrido, las variables de estación origen, estación destino, género y edad.



Modelo Arbol de Decisión

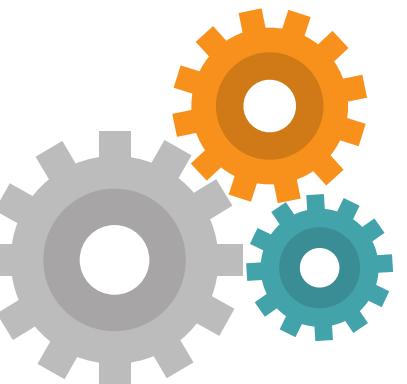


Modelos

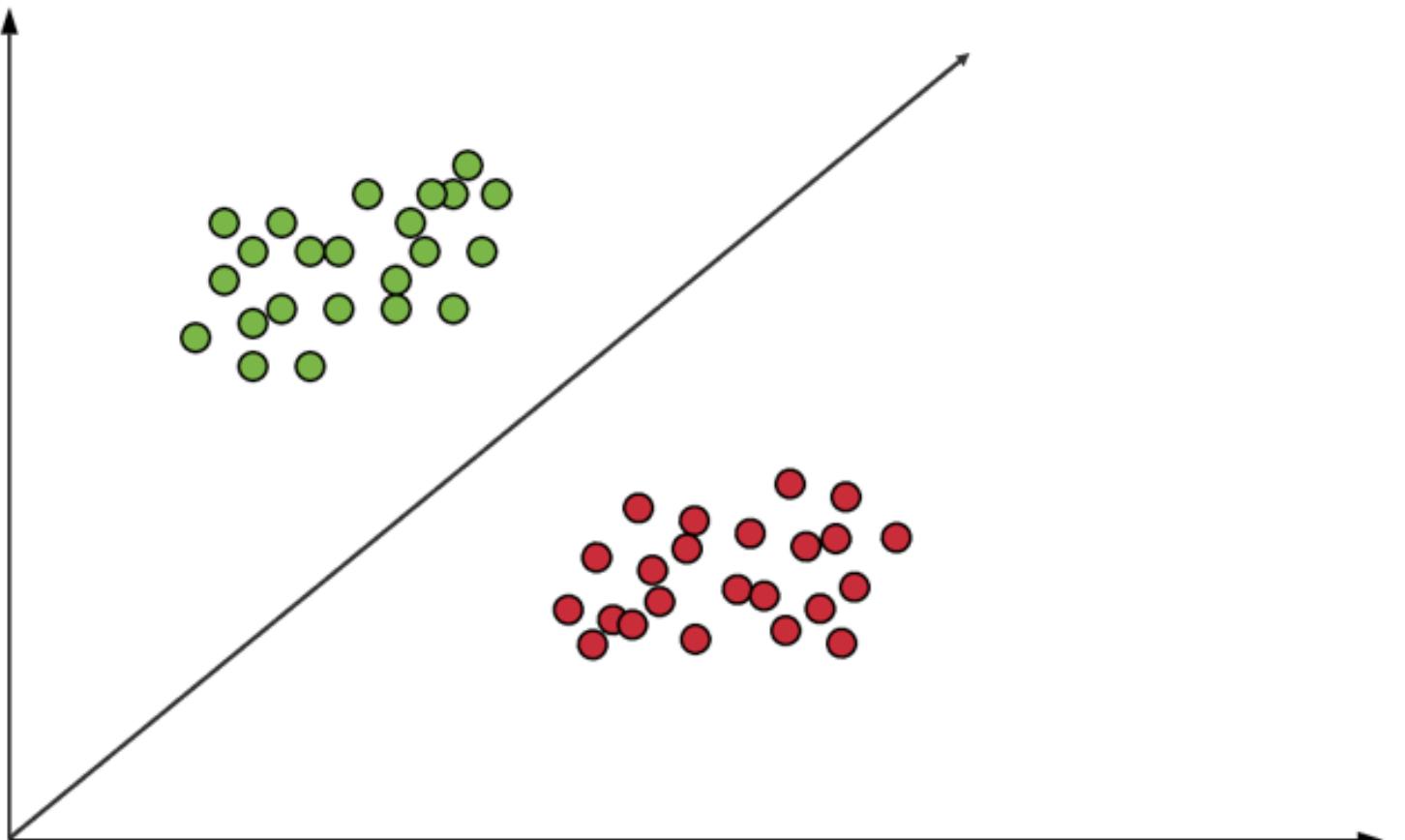


Probamos el modelo de árbol de decisión, iteramos manualmente sobre la profundidad del mismo, hasta obtener el mejor parametro con el método de gridsearch.

No conseguimos buenas métricas.
Obtuvimos un accuracy de: 0.56

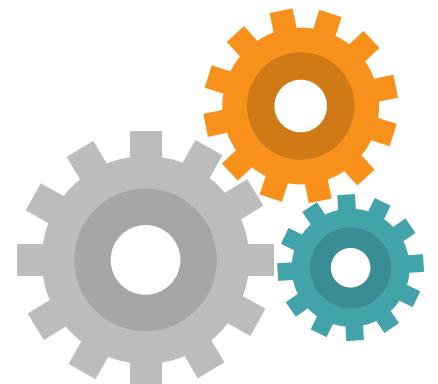
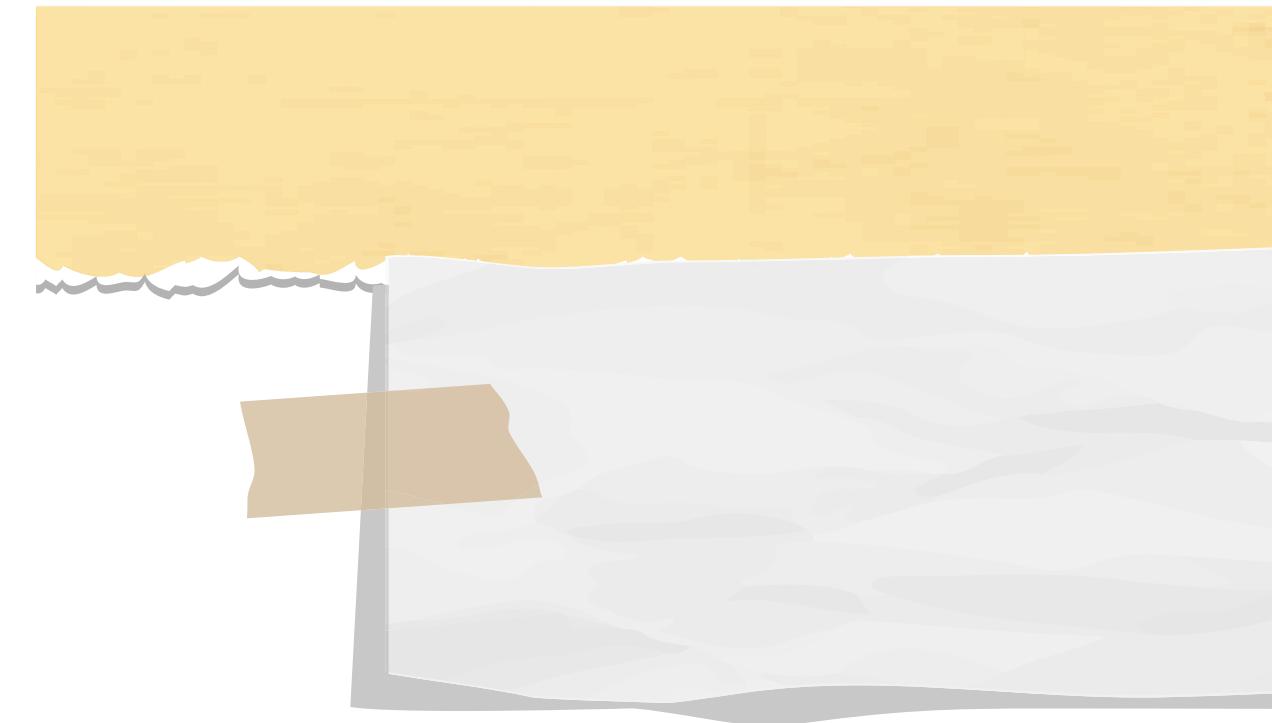


Modelo Support Vector Classifier



Probamos este modelo de clasificación como alternativa al KNN (no logramos usarlo por el costo computacional), iteramos manualmente el parámetro "C" y optimizamos el mismo con el metodo gridsearch. Obtuvimos un accuracy de 0.58

Modelos



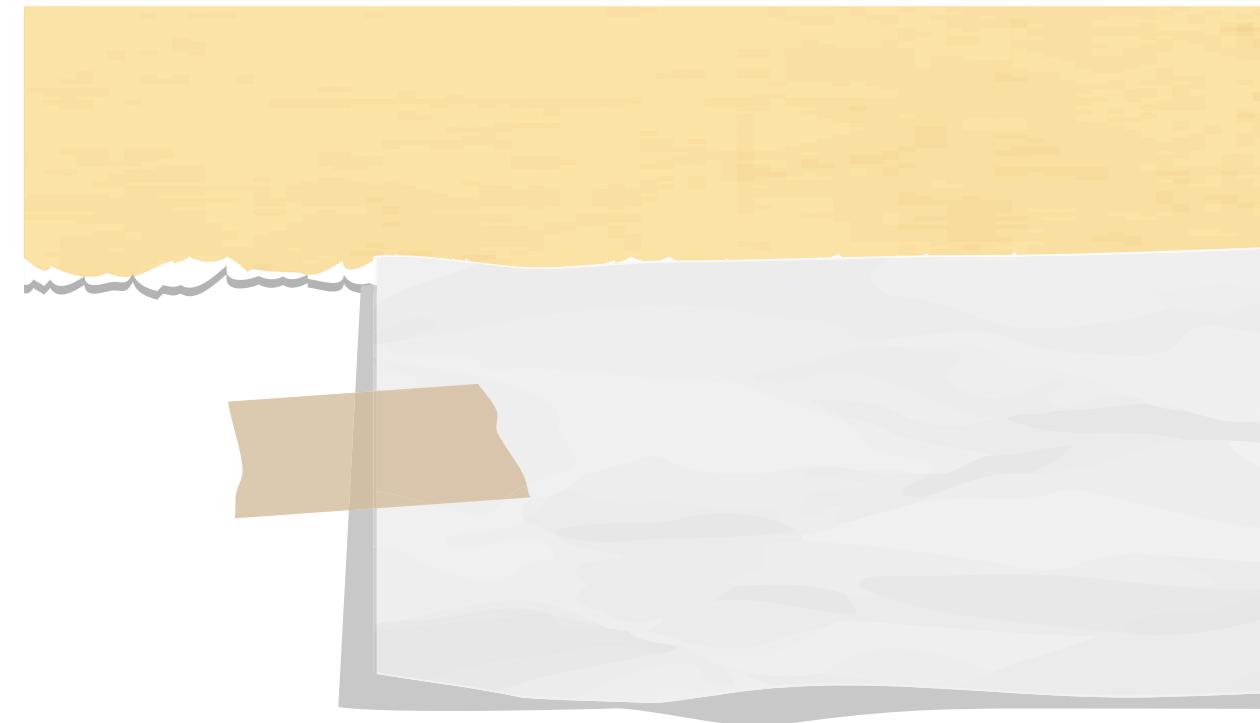
Modelo Random Forest



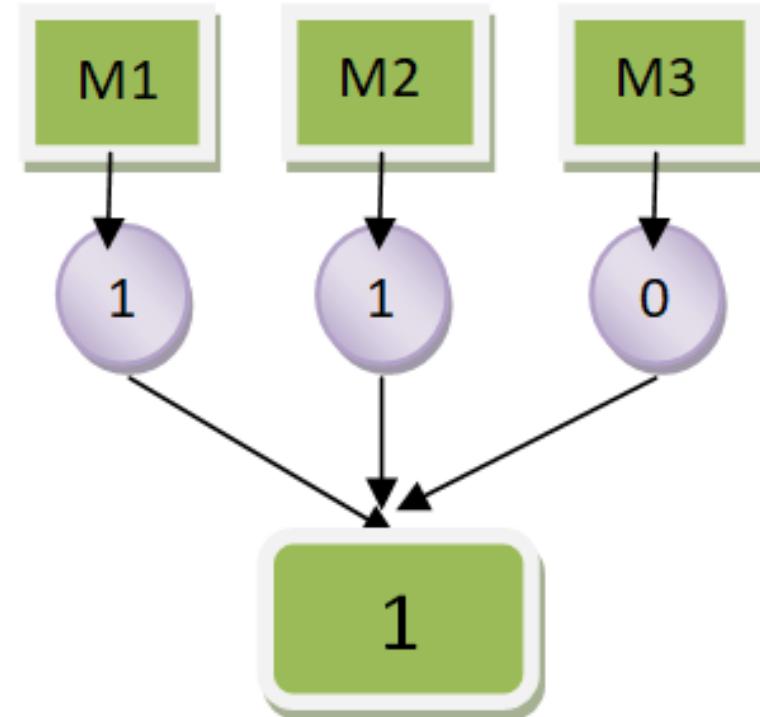
Random Forest

Por ultimo utilizamos el modelo de Random Forest, tambien optimizando los hiperparametros con el mismo método de gridsearch, obteniendo un accuracy de 0.54

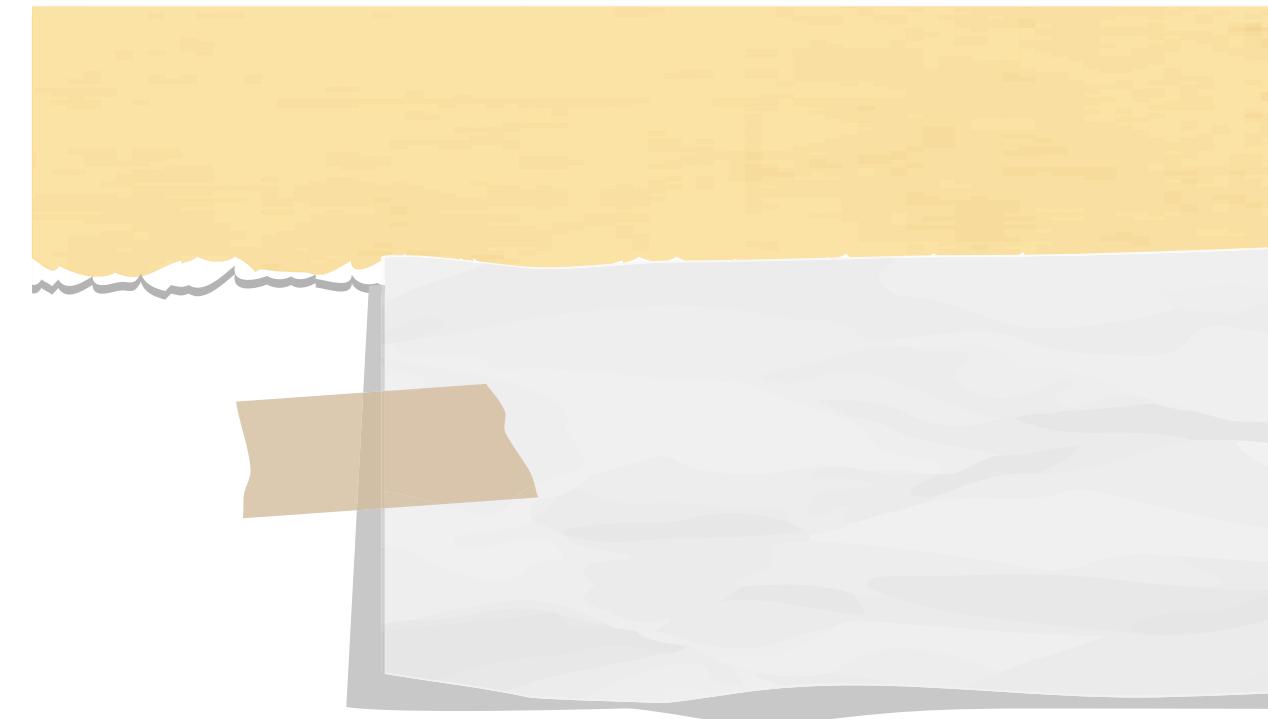
Modelos



Ensamble de modelos



Modelos



Para obtener mejores métricas probamos el ensamble de modelos mediante el método de Voting Classifier, el cual nos arrojó un accuracy del 0.56

Intentamos utilizar el método de Boosting, pero no logramos correrlo debido a su gran costo computacional



Conclusiones

Con el enfoque planteado de separar en 2 categorías el tiempo de recorrido, los datos que tenemos disponibles y las optimizaciones aplicadas a no logramos establecer buenas predicciones con los modelos utilizados.

A su vez respondiendo los objetivos secundarios, logramos determinar que las cuatro estaciones mas utilizadas son: Parque Centenario, Pacífico, Parque Las Heras y Plaza Italia y que el porcentaje de uso por género es 63,2% masculino y el 36,8% femenino

Como futuras líneas para lograr mejores resultados se podría intentar sumando variables al análisis como puede ser el estado del clima, el día de la semana, y los horarios de los recorridos.

