Estudio de la ubicación de los monumentos en la Ciudad Autónoma de Buenos Aires a partir de Machine Learning

Alejo Alvarez – Juan Pablo Alfonso
Universidad Tecnológica Nacional, Facultad Regional Buenos Aires, Argentina

Resumen

Este artículo va a estar conformado por dos partes: en primer lugar, un análisis exploratorio de los distintos datos informados por el portal oficial del gobierno sobre los monumentos con el fin de detectar su composición y distribución dentro de la Ciudad. Posteriormente el objetivo es aplicar modelos de machine learning de aprendizaje supervisado como lo son Support Vector Machine (SVM), Logistic Regression, KNN classification y Random Forest a fines de predecir la calle donde se ubican los monumentos.

Palabras clave: análisis, machine learning, datos, monumentos, distribución.

1 INTRODUCCIÓN

El presente informe consiste en el análisis de la distribución geográfica de los monumentos agrupados por comunas y los materiales por los cuales están conformados.

El objetivo será entender cuales son las comunas con mayor cantidad de monumentos y cuales son los materiales más utilizados en los mismos.

Partiremos de un dataset del portal de datos abiertos del Gobierno de la Ciudad de Buenos Aires el cual se irá puliendo a fines de lograr los objetivos previamente mencionados, extrayendo información que no aporta valor a nuestro análisis.

Por último, con modelos de aprendizaje supervisado intentaremos predecir las calles en las que los monumentos se encuentran.

2 DESCRIPCIÓN DEL DATASET

El Dataset utilizado para realizar el análisis es de carácter público a través de la página oficial de “Buenos Aires Data” la cual brinda datos generados, guardados y publicados por el Gobierno de la Ciudad de Buenos Aires.

El dataset escogido originalmente está conformado por 2233 samples y 17 features.

Éstas features son:

ID: Número único que sirve para identificar y enumerar cada monumento.

OBJETO_OBRA: Indica que tipo de objeto es el monumento (estatua, mástil, placa, etc)

CANTIDAD_OBJETOS: Indica la cantidad de objetos que componen la obra.

MATERIAL: Indica el material con el cual está hecho el monumento.

DENOMINACION_SIMBOLIZA: es la razón a la que hace alusión la obra.

AUTORES: Identifica al autor del monumento.

UBICACIÓN: Indica la ubicación de la obra.

OBSERVACIONES: Remarca alguna observación sobre la ubicación.

DIRECCION_NORMALIZADA: Es la dirección donde se encuentra la obra.

CALLE: Nombre de la calle donde se ubica el monumento.

ALTURA: Número de la dirección donde se encuentra la obra.

BARRIO: Nombre del barrio donde se ubica el monumento.

COMUNA: Indica el número de comuna donde se encuentra la obra.

CODIGA_POSTAL: Es el código postal del barrio donde se encuentra el monumento.

CODIGO_POSTAL_ARGENTINO: Es el código postal expresado de forma completa.

LATITUD: Coordenadas geográficas del lugar donde se encuentra la obra.

LONGITUD: Coordenadas geográficas del lugar donde se encuentra la obra.

3 ANÁLISIS EXPLORATORIO DE DATOS (EDA)

El objetivo de nuestro EDA es mostrar cómo se agrupan los monumentos según las 15 comunas que hay en la Ciudad de Buenos Aires, la distribución de los materiales utilizados, mostrar en promedio la cantidad de objetos de los monumentos por comunas, localizar geográficamente cada monumento y por último entender para cada distribución por comunas que cantidad de material se corresponde con cada una.

Primeramente analizamos con qué tipo de dataset vamos a trabajar, viendo sus dimensiones y definiendo qué variables dejamos fuera del mismo. Procedemos entonces a una limpieza de las columnas "ID", "AUTORES", "CÓDIGO POSTAL ARGENTINO", "DENOMINACION_SIMBOLIZA", "OBJETO_OBRA", "UBICACIÓN", "OBSERVACIONES" y "DIRECCION_NORMALIZADA".

Con el dataset ya con un primer filtrado, pasamos a la instancia de revisión de NaN´s en nuestra variable objetivo "COMUNA" y eliminamos las filas con NaN´s.

Empezamos entonces con nuestro primer análisis en el que analizamos la distribución de monumentos por comunas.

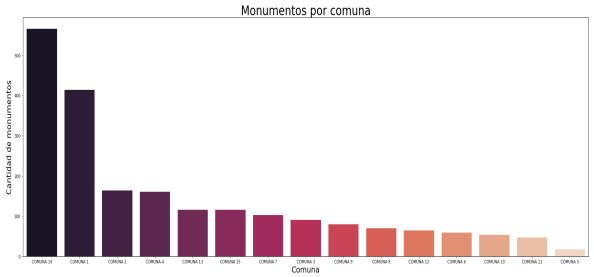

Figura 1

Posteriormente seguimos con el desarrollo de nuestro próximo análisis en el que queremos entender cómo es la distribución de la cantidad de monumentos según el tipo de material. En una primera aproximación a este análisis notamos el número de valores diferentes que aporta la categoría "MATERIAL", siendo 105 diferentes. Los cuales buscamos los más representativos ocupando >1% del total de muestras, nos quedamos así con 12 registros y catalogamos el resto de los materiales bajo la etiqueta "Otros".

Procedemos a realizar el análisis de distribución de los materiales.

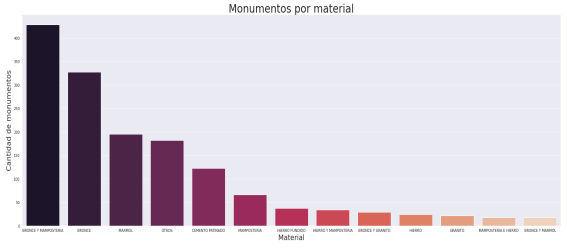

Figura 2

Luego intentamos visualizar la cantidad de objetos promedio por monumento para cada comuna, este caso debe ser analizado particularmente, el boxplot, para la mayoría de los casos queda como una única línea a la altura de valor 1, esto se debe a que la gran mayoría de los monumentos son una única estructura, como lo son estatuas o monolitos, salvo casos muy particulares.

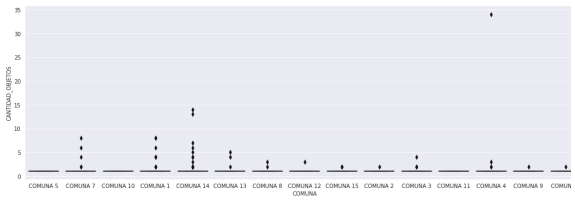

Figura 3

En otra instancia de análisis, buscamos mostrar donde se ubican geográficamente los monumentos mediante las variables “LATITUD” y “LONGITUD”, lo que nos permite visualizar una dispersión con forma similar a la Ciudad de Buenos Aires.

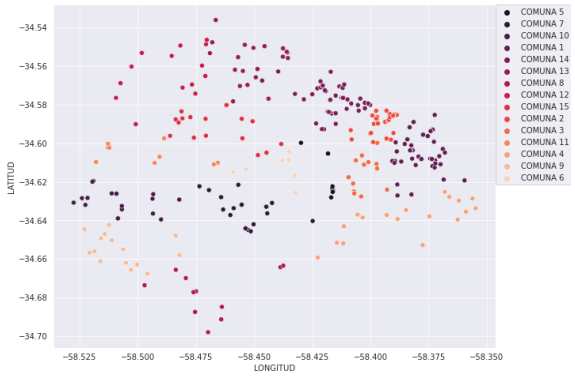

Figura 4

Por último, combinamos mediante una tabla pivot los 2 primeros análisis que realizamos, logrando así un entendimiento de la cantidad de monumentos por comunas y que composición de materiales corresponden a cada una.

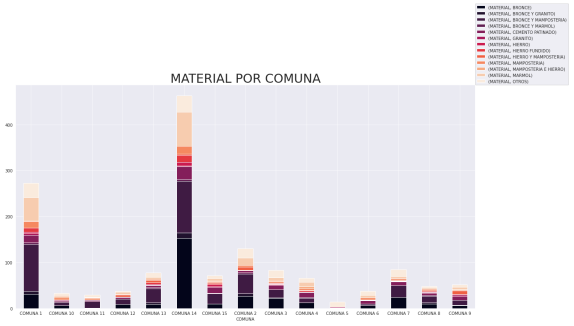

Figura 5

4 MODELO DE APRENDIZAJE

Una vez finalizado el EDA procedemos a aplicar 4 métodos de aprendizaje supervisado para poder predecir la calle en la que el monumento se encuentra según altura, código postal y longitud, esto fue así debido a que la matriz de correlación nos mostraba cuáles son las variables que más relación tienen entre sí.

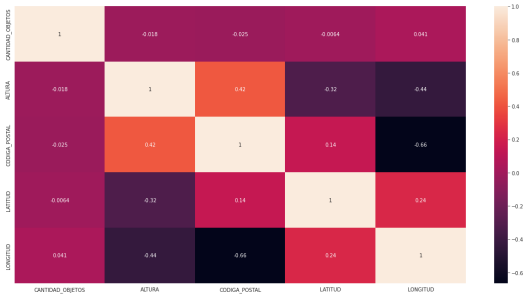

Figura 6

El tamaño de la muestra de entrenamiento definido es de un 25% y establecemos 15 vueltas de entrenamiento

Primero utilizamos el método de Logistic Regression y obtuvimos un Score del modelo de 0.52.

Luego, utilizamos 3 modelos de aprendizaje que nos arrojaron los mismos valores de score, estos métodos son K-Nearest Neighbor, Support Vector Machine y

Random Forest. El score obtenido por estos tres métodos es de 0.93.

También hay que destacar que todos los modelos mencionados previamente fueron acompañados también con Grid Search Cross Validation para poder darle al modelo los hiperparámetros que mejor se ajustan a cada uno de los mismos.

Por último evaluamos, los 3 modelos que nos dieron el score de mayor puntaje en Train los utilizamos y probamos con el test, lo que para los 3 modelos nos dan un accuracy de 0.9095744680851063

	Modelo	Accuracy	Train
0	KNN	0.909574	0.93
1	SVM	0.909574	0.93
2	Random Forest	0.909574	0.93

Figura 7

5 DISCUSIÓN Y CONCLUSIONES

Luego de los distintos análisis llevados a cabo con los modelos de aprendizaje podemos llegar a dos tipos de conclusiones:

Por un lado, el EDA, nos indica que la mayor cantidad de monumentos se encuentra en la comuna 14 y le sigue la comuna 1, además, es apreciablemente mayor la cantidad que poseen ambas en comparación con el resto de las comunas. El material más empleado para la construcción de los monumentos en caba es Bronce y Mampostería , seguido del Bronce y luego en tercer lugar se ubica la categoría que catalogamos como “OTROS”, lo que indica que una gran porción de los monumentos de la Ciudad es de la más diversa composición, desde otro punto de vista esto se correlaciona con la realidad, ya que gran cantidad de los monumentos que uno puede observar en las calles son placas conmemorativas o cuentan con la placa que explica la estatua, monolito, etc. También la cantidad de objetos por monumento que predomina es 1, esto se debe a que la mayoría de los monumentos son estructuras únicas o placas.

En cuanto a la aplicación de modelos de aprendizaje, podemos asegurar que para predecir en qué calle va a estar ubicado el monumento tanto el método de K-Nearest Neighbor como el Support Vector Machine y el Random Forest son igual de precisos. Ante esta particularidad el modelo más óptimo para llevar a cabo es el que produce menor costo de computación. En este caso el modelo con mayor velocidad de reacción es el de Support Vector Machine.

A modo de cierre, podemos decir que el dataset utilizado no es el más óptimo para los modelos de predicción que establecimos o que conceptualmente no seleccionamos las variables indicadas dentro del mismo para este análisis.

6 REFERENCIAS

1. Pattern Recognition and Machine Learning - Christopher Bishop http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf

2 .http://www.Stackoverflow.com

3. https://github.com/clusterai

4 .https://scikit-learn.org/stable/