INSTITUTO TECNOLÓGICO DE BUENOS AIRES – ITBA

ESCUELA DE (INGENIERÍA Y TECNOLOGÍA – INGENIERÍA Y GESTIÓN - POSTGRADO)

Análisis de la Utilización de Taxis en la Ciudad de Buenos Aires

AUTOR/ES: Pugliese Franco (Leg. Nº 103777)

DOCENTE/S TITULAR/ES O TUTOR/ES: Aizemberg, Diego Ariel

TRABAJO FINAL PRESENTADO PARA LA OBTENCIÓN DEL TÍTULO DE ESPECIALISTA EN CIENCIA DE DATOS

BUENOS AIRES

SEGUNDO CUATRIMESTRE, 2019

ABSTRACT

Con el proyecto Uber Movement las ciudades del mundo pueden obtener información necesaria que les permita identificar puntos neurálgicos donde adaptar sus infraestructuras con objetivo de llevar a cabo una optimización del flujo de tráfico [8].

En la Ciudad de Buenos Aires, la aplicación BA Taxi brinda similar plataforma a la de Uber. El Gobierno entrega el dataset a usuarios finales, se procede a realizar inicialmente un análisis descriptivo y se generan nuevas variables a partir de las otorgadas. Finalmente se analiza la existencia de combinación de variables tal que se prediga viajes de mala calidad, es decir, viajes en donde no se presente una correlación directa entre tiempo insumido y distancia recorrida.

Luego de la clara detección de ciertos patrones clave para la identificación de viajes malos, se considerará la investigación como piedra angular para futuros estudios incluyendo nuevos conjuntos de datos cómo ser: meteorología, obras viales o protestas.

ÍNDICE

CAPÍTULO I: Introducción

Las grandes ciudades del mundo inicialmente fueron diseñadas y construidas para ser transitadas por automóviles. En el presente, la gran adopción del uso de los mismos causó el colapso del tránsito produciendo una consecuente contaminación del medioambiente. Nuevas alternativas de movilidad inteligentes surgen para contrarrestar lo anteriormente mencionado.

Empresas como Uber/Lyft/Cabify lideran esta transformación en las grandes ciudades fomentando el uso de *carpooling* y tecnologías verdes de locomoción [1]. La instalación de las mencionadas empresas suelen causar disrupción en los países cuestionando, dentro de otros tópicos, el *status quo* en materia de transporte y traslado de pasajeros.

Localmente en Argentina y, más precisamente en la Ciudad de Buenos Aires, como estrategia al avasallante establecimiento de Uber se instaura la aplicación BA Taxi brindando similar plataforma a la de Uber segmentando su operación en los Taxis de la Ciudad de Buenos Aires [3].

En el marco de políticas públicas y con el objetivo de fomentar la transparencia en sus gestiones, los gobiernos del mundo exteriorizan sus datos recolectados por diversas fuentes poniéndolo a disposición de sus ciudadanos.

En el presente estudio se pretende procesar y analizar los viajes efectuadas por Taxis bajo la aplicación BA Taxi.

CAPÍTULO II: Antecedentes

Diversos estudios de análisis de flujo de tráfico relacionados a medios de transporte han sido llevados a cabo tanto por gobiernos como empresas privadas.

En materia de medios de transporte como ser el taxi, con el ya bien conocido crecimiento y adopción de empresas como Uber, Lyft,Cabify y el boom en tecnologías para la explotación de grandes volúmenes de datos, es de especial interés su estudio.

En este caso, Uber se destaca con el proyecto Uber Movement que permite a las ciudades adaptar sus infraestructuras e invertir en soluciones que las hagan más eficientes en materia de transporte nutriéndose de la información brindada por esta empresa. Para ello Uber calcula el tiempo insumido por viajes entre origen y destino, velocidad promedio y finalmente calcula una velocidad llamada libre de flujo permitiendo a las ciudades tomar decisiones en base a los datos mencionados [8].

Estas empresas pretenden llevar el transporte al próximo nivel, concientizando a usuarios sobre el uso racional de automóviles, fomentando el *carpooling* y tecnologías verdes de locomoción [1]. Hoy día lo anteriormente mencionado se encuentra en discusión, incluso estudios realizados en San Francisco (cuna de estas empresas) manifiestan que no sólo no mejoró el tráfico sino que lo perjudicó. El indicador "tiempo de demora de vehículos" caracterizado por la comparación entre tiempo de viajes en situaciones de congestión versus el libre flujo alcanzando un incremento de un 62% según este estudio [2].

CAPÍTULO III: Definición del problema

Problema

Analizar el uso del sistema BATaxi con el objetivo de la detección de patrones y anomalías existentes en el sistema.

CAPÍTULO IV: Objetivos

Objetivo general

Análisis de la correlación entre las variables del dataset BATaxi, identificando y clasificando patrones de uso.

Objetivos específicos

- Obtener dataset a utilizar.
- Seleccionar herramienta de análisis/visualización.
- Enriquecer dataset de varias fuentes.
- Extraer conclusiones de la población objeto de estudio para la toma de decisiones.
- Comparar los resultados del presente estudio con otros estudios realizados.

CAPÍTULO V: Hipótesis

Hipótesis

Existencia de viajes cuya distancia recorrida no se encuentra enteramente correlacionada con el tiempo de viaje, factores exógenos a los viajes puedan incidir en los mismos.

CAPÍTULO VI: Diseño de la Investigación

La presente investigación tiene como objetivo realizar un análisis descriptivo de tendencias en la utilización del sistema BATaxi.

- Determinación de dataset a utilizar.
- Análisis preliminar de dataset.
- Enriquecimiento de dataset generando nuevos atributos en el mismo aplicando *featuring engineering*.
- Interpretar patrones de utilización del sistema BATaxi.
- Comparar los resultados del presente estudio con otros estudios realizados.

Muestras que serán relevadas

Se analizarán patrones de utilización del sistema BATaxi en una muestra de viajes otorgada por el Gobierno de la Ciudad de Buenos Aires.

Método de relevamiento

Datasets provisto por el Gobierno de la Ciudad de Buenos Aires en una hackathon sobre Ciudades Inteligentes [9].

Variables a medir

Viajes del sistema BATaxi:

- ID de Taxista.
- Fecha Inicio y Fin del Viaje.
- Duración del Viaje.
- Latitud/Longitud Origen,
- Latitud/Longitud Destino.
- Cantidad de pasajeros transportados.

Métodos de análisis estadísticos

- Analizar patrones de uso del sistema BATaxi.
- Identificar anomalías en el sistema BATaxi.

CAPÍTULO VII: Alcances y limitaciones

- El Dataset obtenido por parte del Gobierno de la Ciudad de Buenos Aires se encuentra limitado a los meses Mayo/Junio/Julio/Agosto del año 2017.

CAPÍTULO VIII: Resultados

- Análisis preliminar de dataset:

  Se procede al análisis preliminar del dataset recibido por el Gobierno de la Ciudad de Buenos Aires con el objetivo de identificar la consistencia y la distribución de los datos obtenidos.

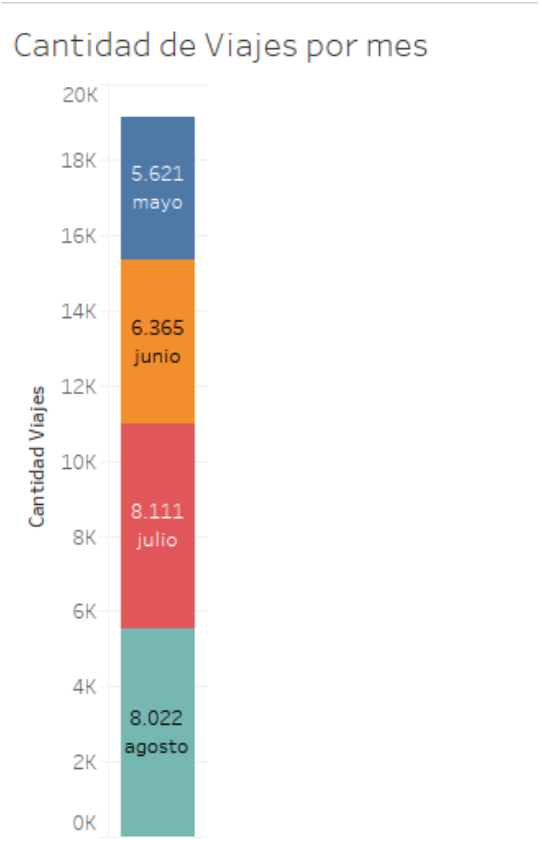
Atributos contenidos en el dataset y su correspondientes descripción:

bataxi-2019-10-27_084052.tsv

Viajes efectuadas por el sistema BA Taxi en 2017 desde Mayo a Agosto.

COLUMN NAME	TYPE	DESCRIPTION
# id_viaje_r	integer	ID único de viaje.
# id_taxista_r	integer	ID del taxista que realizó el viaje.
fecha_inicio	datetime	Fecha y hora inicio del viaje.
fecha_fin	datetime	Fecha y hora fin del viaje.
# duracion	integer	Duración en segundos del viaje.
# origen_viaje_x	decimal	Longitud origen del viaje.
# origen_viaje_y	decimal	Latitud origen del viaje.
# destino_viaje_x	decimal	Longitud destino del viaje.
# destino_viaje_y	decimal	Latitud destino del viaje.
# cantidad_pasajeros	integer	Cantidad de pasajeros transportados.

Distribución de viajes por mes en el año 2017:



El dataset contiene aproximados 19K registros con una distribución balanceada de cantidad de viajes por mes.

- Enriquecimiento de dataset original:

Una vez efectuado la identificación de las variables correspondientes al dataset original, se busca la generación de nuevas variables que permiten entender con mayor facilidad el contexto del problema. Aplicando *feature engineering*[1] se logran las siguientes variables derivadas:

turno	string	Turno del viaje (Madrugada,Mañana,Tarde,Noche). (Campo inferido).
dia_semana	string	Día de la semana en la que se produjo el viaje (campo inferido).
laborable	string	Lunes a Viernes -> Laborable. Sabado y Domingo -> no laborable.
distancia	decimal	Distancia recorrida en metros (campo inferido).
barrio_origen	string	Barrio origen (campo inferido)
barrio_destino	string	Barrio destino (campo inferido)
good_bad	boolean	Clasificación de viaje Bueno/Malo. Se traza recta de regresión, por arriba de ella se considera un viaje malo y por debajo bueno.
dur_cluster	string	Categorización por clustering de la variable duración.
dis_cluster	string	Categorización por clustering de la variable distancia.

Turno/día de la semana/día laborable/no laborable: Ambas variables categóricas derivadas de la hora y fecha del viaje respectivamente.

[1] https://en.wikipedia.org/wiki/Feature_engineering El codigo pyton, se encuentra disponible en [10]

```
# Cálculo turno/día de la semana/laborable/no laborable
from datetime import date

def turno(hora):
  if ((hora >= 0) & (hora <= 5)):
    return('Madrugada')
  elif ((hora > 5) & (hora <= 12)):
    return('Mañana')
  elif ((hora > 12) & (hora <= 20)):
    return('Tarde')
  else:
    return('Noche')

def laborable(diaSemana):
  if ( (diaSemana == 'Saturday') | (diaSemana == 'Sunday') ):
    return('No Laborable')
  else:
    return('Laborable')

df['turno'] = df.apply(lambda x: turno(pd.to_datetime(x['fecha_inicio']).hour), axis=1)
df['dia_semana'] = pd.to_datetime(df['fecha_inicio']).dt.day_name()
df['laborable'] = df.apply(lambda x: laborable(x['dia_semana']), axis=1)
```

Distancia: Distancia en metros recorrida por el viaje. Variable generada a partir de la latitud y longitud origen/latitud y longitud destino.

```
# Cálculo distancia (lineal)
!pip install geopy
from geopy.distance import geodesic
from geopy.point    import Point

for index, row in df.iterrows():
  lon_ori = row['origen_viaje_x']
  lat_ori = row['origen_viaje_y']
  lon_des = row['destino_viaje_x']
  lat_des = row['destino_viaje_y']

  punto_ori = Point(latitude=lat_ori, longitude=lon_ori)
  punto_des = Point(latitude=lat_des, longitude=lon_des)

  df.loc[index, 'distancia'] =  int(geodesic(punto_ori, punto_des).meters)
```

Barrio Origen/Barrio Destino: Se utilizan las coordenadas origen y destino y se las incluye dentro del mapa de la Ciudad de Buenos Aires [4, 7] y Provincia de Buenos Aires. Ambos mapas se encuentran sectorizados en Barrios/Partidos, logrando de esta forma asignar Barrio de Origen y Destino a los viajes en cuestión. Dicho trabajo se realiza mediante la aplicación de consultas espaciales en PostGIS.

Clustering distancia/duración: Se busca realizar una categorización de los viajes de acuerdo a su distancia y duración para lo cual se efectúa un algoritmo de Clustering sobre las variables mencionada anteriormente utilizando la aplicación de QGIS.

Ambas variables se categorizan en Short/Medium/Large de acuerdo a los siguientes valores obtenidos del análisis de k-Means.

Branch: master ▾ bataxi / cluster_distancia.csv Find file Copy path

fpuglie Update and rename cluster_distancia.tsv to cluster_distancia.csv 69b7bc5 38 minutes ago

1 contributor

4 lines (4 sloc) 68 Bytes Raw Blame History

Search this file...

	metros	km	grupo
2	3875	3.9	short
3	7890	7.9	medium
4	37900	37.9	large

Branch: master ▾ bataxi / cluster_duracion.csv Find file Copy path

fpuglie Update cluster_duracion.csv 9abb639 37 minutes ago

1 contributor

4 lines (4 sloc) 73 Bytes Raw Blame History

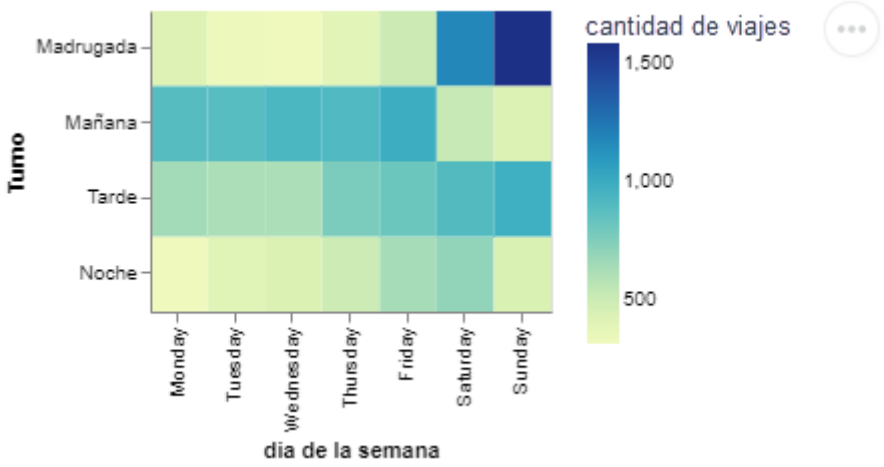
Search this file...

	segundos	minutos	grupo
2	779	13	short
3	1336	22.3	medium
4	3740	62.3	large

- Interpretar patrones de utilización del sistema BATaxi:

  Una vez obtenidas la variables derivadas, se crea el siguiente heatmap con el objetivo de identificar cuáles son los días y turnos dónde se encuentra la mayor concentración de viajes.

```
# Cálculo distancia (lineal)
!pip install geopy
from geopy.distance import geodesic
from geopy.point    import Point

for index, row in df.iterrows():
  lon_ori = row['origen_viaje_x']
  lat_ori = row['origen_viaje_y']
  lon_des = row['destino_viaje_x']
  lat_des = row['destino_viaje_y']

  punto_ori = Point(latitude=lat_ori, longitude=lon_ori)
  punto_des = Point(latitude=lat_des, longitude=lon_des)

  df.loc[index, 'distancia'] =  int(geodesic(punto_ori, punto_des).meters)
```

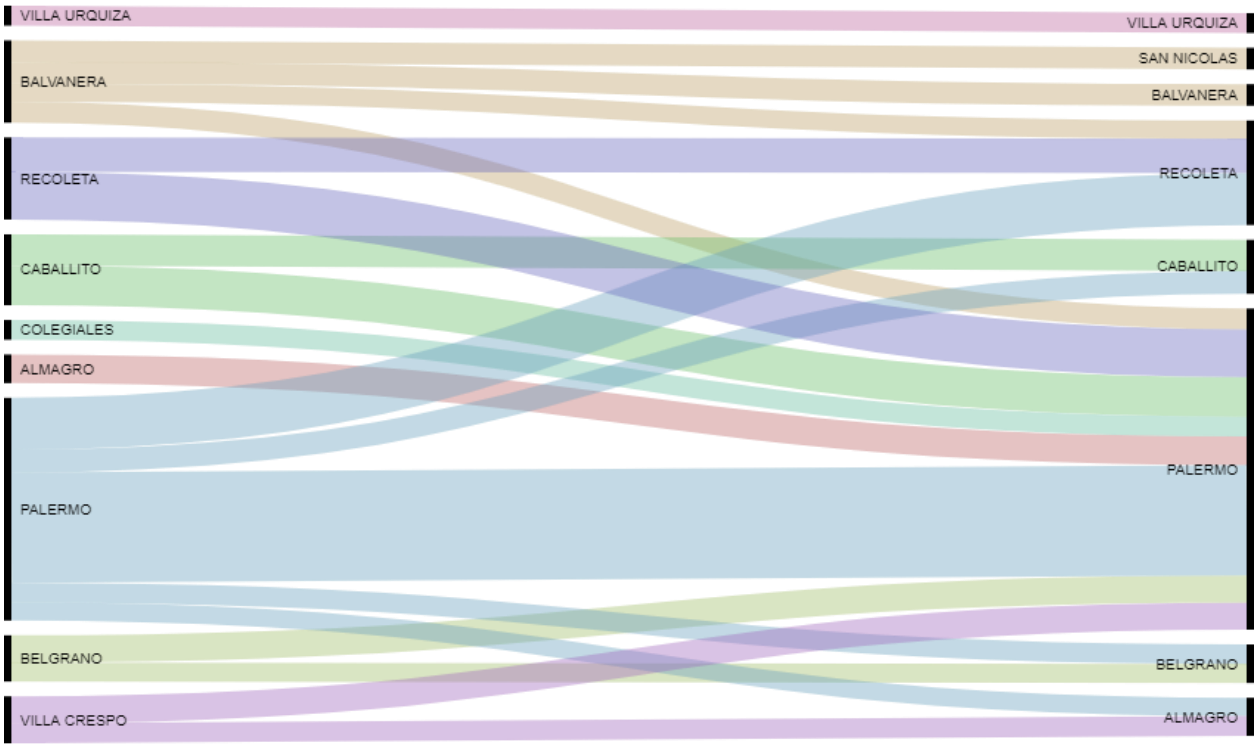


De acuerdo al gráfico anterior, se puede interpretar que la mayor concentración de viajes se encuentra a la madrugada del sábado y del domingo. Esto es consistente con la mayor afluencia de público a la noche de los días viernes y sábados.

En segundo lugar, siguiendo a los sábados y domingos, se observa una alta concentración de lunes a viernes a la mañana. Esto se encuentra alineada a la mayor afluencia de personas dirigiéndose a sus respectivos trabajos.

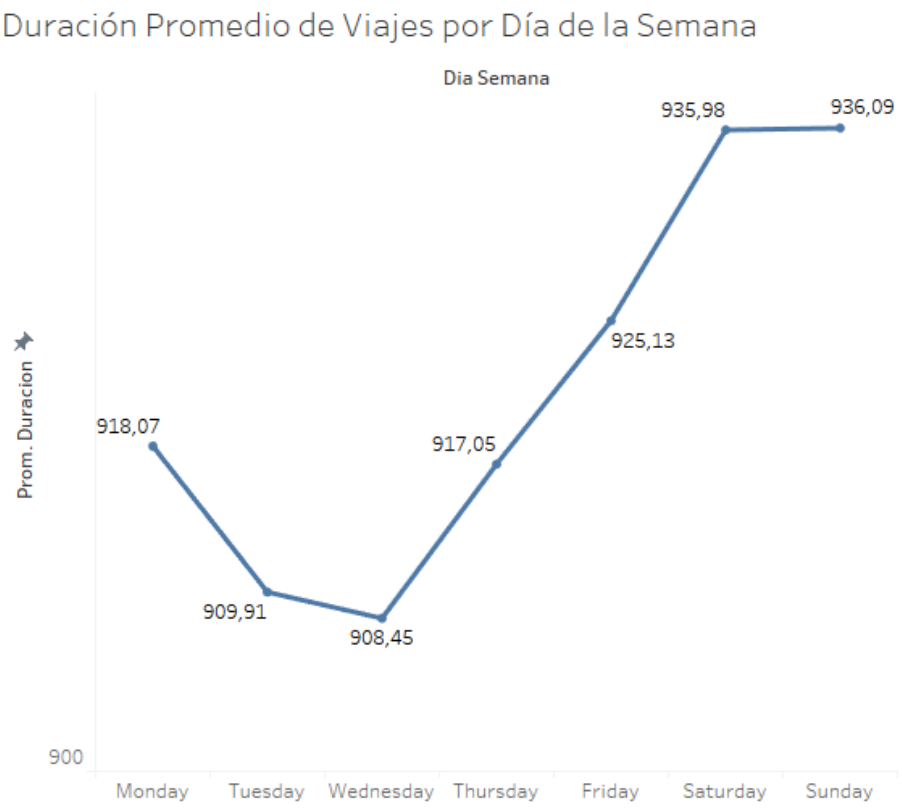
Por el contrario, la menor concentración se ubica los domingos a la noche y días de semana a la madrugada.

Top 20 Barrios Origen y Destino: Mediante un diagrama de Sankey (realizado con RawGraph [5]) podemos visualizar cómo un flujo de información se traslada desde un origen hacia un destino. En el presente caso, utilizamos barrio origen y destino.



Se detecta claramente que Palermo es el barrio con mayor uso de Taxis tanto de Origen como Destino seguido por los barrios Recoleta/Caballito/Balvanera como Origen de Viaje.

Duración promedio de viajes por día de las semana:

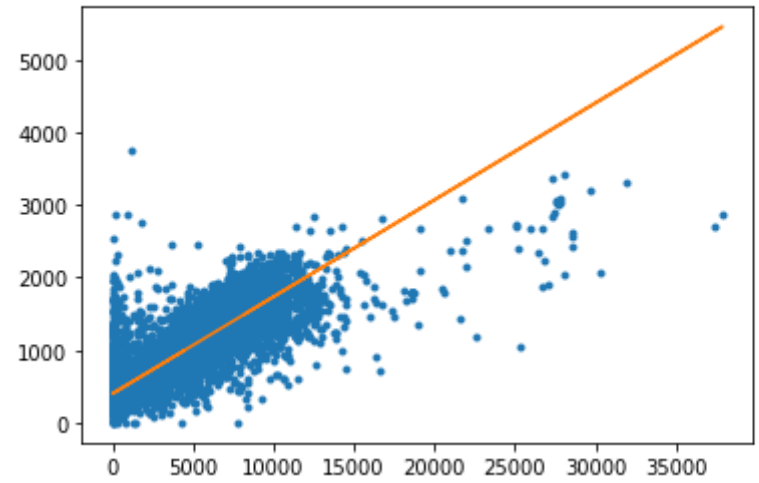


Se observa un incremento en la duración promedio de los viajes a partir del jueves con una mayor incidencia sábados y domingos.

Análisis de correlación entre variables duración y distancia:

```
# ScatterPlot y recta de regresión
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np


x2 = df.distancia
y2 = df.duracion

m, b = np.polyfit(x2, y2, 1)

plt.plot(x2, y2, '.')
plt.plot(x2, m*x2 + b, '-')

print("m=", m," b=", b)

plt.show()
```

```
m= 0.13340144265978157  b= 398.93067540713713
```



Se efectúa un scatterplot con x = distancia (en metros), y = duración (en minutos) registrándose una relación lineal directamente proporcional entre ambas variables.

Lo mencionado anteriormente se encuentra también explicado por un alto R2 y bajo P-Value. Es decir que la duración es explicada en un 69% por la distancia recorrida.

OLS Regression Results

Dep. Variable:	duracion	R-squared:	0.690
Model:	OLS	Adj. R-squared:	0.690
Method:	Least Squares	F-statistic:	4.260e+04
Date:	Sun, 10 Nov 2019	Prob (F-statistic):	0.00
Time:	16:14:46	Log-Likelihood:	-1.3258e+05
No. Observations:	19148	AIC:	2.652e+05
Df Residuals:	19146	BIC:	2.652e+05
Df Model:	1		
Covariance Type:	nonrobust		

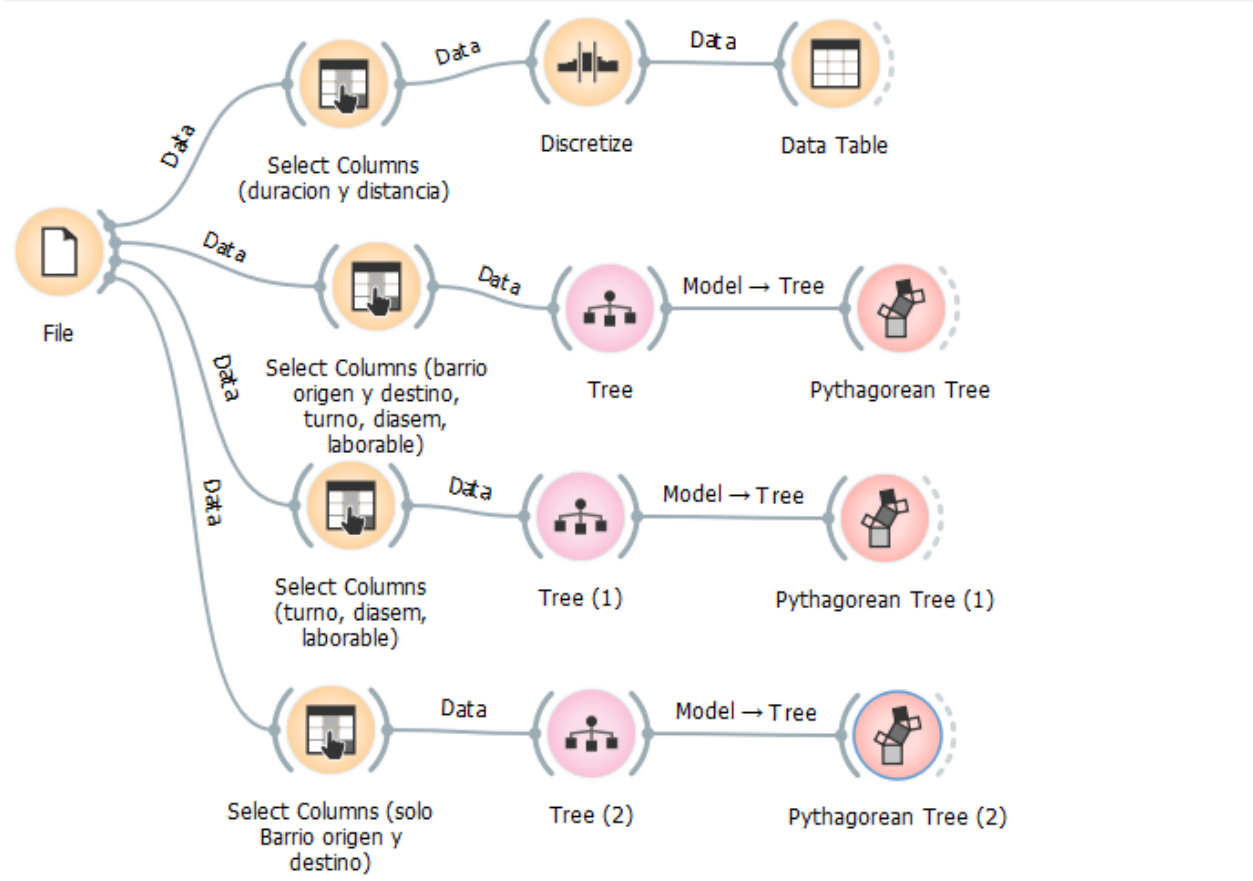
	coef	std err	t	P>|t|	[0.025	0.975]
const	398.9307	3.100	128.678	0.000	392.854	405.007
distancia	0.1334	0.001	206.400	0.000	0.132	0.135

Omnibus:	3622.529	Durbin-Watson:	1.985
Prob(Omnibus):	0.000	Jarque-Bera (JB):	108140.820
Skew:	-0.016	Prob(JB):	0.00
Kurtosis:	14.642	Cond. No.	8.37e+03

Clasificación de viaje bueno/malo: Luego de trazar la recta de regresión se clasifican a los viajes por encima de la recta de regresión como malos y por debajo como buenos.

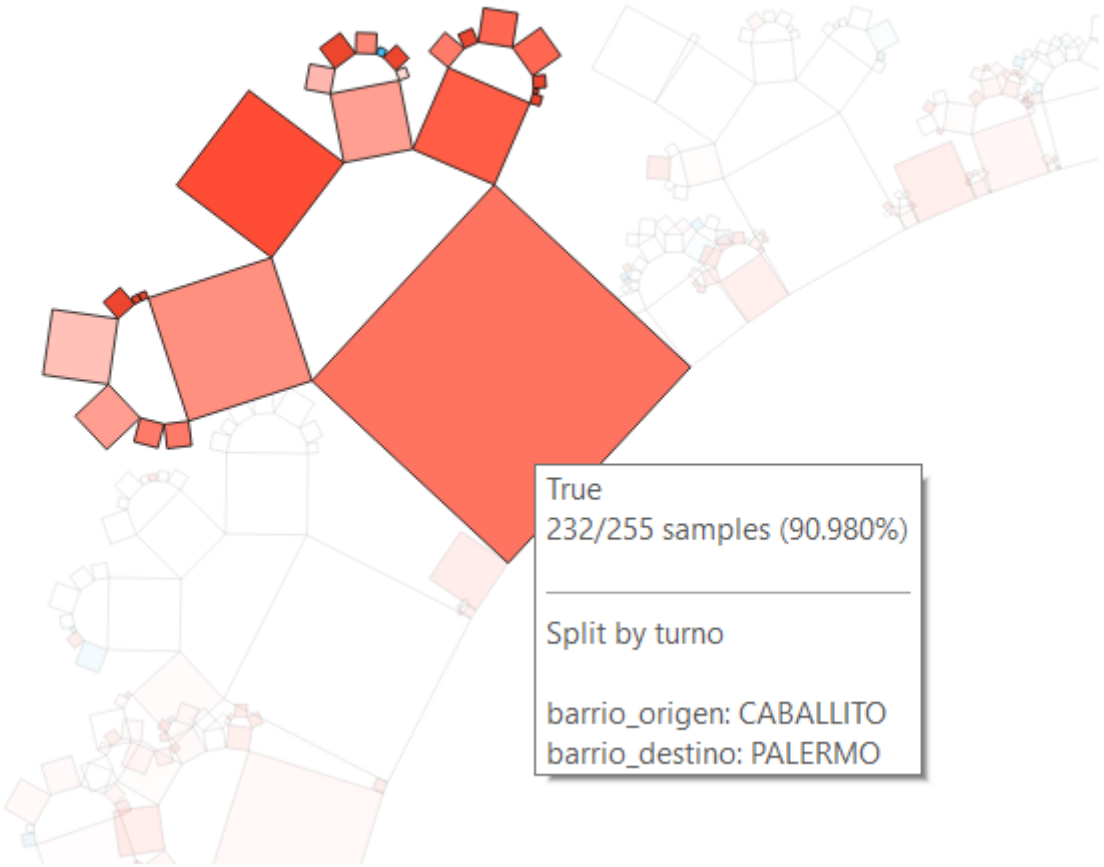
Inferencia de la variable target viaje bueno/malo: Una vez enriquecido el dataset se procede a identificar características de viajes que permitan inferir con las variables del tipo *feature* la variable objetivo o *target* viaje bueno/malo.

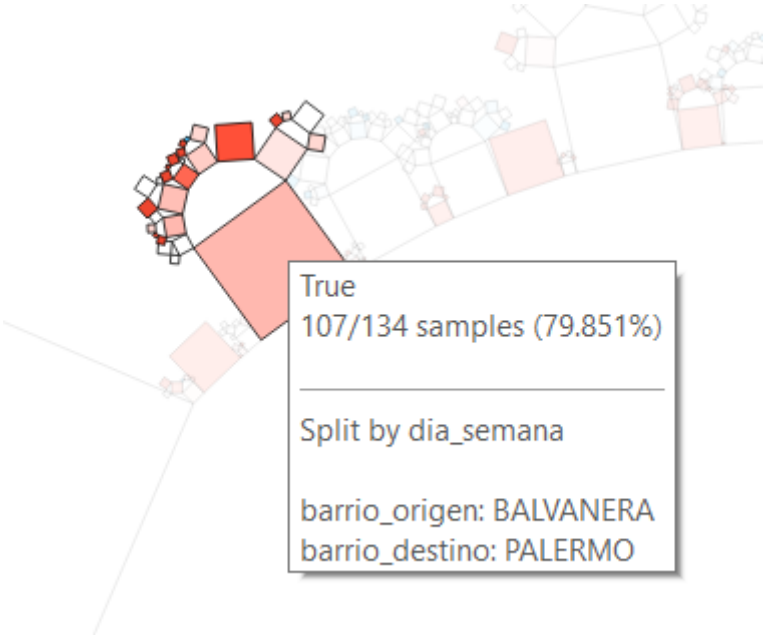
Utilizando la herramienta de data mining Orange [6] se selecciona del dataset distintas combinaciones de campos y aplicando la técnica de árbol Pitagórico se buscan combinaciones de las variables que segmenten con un alto porcentaje (> 70) viajes considerados “malos”.



Se concluye que si un viaje parte desde Caballito a Palermo, en el 91% de los casos el viaje es considerado malo.

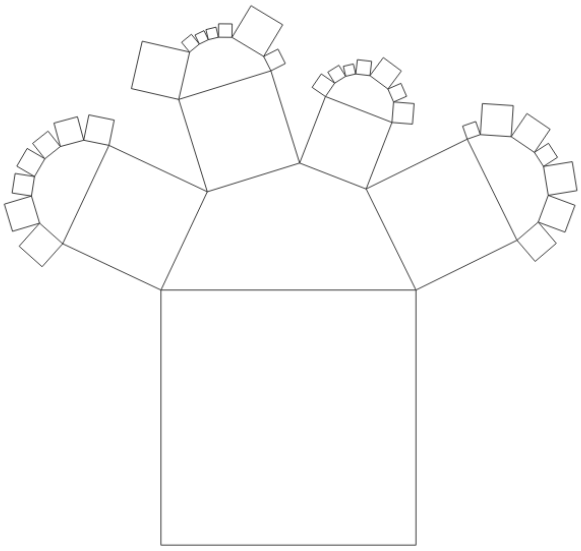
De mismo modo si un viaje parte de Balvanera hacia Palermo, en el 79% de los casos el viaje también es considerado malo.





Por el contrario, si en el dataset se utiliza las variables turno/dia semana/laborable se visualiza que el árbol no corta en ningún momento al dataset en partes mayores al 70%. Esto explica que las variables en cuestión no predicen la variable target. Nótese en el siguiente árbol en ningúno de los casos sus hojas son pintadas.

Más aún, si se seleccionan todas las variables del dataset, efectivamente el árbol se genera sólo cortando por barrio origen y destino. Esto confirma la hipótesis que el barrio origen y destino incide notablemente en la clasificación de un viaje.



CAPÍTULO IX: Discusión de los resultados

En el estudio en cuestión podemos llegar a la conclusión que viajes desde Balvanera/Caballito hacia Palermo tendrán altas probabilidades de ser considerados “malos” correlacionando tiempo/duración y las variables semana/día es laborable/turno del día no tienen incidencia sobre esta decisión.

Es preciso formular las siguientes hipótesis sobre estas situaciones:

- ¿Existen buenas condiciones de rutas entre los destinos?
    - ¿Excesos de lomas de burro o baches?
- ¿Existen buenas comunicaciones entre los destinos?
- ¿Existirán obras viales constantes dentro de esas fechas?
- ¿Podrá incidir alguna condición climática?
- ¿Existirán protestas en la vía pública?

CAPÍTULO X: Conclusiones y trabajos futuros

Sin lugar a dudas el crecimiento exponencial de datos de diversas fuentes y su disponibilización de forma abierta otorga la posibilidad concreta a usuarios finales de efectuar análisis diversos que, o bien no eran posibles procesar por la tecnología al momento existente, o bien quedaban en manos de gobiernos o privados.

Para lograr caracterizar con mayor precisión el flujo de viajes entre puntos del mapa y detectar posibilidades de mejoras de comunicación entre barrios será necesario recabar datos de fuentes extra que permitan refutar las hipótesis en **capítulo IX - Discusión de Resultados** cómo ser meteorología/obras viales/protestas/etc.

Este será el desafío planteado para futuros trabajos en materia de transporte en los Taxis de la Ciudad de Buenos Aires.

REFERENCIAS

[1] https://blueandgreentomorrow.com/transport/top-5-eco-friendly-transportation-methods-you-can-feel-great-about/

[2] https://www.theverge.com/2019/5/8/18535627/uber-lyft-sf-traffic-congestion-increase-study

[3] http://www.asociaciontaxistasdecapital.com.ar/ba-taxi/

[4] https://data.buenosaires.gob.ar/

[5] https://rawgraphs.io/

[6] https://orange.biolab.si/

[7] https://usig.buenosaires.gob.ar/ (Unidad de Sistemas de Información Geográfica CABA)

[8] https://movement.uber.com/?lang=en-US

[9] Repositorio GitHub, con los datos de bataxi, https://github.com/fpuglie/bataxi

[10] Código fuente python, generación de nuevas variables y gráficos. http://bit.ly/bataxi-fpuglie