# UNIVERSIDAD TECNOLÓGICA NACIONAL
# FACULTAD REGIONAL BUENOS AIRES

**Carrera:** **INGENIERÍA INDUSTRIAL**

**Materia:** **CIENCIA DE DATOS**

# TRABAJO PRÁCTICO FINAL

## *Predicción del flujo de ciclistas en la Ciudad de Buenos Aires*

## Equipo de trabajo:


Lema, Federico Ariel

Legajo: 155.590-0

fedelema96@gmail.com

Angeles, German

Legajo: 156.188-1

german_1997_20@hotmail.com


Grupo 16

## 1. OBJETIVO

Nuestro trabajo se basará en analizar el volumen de ciclistas en la Ciudad de Buenos Aires para poder observar los barrios o zonas donde hay mayor flujo y así determinar cuál será el flujo de ciclistas en los puntos en donde no tenemos información.

Así mismo, analizaremos los bicicleteros activos de la Ciudad, para ver si corresponden con este flujo y evaluar donde correspondería colocar nuevos sin pisarnos con los que ya están activos.

## 2. DATASETS UTILIZADOS

Para realizar lo propuesto en el objetivo, vamos a trabajar con 2 datasets principales y 1 secundario.

Uno será de volumen de ciclistas anuales, el cual inicialmente tenía 9004 samples (m) y 6 features (n), pero por motivos de practicidad lo modificamos, agrupando grupos similares, obteniendo un nuevo dataset de 794 samples y 6 features.

Las samples o filas son las instancias del dataset, las que obtenemos de una muestra de datos, mientras que las features o columnas son las variables que definen a cada sample. La cantidad de features equivale a la cantidad de dimensiones que describirán a cada sample.

Las features del dataset de ciclistas son:

- Cruce: son los nombres de las dos calles que intersectan en esa esquina.
- Longitud: formando las coordenadas del cruce junto con la latitud.
- Latitud: formando las coordenadas del cruce junto con la longitud.
- Cantidad de registros: es el promedio de los registros obtenidos del flujo de ciclistas por cada punto, entre los años 2013 y 2018.
- Mañana: es el promedio de los registros obtenidos del flujo de ciclistas por cada punto, tomados a la mañana, entre los años 2013 y 2018.
- Tarde: es el promedio de los registros obtenidos del flujo de ciclistas por cada punto, tomados a la tarde, entre los años 2013 y 2018.

El otro dataset será de bicicleteros públicos, el cual posee 199 samples y 10 features.

Las features del dataset de bicicleteros públicos son:

- Longitud: coordenadas del bicicletero
- Latitud: coordenadas del bicicletero
- Nombre: como se designó a este bicicletero
- Domicilio: el cruce donde se ubica el mismo
- Imagen: archivo de la imagen del bicicletero (la eliminaremos)
- Automatización: indica si el bicicletero es automático o manual
- Observación: dato de información sobre el bicicletero, como la fecha en que pasó de ser automático a manual
- Número de estación: hay un número asignado a cada bicicletero
- Horario: la disponibilidad horaria en la que opera

- Dirección normalizada: es el domicilio escrito de forma normalizada

Para nuestro análisis, vamos a eliminar las columnas de imagen, automatización, observación y número de estación, ya que no las consideramos relevante.

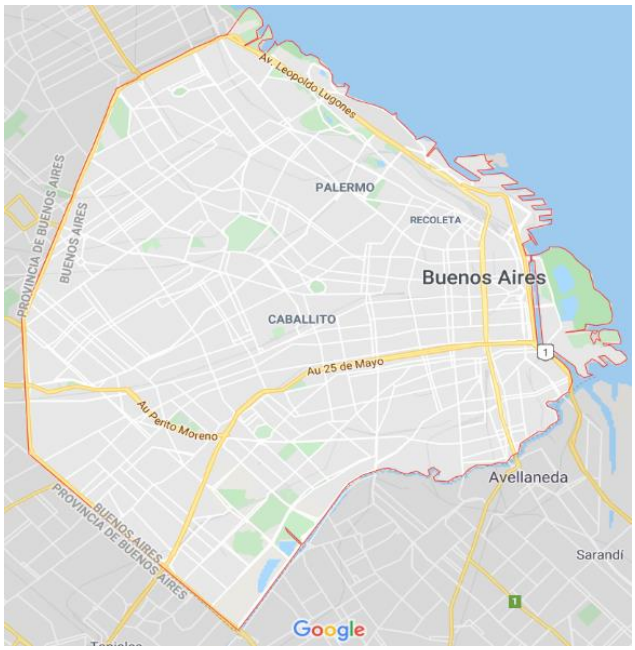
Por último, utilizaremos un dataset secundario de lugares para dejar la bicicleta, con 938 samples y 16 features, para ver también si se corroboran estos con el flujo de los ciclistas. No ahondaremos mucho en las distintas features de este dataset ya que no lo consideramos relevante para este reporte.

## 3. EDA (Exploratory Data Analysis)

Comenzamos haciendo un poco de limpieza en los dataset, eliminando filas que contenían algunos valores nulos (conocidos como NaNs) y eliminando columnas que no considerábamos relevantes para nuestro análisis (mencionadas anteriormente).

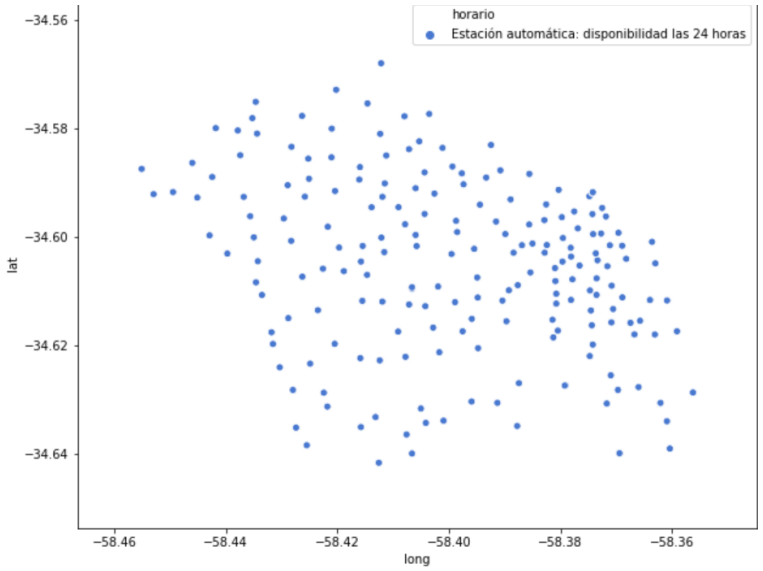
En el de ciclistas pudimos observar que había 48 filas que no tenían la información de latitud y longitud, por lo que procedimos a borrarlas ya que no podíamos utilizarlas y no representaban un porcentaje significativo del total.

```
cruce                0
long                48
lat                 48
Cantidad_registros   0
mañana               0
tarde                0
dtype: int64
```



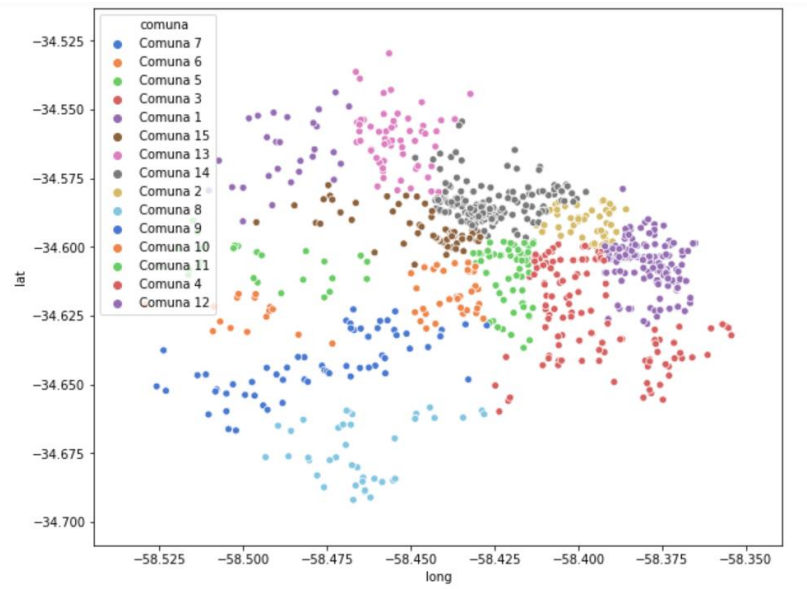

Luego, hicimos algunos gráficos para poder visualizar las distribuciones de los datos de los distintos datasets. Podemos entender la ubicación de los mismos al comparar los gráficos con un mapa de la Ciudad de Buenos Aires.

A. Scatterplot con las coordenadas de los distintos bicicleteros públicos, agrupados por el horario de cada uno:




Con este gráfico notamos que los bicicleteros estaban bastante distribuídos en general y que se agrupaban más por la zona del centro. También se pudo determinar que todos tenían una disponibilidad de 24 horas, por lo que ya no quedan bicicleteros manuales que trabajan menos de 24 horas por día.

B. Scatterplot con las coordenadas de los distintos lugares para dejar la bicicleta, agrupados por comuna:

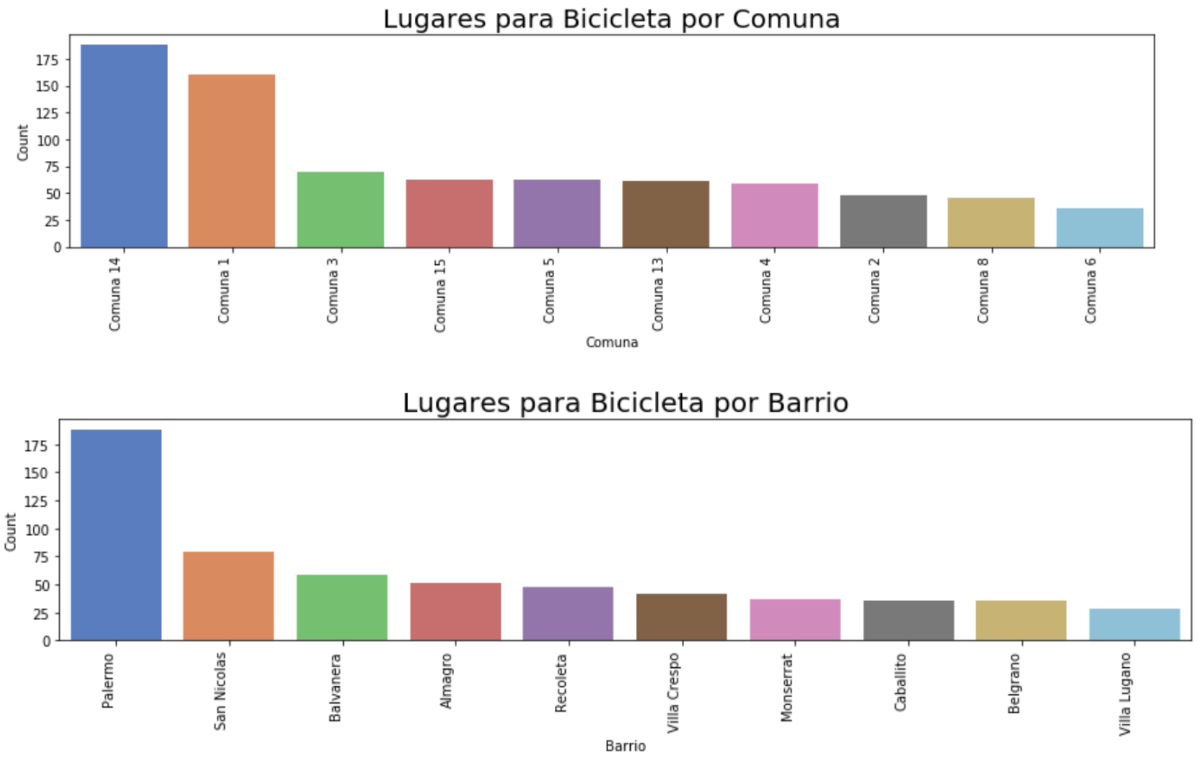



Acá se puede visualizar que los lugares disponibles para dejar la bicicleta están bastante más agrupados por la zona de Palermo, Recoleta.

Elegimos distribuirlos por comuna para poder ver mejor los grupos, ya que al dividir por barrio quedaba muy confuso el mapa.

Complementario a esto, hicimos un ranking para ver donde había más cantidad de estos lugares.

C. Countplot de la cantidad de lugares para dejar la bicicleta, agrupado por comuna y por barrio:




Realizamos el countplot para que nos muestren las 10 comunas/barrios con mayor cantidad de estos lugares. Pudimos precisar que la Comuna con mayor cantidad es la 14, seguida por la 1; mientras que el barrio con mayor cantidad es Palermo, seguido por San Nicolás.

Después de analizar estos datasets, hicimos mapas de calor para ver el flujo de ciclistas promedio en la ciudad entre 2013 y 2018.

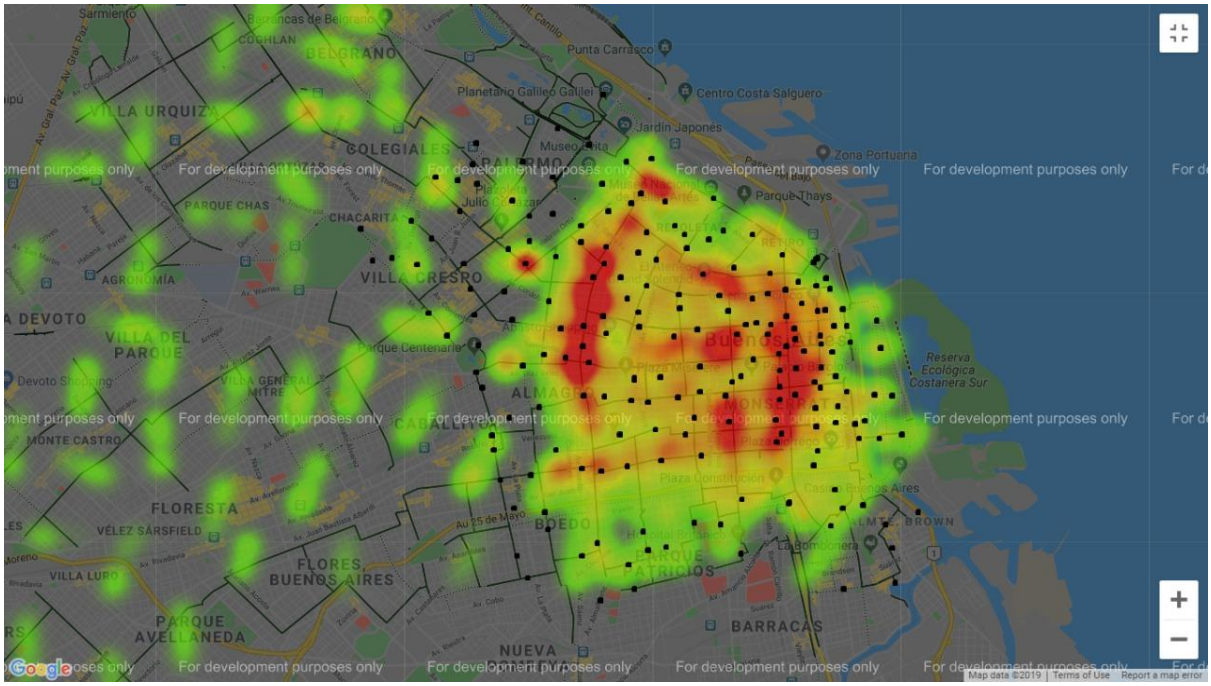
Por un lado, hicimos un mapa general de flujo promedio diario y, por otro lado, hicimos 2 mapas separados: flujo por la mañana y flujo por la tarde.

A los mapas obtenidos les hicimos algunas modificaciones para poder visualizarlos mejor:

- Colocamos mapas de Google Maps de la ciudad de fondo, para ver la distribución exacta en cada zona.
- Agregamos los puntos en donde se encuentran los bicicleteros públicos, para poder ver si coincide la mayor densidad de estos con las zonas de mayor flujo o si será necesario colocar nuevos bicicleteros en las cercanías de estas zonas.
- Agregamos las bicisendas de la ciudad, para poder ver si el flujo tiene relación con estas y para verificar que los bicicleteros estén próximos a las bicisendas.

Los mapas finales obtenidos son los siguientes:

*FLUJO PROMEDIO DIARIO*




En la zona céntrica pudimos ver que había mucha cantidad de bicicleteros en la zona más cercana a Puerto Madero que por ahí no eran tan necesarios como en otras zonas con más flujo.

Hay zonas rojas aisladas donde podría incluirse un biciletero.

Pudimos ver también que la cantidad de bicisendas están en mayor cantidad en la zona céntrica, lo que explica el flujo mayor, ya que los ciclistas deben sentirse más seguros navegando por caminos designados especialmente para ellos en vez de ir por entre medio de los autos.

*FLUJO A LA MAÑANA*

*FLUJO A LA TARDE*


Se puede evidenciar que hay un mayor flujo de ciclistas a la tarde con respecto a la mañana.

Podemos deducir de esto que a la mañana la gente utiliza la bicicleta para transportarse: como para ir a trabajar o a la facultad; mientras que a la tarde esta gente hace este camino de retorno a sus hogares.

Sumado a esto, deducimos que a la tarde se agrega una mayor cantidad de gente que utiliza la bicicleta de modo recreativo, como para salir a pasear, ya que es un horario que el general de la gente suele tener más disponible que la mañana.

Por esto es que en el mapa vemos mayores densidades a la tarde (las densidades más intensas son de color rojo, luego aparecen colores anaranjados hasta las densidades medias, que son amarillas y, por último, las más bajas son verdes).

## 4. MACHINE LEARNING

Una vez analizados nuestros datos, quisimos generar un modelo que pueda tomar los datos existentes de las coordenadas y predecir cuál sería el flujo en esa ubicación.

Para esto, aprovechamos que ya teníamos coordenadas con su respectivo flujo calculado y decidimos entrenar un modelo usando estos datos para que luego pueda realizar esta predicción.

Por lo tanto, comenzamos dividiendo nuestro dataset de flujo de ciclistas en dos:

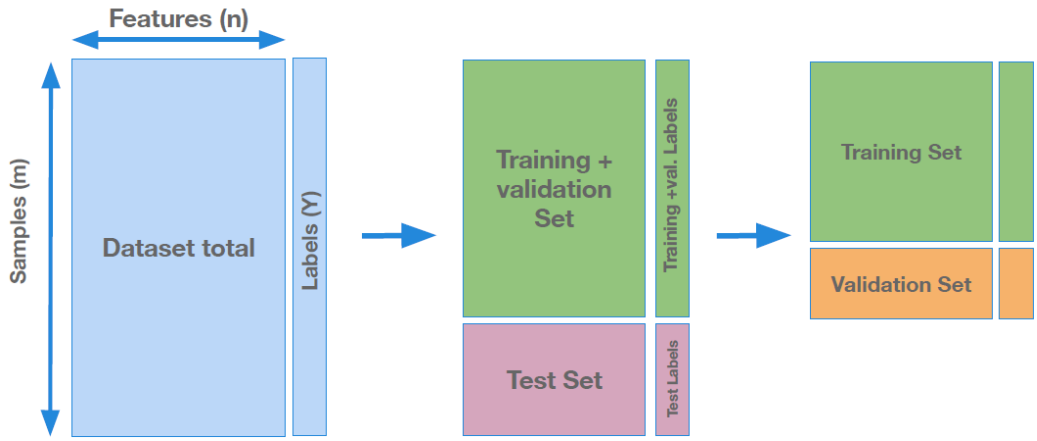
- TRAIN: Estos datos serán usados para el entrenamiento del modelo. Dentro del train vamos a tener los datos que nos servirán de input (X), que serán las coordenadas: latitud y longitud.

Y también tendremos el output o la etiqueta (Y), la cual dependerá de los datos de input X. Usaremos para el entrenamiento el 75% del dataset.

- TEST: Estos datos serán usados como prueba para ver qué tan bien se adaptaron las predicciones del modelo a las etiquetas reales. Por lo que separaremos nuestros datos en X e Y, al igual que en train, pero solo le daremos al modelo los X. Usaremos como prueba el 25% restante de nuestro dataset.

Por lo tanto, el modelo obtendrá la información X TRAIN, Y TRAIN y X TEST. Y nos devolverá una predicción (Y PRED), las cuales compararemos con Y TEST para ver qué tan preciso es y el error obtenido.

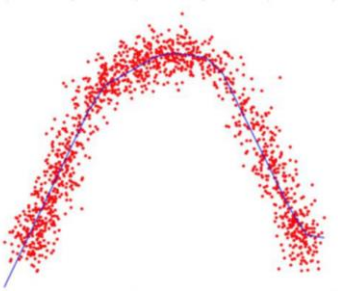
A su vez, subdividiremos el TRAIN en un TRAINING SET y un VALIDATION SET, lo que nos servirá para ver si el modelo funciona bien o si solo lo hace porque tuvimos suerte en la forma en la que lo partimos para entrenarlo.

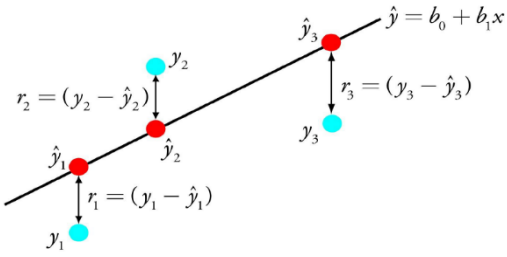



**El clasificador aprenderá la regla de decisión utilizando el train set (samples + labels). Luego clasificará las muestras de test (sin mirar las labels de test) y se medirá la exactitud de clasificación en testeo.**

Luego de esto, quisimos ver qué funciones de regresión funcionaba mejor con nuestro modelo y nos arrojaban el menor error en las predicciones.

Las funciones regresoras son aquellas que, a partir de ciertas variables X de una muestra, estiman un valor de la variable dependiente Y de esa misma muestra. Esta se ajustará a los datos de manera de reducir el error entre la predicción y el valor real.





Para medir el error entre la predicción y el valor real lo podemos hacer de tres formas:

- MSE (Error cuadrático medio)
- RMSE (Raíz cuadrada del error cuadrático medio)
- MAE (Media del error)

Vamos a utilizar para comparar nuestros modelos de regresión el RMSE, ya que es el más significativo.

$$\boldsymbol{MAE} = \frac{|\sum (\widehat{y}_t - y_t)|}{n}$$

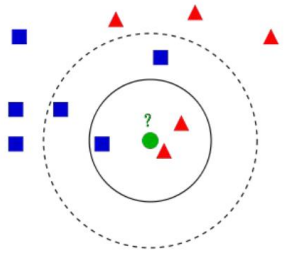
$$\boldsymbol{MSE} = \frac{\sum (\widehat{y}_t - y_t)^2}{n}$$

$$\boldsymbol{RMSE} = \sqrt{\frac{\sum (\widehat{y}_t - y_t)^2}{n}}$$

Decidimos implementar 3 modelos distintos de regresión, y en base al RMSE obtenido de cada uno, nos quedamos con el mejor. Estos modelos son:

KNN Regression

Este modelo toma los K vecinos más cercanos por distancia euclídea y en base a estos predice el resultado (interpolando los Y de los vecinos).



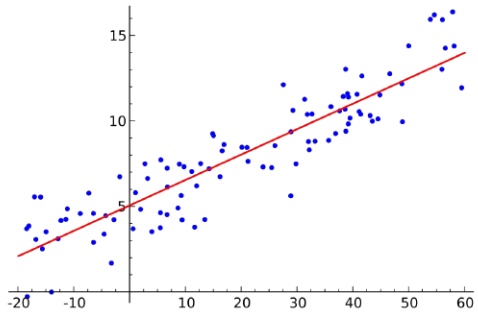

Usamos este modelo y nos predijo los resultados con un error

**RMSE= 19,99**

$$d(x_a, x_b) = \sqrt{(x_{a1} - x_{b1})^2 + (x_{a2} - x_{b2})^2 + ... + (x_{ap} - x_{bp})^2}$$

Regresión lineal

La regresión lineal es una función lineal que se construye calculando los parámetros Beta asociados a cada variable.




Para obtener los valores de los parámetros del modelo utilizamos Mínimos cuadrados y obtenemos una única solución.

$$\min_{\beta} \|X_w - y\|^2 \longrightarrow \hat{\beta} = (X^T X)^{-1} X^T y$$

$$\widehat{y} = f\,(x, \beta)$$

$$\hat{y}(x, w) = w_0 + w_1 x_1 + w_2 x_2 + ... + w_p x_p$$

En este caso el error obtenido **RMSE= 21,63**

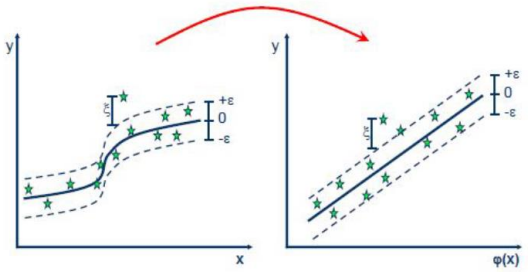
Support Vector Regression (SVR)

SVR busca determinar un margen (o radio) como función de costo y trata de que todas las muestras estén dentro del margen.

Las muestras que caigan fuera del margen tendrán un “Xsi mayor a 0” y las que estén dentro del tubo tendrán un “Xsi=0”.

El hiper-parámetro es una función que penaliza muestras fuera del margen.

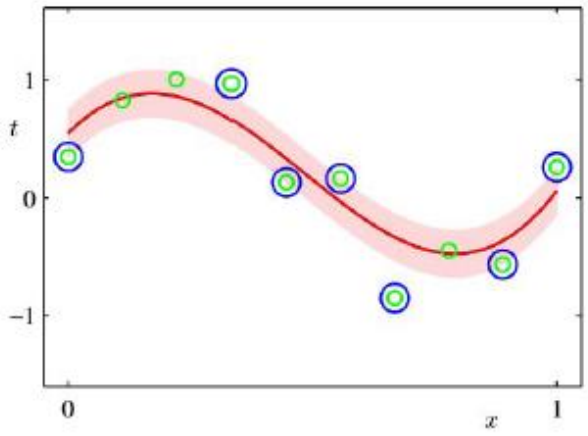
Si el error entre predicción (Y PRED) y el valor real de Y es menor que Epsilon, entonces se determina que el error es 0.

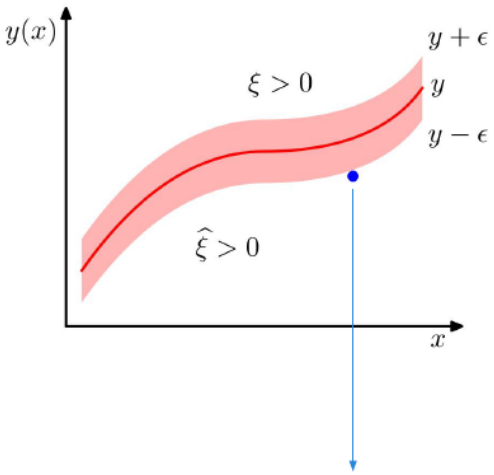



$$C\sum_{n=1}^{N}\xi_n + 1/2\,\|w\|^2$$

$$min\frac{1}{2}\|w\|^2 \qquad \begin{aligned} y - wx_i - b &< \varepsilon \\ -y + wx_i + b &< \varepsilon \end{aligned}$$

Sólo algunas muestras definirá el margen de predicción y serán llamadas “support vectors”.




Al igual que en SVM, podemos “permitir/flexibilizar muestras fuera del “tubo” de predicción del modelo.

El error obtenido en este caso es **RMSE= 24,99**

En base a los errores obtenidos con cada modelo, elegimos quedarnos con el KNN ya que en el se registro el menor error de los tres (RMSE= 19,99).

## 5. CONCLUSIÓN

Luego de realizar este estudio, pudimos determinar que el algoritmo no es tan bueno como queríamos para determinar los flujos, ya que no tenía información suficiente en cuanto a cantidad de muestras y en cuanto a cantidad de dimensiones para el análisis.

Vimos que para poder mejorar este análisis, además de agregar mediciones en más ubicaciones para poder disminuir nuestro error, se podría medir otras variables que ayuden al modelo a predecir mejor como las siguientes:

- Si hay o no una bicisenda que pase por esa esquina
- Distancia al bicicletero más cercano
- Si hay cerca de esta ubicación algún edificio de importancia, como universidades o empresas de gran envergadura

Creemos que estas variables pueden ayudar más al modelo a poder definir el flujo ya que los ciclistas suelen circular más por las zonas dónde hay bicisendas o bicicleteros cercanos. También consideramos que una gran parte del flujo se debe a que las personas se trasladan en bicicleta para poder concurrir a sus trabajos o al lugar dónde estudian.