Análisis del arbolado público lineal en la Ciudad de Buenos Aires

Ciencia de datos
Universidad Tecnológica Nacional
Facultad Regional Buenos Aires

Abstract
El presente trabajo está formado por dos partes: primero un análisis exploratorio de los datos, y luego la aplicación de un modelo de machine learning (logistic regression), para lograr predecir información acerca de los datos analizados.

1 INTRODUCCIÓN

El objetivo de este trabajo es analizar los datos del arbolado público lineal existente en la Ciudad Autónoma de Buenos Aires, y mediante las características de cada árbol poder predecir qué tipo de hoja tiene.

2 DATASET

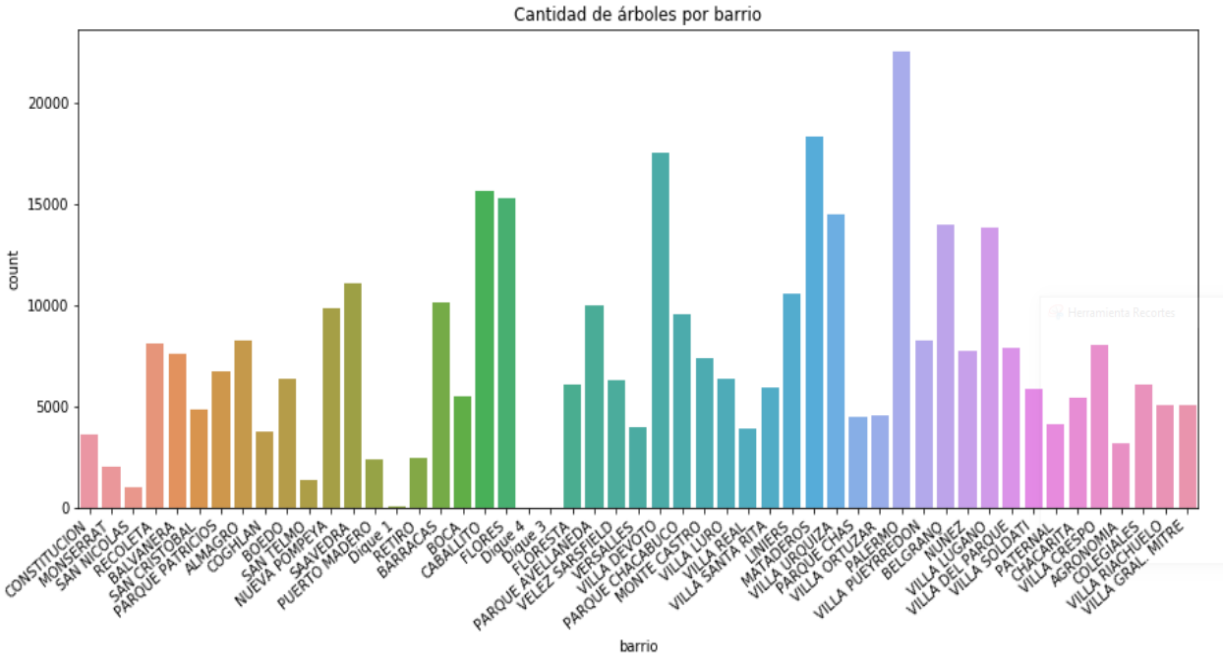
El dataset elegido está formado por 372.699 samples y 20 features. Éstas features son: coordenadas de localización de cada árbol (latitud y longitud), tipo de evento relevado, id del árbol, altura, diámetro, inclinación, id de la especie, nombre de la familia del árbol, nombre del género del árbol, nombre científico del árbol, nombre común del árbol, tipo de follaje, origen del árbol, manzana en la que está ubicado, barrio, comuna, calle, número de chapa frontal, número de chapa alternativa y longitud de la calle donde se encuentra el árbol.

https://data.buenosaires.gob.ar/dataset/arbolado-publico-lineal

3 ANÁLISIS EXPLORATORIO DE DATOS

Se comenzó realizando el EDA (Análisis exploratorio de datos). Se buscaron los NaNs y se eliminaron las filas que los contenían. Luego se eliminaron las features que no aportarían nada al momento de aplicar el modelo.

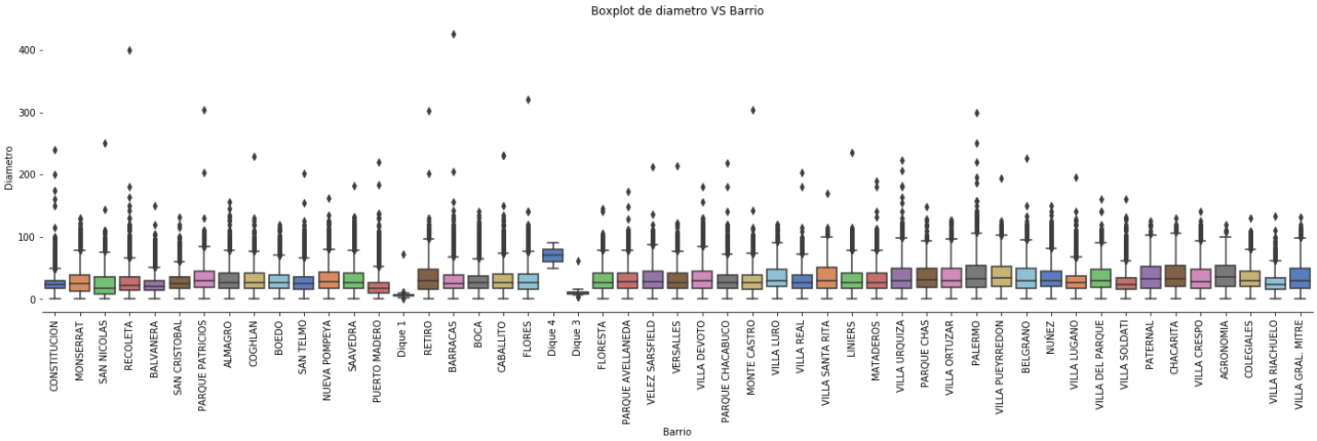
Se realizó un gráfico para visualizar la cantidad de árboles por barrio.

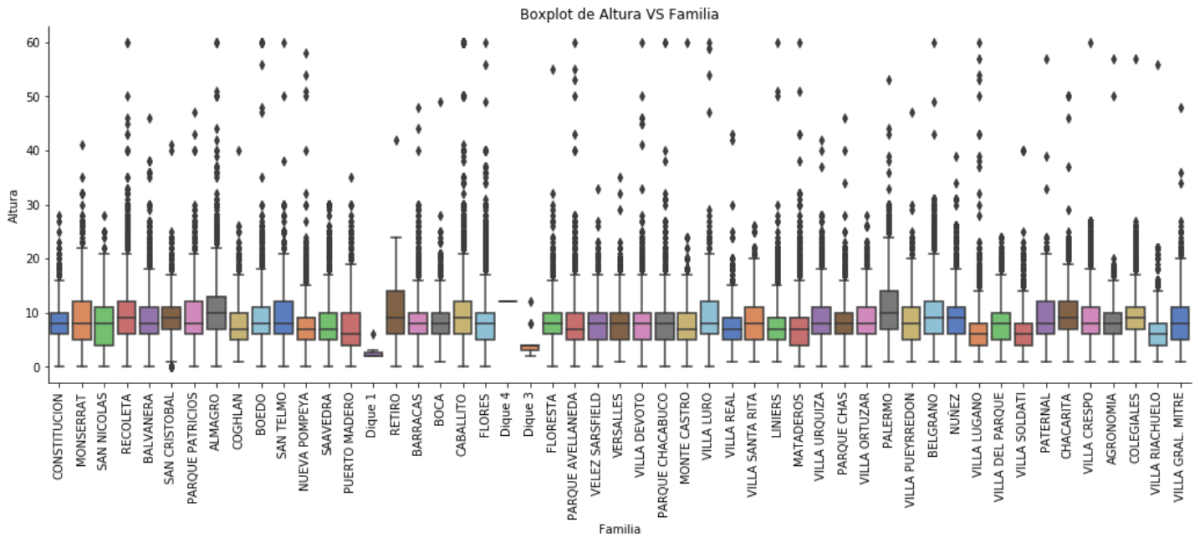


Observamos que los 5 barrios con mayor cantidad de árboles son:

1. PALERMO 22483
2. MATADEROS 18289
3. VILLA DEVOTO 17537
4. CABALLITO 15660
5. FLORES 15278

Se realizaron dos boxplot: uno para los diámetros por barrio, y otro para las alturas por barrio.

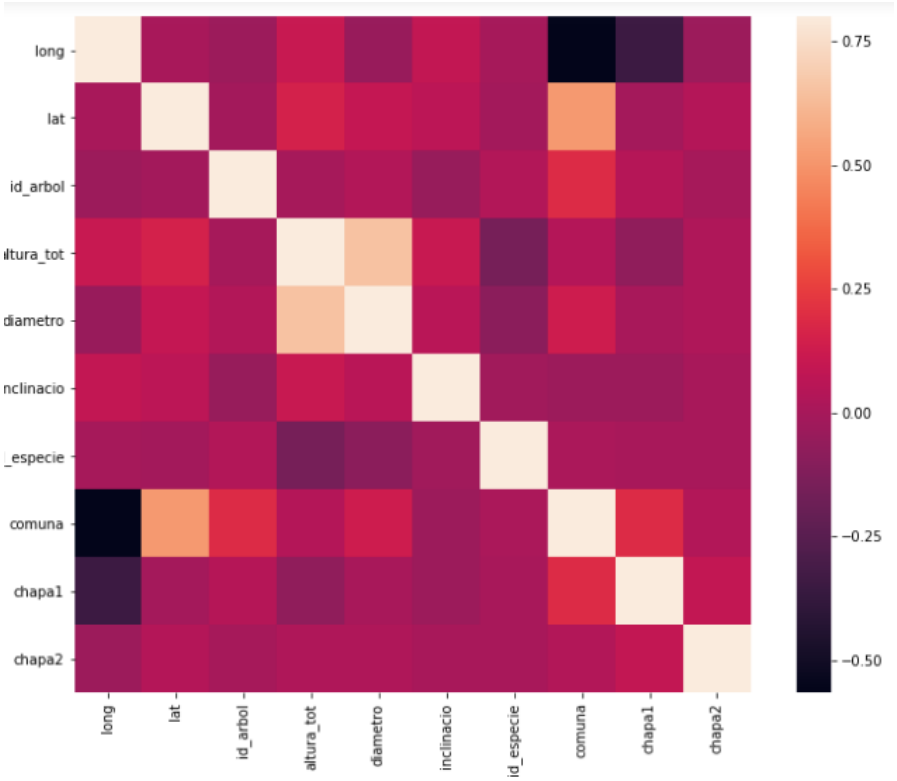




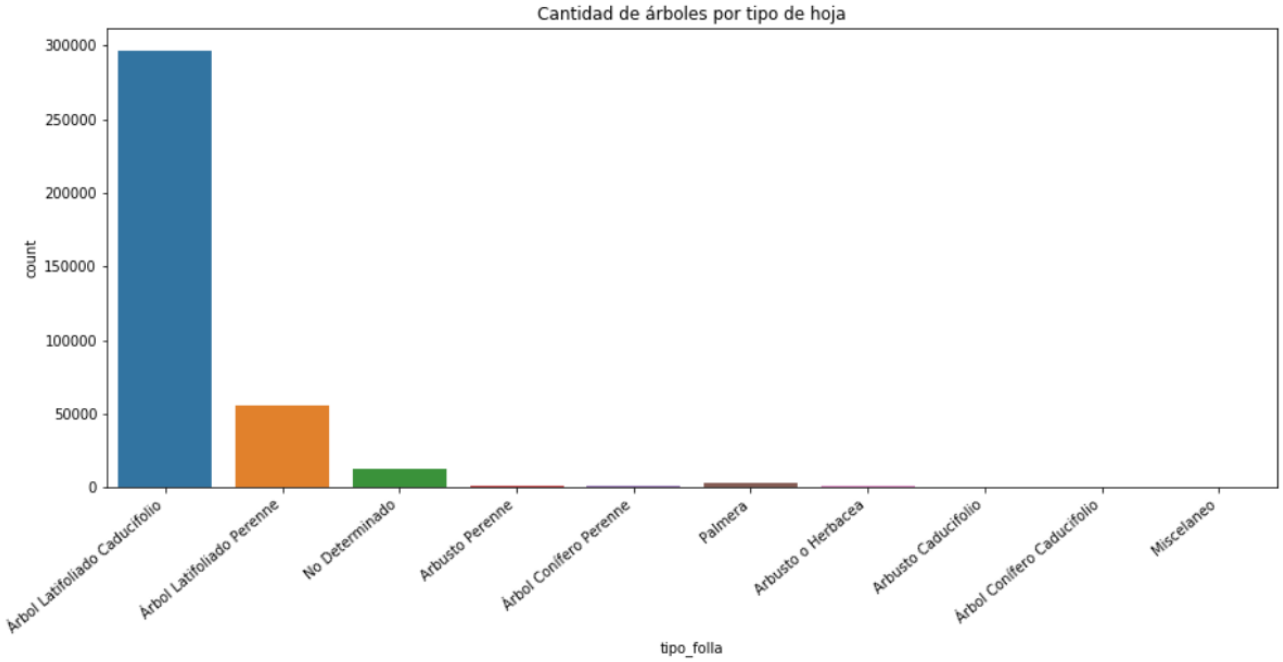
Se observó que los tipos de follaje son: Árbol Latifoliado Caducifolio, Árbol Latifoliado Perenne, No Determinado, Arbusto Perenne, Árbol Conífero Perenne, Palmera, Arbusto o Herbácea, Arbusto Caducifolio, Árbol Conífero Caducifolio y Miscelaneo.

El origen de los árboles puede ser 'Exótico', 'Nativo/Autóctono' o 'No Determinado’.

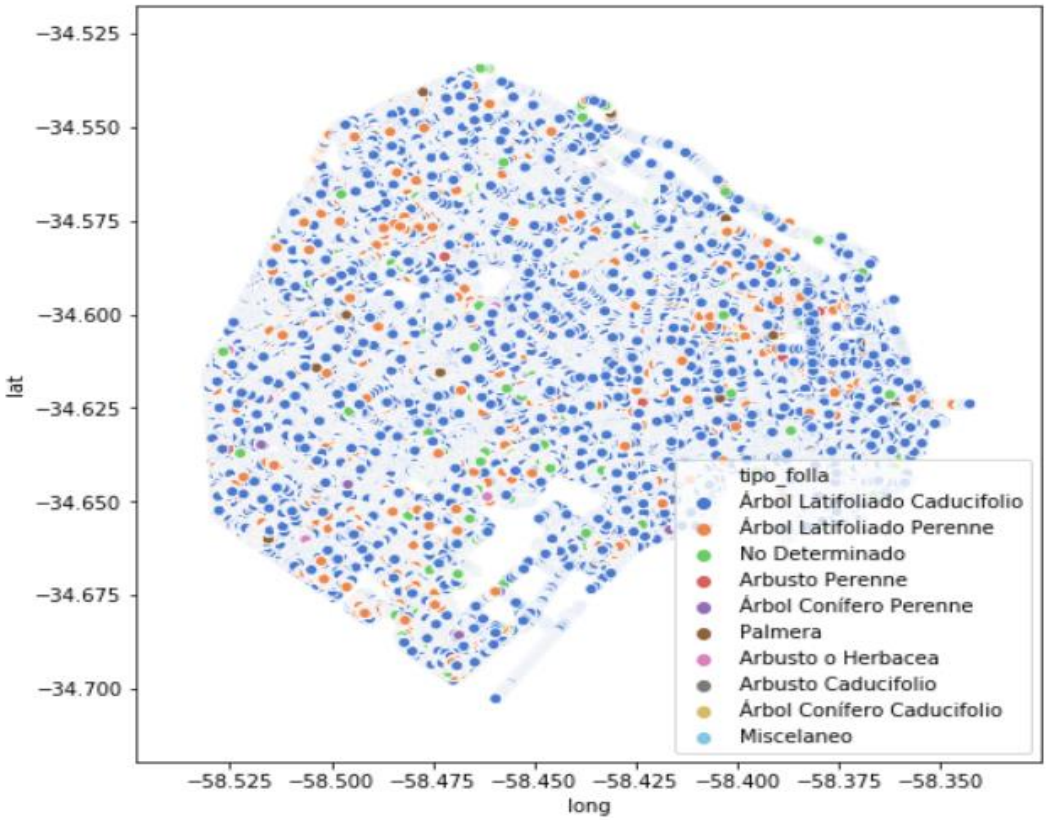
Mediante la realización de un *Heatmap* observamos que hay una alta relación lineal entre la altura y el diámetro de los árboles.



Se realizó un countplot para conocer la cantidad de árboles por cada tipo de follaje existente.



Mediante un scatter plot realizamos un gráfico con los datos de la latitud y longitud de cada árbol para ubicarlos en un mapa de la Ciudad de Buenos Aires.



Se realizó una logistic regression, buscando predecir el tipo de follaje usando: latitud, longitud, altura del árbol, diámetro, inclinación, origen y comuna. Se determinó un tamaño para el test del 70%.

Luego de aplicar el modelo, se obtuvo una accuracy del 91,6%.

Matriz de Confusión Normalizada

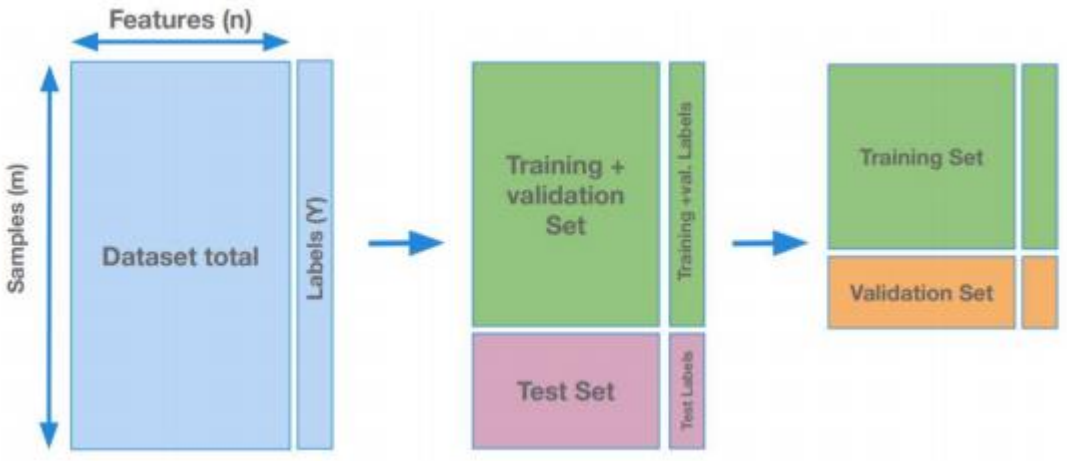
4 MATERIALES Y MÉTODOS

Logistic Regression es un método de clasificación lineal de aprendizaje no supervisado. Es una regresión lineal precedida de una función de activación sigmoide, lo que genera que el output sea binario y no continuo como una regresión normal.

A cada muestra clasificada, le asigna una probabilidad de pertenecer a cada clase existente en el problema. Si la probabilidad es mayor a cierto threshold entonces pertenece a una clase y viceversa.

El regresor logístico debe aprender un parámetro interno por cada dimensión del vector de entrada. Para eso calculará el gradiente del error de clasificación y tratará de minimizarlo.

Para poder realizar la regresión antes hay que realizar algunos pasos previos. Primero separamos la variable dependiente e independiente para poder después separar en train y test los datos. Con esto el clasificador aprenderá la regla usando el train set y luego clasificará las muestras del test y se medirá la exactitud.



Una vez finalizado este paso es necesario escalar los datos. El método asume que cada feature de manera individual responde a una distribución de probabilidad normal y busca estandarizar los valores afectandolos por la media y el desvío standard. Cada feature después de pre-procesarla quedará con una media = 0 y un desvío standard = 1.

Finalmente se aplica el modelo y se calcula el accuracy para ver cuán bien funciona el modelo.

5 CONCLUSIONES

Con lo expuesto podemos concluir que el Logistic Regression aplicado predice correctamente. A partir de los distintos datos de los árboles como la altura, el diámetro, la inclinación, el origen y la ubicación en la ciudad, fue posible predecir el tipo de follaje con una exactitud del 91,6%. Esto podría ser utilizado para;

- Diseño de políticas públicas de arbolado teniendo en cuenta cantidad de árboles por barrio, el tipo de árboles distribuido en la ciudad, para con esto poder planificar a futuro.
- Conociendo el tipo de hoja, se puede conocer en que época del año éstas caerán y planificar la recolección de las mismas.
- Analizar por barrio la presencia de árboles que provoquen algún tipo de alergia a la población, y tomar medidas al respecto.
- En caso de no saber qué tipo de hoja tiene un árbol, con los datos usados como X seríamos capaces de clasificarlo.

6 REFERENCIAS

1) Jerome H. Friedman, Robert Tibshirani y Trevor Hastie “The Elements of Statistical Learning Data Mining, Inference, and Prediction“T.” Página 119
2) Raschka, Sebastian. “Python Machine Learning”
3) https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python