Análisis de establecimiento educativos

Ciudad Autónoma de Buenos Aires

Ailin Capdevila[1] y Iñaki Perez Pace[1]

[1]Ingeniería Industrial, Regional de Buenos Aires, Universidad Tecnologica Nacional, Argentina

Abstract
Este articulo tiene por objetivo entender como se distribuyen los establecimientos educativos en CABA

Keywords
Educación, Escuelas, Establecimiento educativo.

1 INTRODUCTION Y OBJETIVOS

En el estudio que se ha llevado a cabo se busca conocer la manera en que se distribuyen los establecimientos educativos a lo largo de la Ciudad de Buenos Aires según sus categorías de nivel de enseñanza o de tipo de gestión.

Adicionalmente interesará conocer si es posible etiquetar a los establecimientos conociendo sus coordenadas geográficas.

2 DESCRIPCION DE DATASET

Para llevar a cabo el presente análisis se ha tomado información proveniente de la base de datos de la ciudad de buenos aires respecto a los establecimientos educativos que en ella se encuentran.

Al largo del data set se enumeran 2825 establecimientos de todo tipo, categorizados según el tipo de enseñanza que otorgan (inicial, primario, secundario, especial, etc.). También permite diferenciar los establecimientos que son de gestión pública, de los privados. El set de datos posee las coordenadas x e y de cada una de las instituciones para poder ubicarlos geográficamente, junto con la información del barrio, comuna y distrito educativo.

A continuación, se puede visualizar un mapeo de los establecimientos a lo largo de la ciudad

3 PRE-PROCESAMIENTO DE DATOS

Previo a llevar a cabo el análisis de los datos, se debe limpiar y ordenar el set de modo que el estudio no esté sesgado o limitado por algún error en la carga de los datos originales.

En primer lugar, se pudo visibilizar que, al querer ubicar geográficamente a los establecimientos, existía un reducido número de casos considerados outliers. Como puede visualizarse en el gráfico siguiente, no se obtenía una muestra real de la ubicación de cada institución:

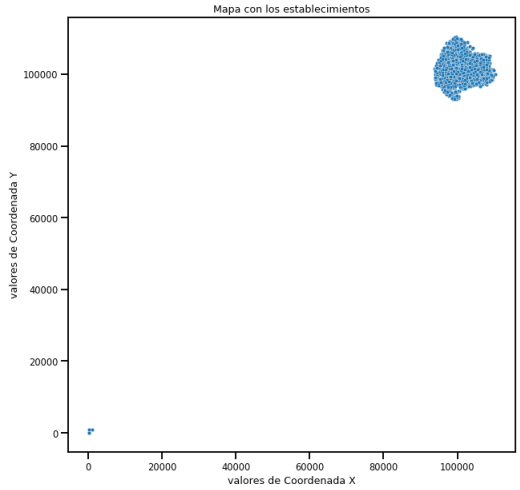


Gráfico 1 - Visualización con outliers

Como se mencionó anteriormente, al ser una cantidad reducida de outliers, se los pudo eliminar sin temor a que genere alguna diferencia en el análisis; resultando el mapa correcto de la siguiente manera:

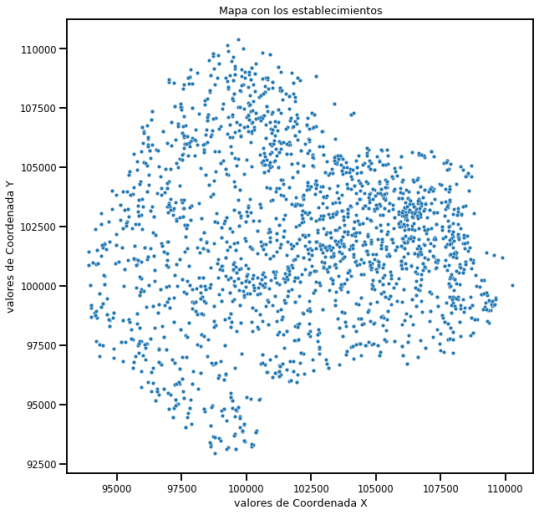


Gráfico 2 - Visualización sin outliers

En segundo lugar, se pudo determinar que existían algunas filas duplicadas; aspecto indeseable debido a que solo interesa poseer una vez cada establecimiento.

Último pero no menos importante, debió estudiarse la existencia o no de valores “NaNs”, es decir valores vacíos no computables los cuales no es posible analizar. Sin embargo, previo a deshacerse de los mismos, debe verificarse que no sean representativos para el set de datos y que no se vea afectado el análisis al eliminarse.

Tabla 1 - Representatividad de NaNs

	Total	Percent
depfun	33	0.011490
tipest_abr	33	0.011490
web_megcba	18	0.006267
email	10	0.003482
nombre_abr	1	0.000348

En la tabla anterior puede visualizarse como la cantidad y su relevancia para el set de datos es muy baja por lo que pueden despreciarse.

4 ANALISIS EXPLORATORIO DE DATOS

Como primera medida interesaba saber que cantidad de establecimientos son de gestión pública y, por ende, cuantos del tipo privada; dándose una clara predominancia por parte de la categoría estatal.

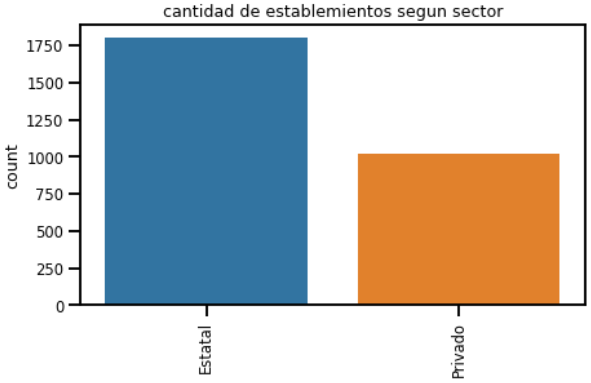


Gráfico 3 - Estatales y privados

Sin embargo, es interesante saber de que manera se encuentran distribuidos estos establecimientos en cada, pudiéndose visualizar donde predomina cada tipo:

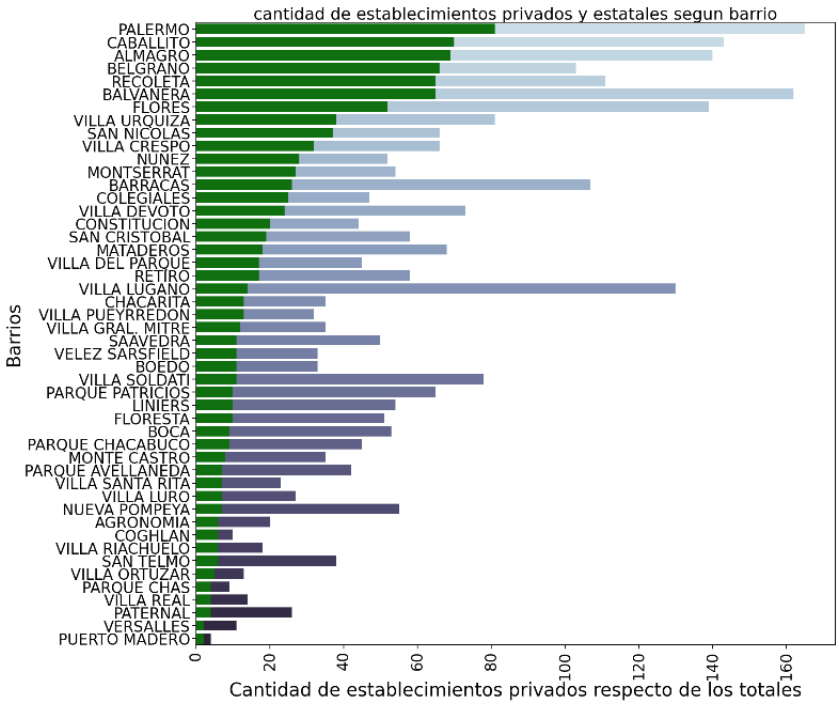


Gráfico 4 - Cantidad de privados por barrio

Por otro lado, resultaba interesante categorizar dichos establecimientos. El set de datos proporcionaba una categorización por niveles de educación demasiado exacta de las instituciones, que escapaba a las necesidades del análisis. A continuación, se puede ver la categorización original sin entrar en el detalle de cada uno, ya que se recategorizarán en grupos más abarcativos:

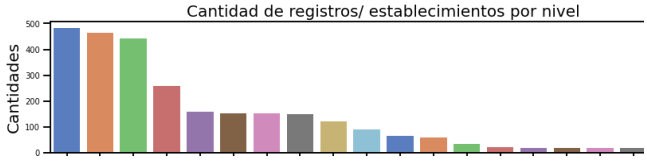
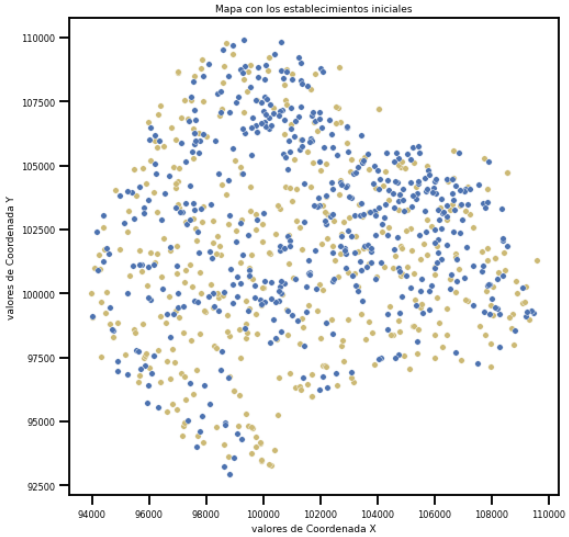


Gráfico 5 - Cantidad por nivel

Todas esas categorías de nivel pudieron en 4 finales: Educación Inicial, Primaria, Secundaria y Especial.

Para visualizar como se distribuyen a lo largo de la ciudad los establecimientos que se engloba dentro de cada categoría, se desarrollaron los siguientes mapas representativos:



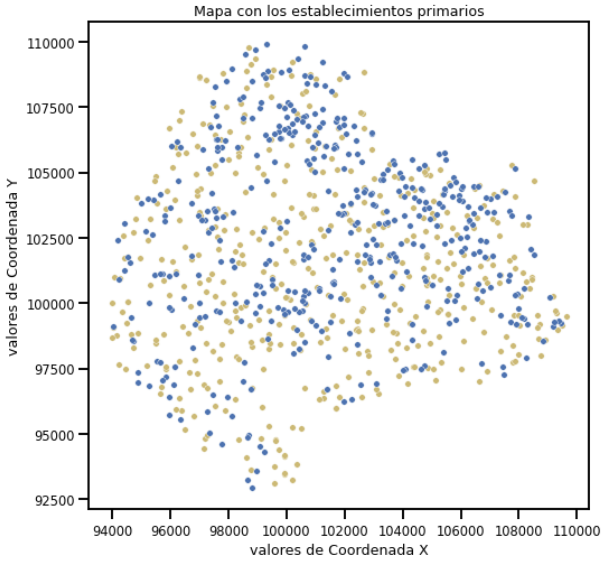
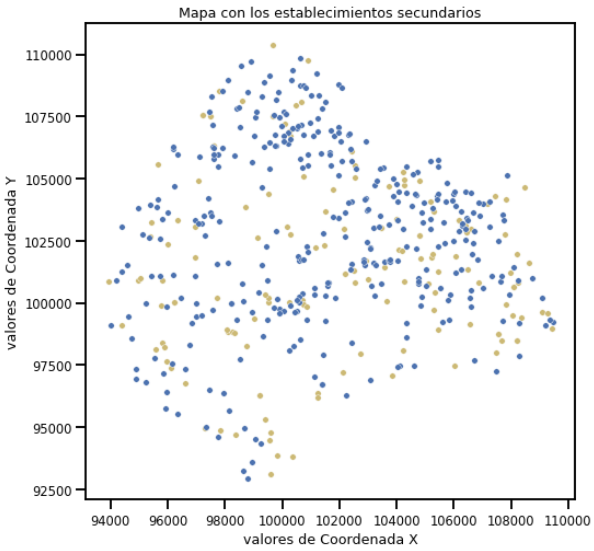
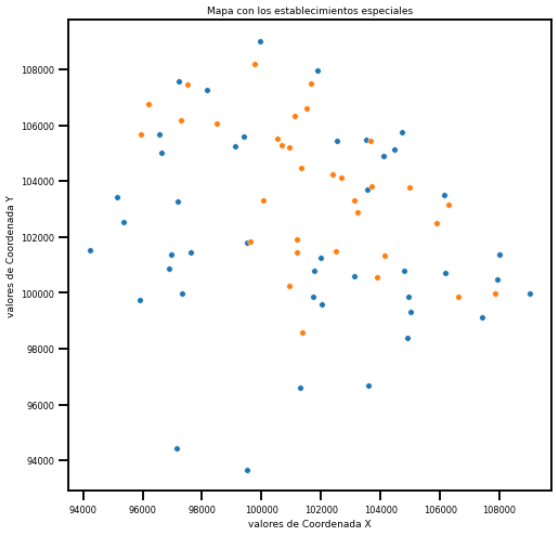


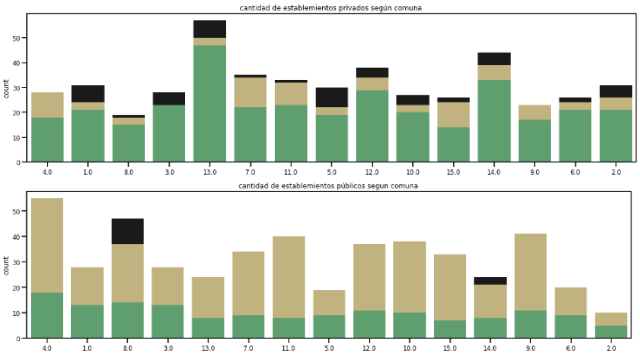
Gráfico 6 - Ubicación geográfica estatales y privados





Privado Estatal

Una vez conocida la distribución en la Ciudad de cada tipo de establecimiento, es interesante agruparlos según lo visto anteriormente (es decir, tipos de niveles y gestión) pero mostrándolos por la comuna en que se encuentran.



- **Inicial** • **Primario** • **Secundario**

5 MÉTODOS

Luego del análisis de los datos, se decidió llevar a cabo dos clasificaciones.

La primera, busca conocer de antemano si un establecimiento será público o privado según las coordenadas geográficas x e y que este posea.

La segunda, por el contrario, pretende saber en qué comuna estará ubicado el establecimiento según sus coordenadas x e y.

Para llevarlas a cabo, se utilizaron dos modelos de clasificación de aprendizaje supervisado

5.1 K-Nearest Neighbors

Más conocido como KNN, este modelo, clasifica un determinado dato, según tenga "K" vecinos más próximos a un grupo u a otro a partir de la "distancia" entre los elementos mencionados.

Nos pareció adecuado usar ese método para clasificar las instituciones por sus comunas correspondientes.

5.2 Support Vector Machines

El SVM, genera un hiperplano que pretende separar las diferentes clases que puedan existir.

6 EXPERIMENTOS Y RESULTADOS

6.1 Clasificación 1

En la primera clasificación nuestro objetivo fue lograr separar los establecimientos privados y estatales, dadas sus coordenadas espaciales para ellos requeríamos métodos como KNN y SVM como describimos antes.

6.2 Clasificación 2

En la segunda clasificación nuestro objetivo fue lograr separar los establecimientos por comunas dadas sus coordenadas espaciales, para ellos también recurrimos a modelos logrados con KNN y SVM como describimos antes.

	C1		C2	
	SVM	KNN	SVM	KNN
score	0,722	0,715	0,943	0,976
acc	0,722	0,717	0,943	0,977

Tabla 2 - Resultados obtenidos

7 DISCUCIÓN Y CONCLUSIONES

La clasificación inicial sobre los colegios privados y estatales nos arrojó una precisión demasiado baja ya que estos no se encuentran fácilmente separables.

Por lo anterior, la segunda clasificación buscaba realizar otro análisis dónde pudiésemos etiquetar las comunas de establecimientos solo por su coordenada, y pudiéramos identificar de mejor manera la potencialidad de las herramientas de clasificación

Sin embargo, los estudios realizados podrían servir a futuro para:

- Se puede observar que zonas tienen menos cantidad de escuelas y se podría correlacionar con la cantidad de habitantes de cada barrio para evaluar la posibilidad de agregar escuelas
- Evaluar necesidades espaciales que requieren más atención en cada barrio.

8 REFERENCIAS

- Data Science Handbook (VanderPlas)
- Introduction to Statistical Learning (Tibshirani)
- Deep Learning Book (Goodfellow)
- Elements of Statistical Learning (Tibshirani)