**Obras Públicas en la Ciudad de Buenos Aires**

Alan Barbera, Ailén Kot, Micaela Tokashiki

Universidad Tecnológica Nacional, Buenos Aires, Argentina

***Abstracto***

En el presente reporte se analizarán las obras que han sido finalizadas dentro del territorio de la Ciudad Autónoma de Buenos Aires (CABA), para luego generar un modelo de clasificación supervisada con el objetivo de predecir el plazo de realización de una obra en función de ciertas características de la misma, tales como ubicación, tipo de obra y demás.

***Palabras claves***

Obras públicas, CABA, construcción, infraestructura.

***1. Introducción***

En el panorama actual en el que nos encontramos, en donde los datos abundan, resulta un punto estratégico poder utilizarlos para la toma de decisiones fundamentada en algo más que en la intuición, en evidencia precisa y concreta.

Por ello es que surge el siguiente trabajo, cuya finalidad es predecir el plazo de realización de distintos tipos de obras públicas en el territorio de CABA. De esta forma, buscaremos conclusiones para facilitar la programación y planificación de nuevas obras a futuro, indicando algunas características de la misma, como el monto de la empresa contratista, la ubicación y el tipo de trabajo a realizar.

***2. Descripción del dataset***

El dataset utilizado para el presente trabajo fue extraído del portal de datos abiertos Buenos Aires Data, en la rama de urbanismo y territorio. El mismo tiene una frecuencia de actualización mensual. Para el desarrollo del reporte se obtuvo la versión actualizada a agosto 2020.

La fuente primaria y la mantención del dataset está a cargo de la Secretaría General y Relaciones Internacionales, la Subsecretaría de Gestión

Estratégica y Calidad Institucional y la Dirección General de Calidad Institucional y Gobierno Abierto.

Los datos iniciales al momento de realización del presente informe están compuestos por 1117 muestras o registros de obras públicas que pueden encontrarse finalizadas, en ejecución o en licitación. Además, cuenta con 36 features o características iniciales, entre ellas el beneficiario, el responsable, el nombre de la obra, el estudio ambiental, etc.

El portal de datos abiertos presenta una breve descripción sobre cada una de las features o características del dataset. De los cuales serán relevantes para nuestro informe los mencionados a continuación.

* Etapa (texto), indica el estado de evolución de la obra
* Tipo (texto), indica el tipo de obra
* Monto\_contrato (número entero), monto en pesos por el cual se firmó el contrato
* Comuna (número entero), indica el número de comuna en donde se encuentra localizada la obra
* Plazo\_meses (número entero) que es el plazo en el cual se concretará la obra

***3. Análisis Exploratorio de Datos***

Luego de importar el dataset y visualizar su dimensión inicial (1117 x 36) lo primero que se realiza es analizar la existencia de valores nulos o NaNs (Not A Number) existentes. Comenzaremos analizándolos por feature y calculando qué porcentaje representan los mismos respectos al total del universo de muestras. De esta forma, observamos que algunas de ellas poseen hasta un 99% de datos nulos, esto nos da un indicio de que se debe actuar con precaución para no eliminar muestras de más al momento de la limpieza de datos.

Los resultados del análisis en profundidad de cada una de las características que sí resultan de interés para el proyecto en particular fueron los siguientes

* Etapa

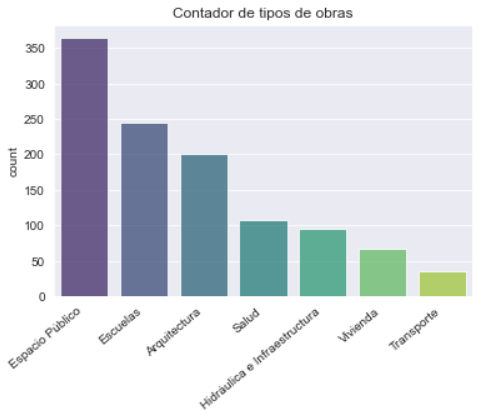
Las obras registradas pueden encontrarse en 3 etapas: finalizadas, en ejecución o en licitación.

Con el presente informe se busca lograr un modelo de aprendizaje supervisado, por lo tanto, la primer limitante que se encuentra es que se requiere que la obra haya sido finalizada para obtener los meses de duración de la obra, las cuales representan el 94% del dataset



* Tipo

En total hay 7 tipos de obras públicas registradas: escuelas, espacio público, vivienda, salud, arquitectura, hidráulica e infraestructura y transporte. Esta feature no posee valores nulos. Lo cual resulta positivo ya que será utilizada luego para realizar nuestro modelo de machine learning. Al tener data en todas las samples no habrá problemas cuando durante la limpieza, sin embargo, hay que tener cuidado al momento de aplicar cualquier modelo de aprendizaje ya que si eliminamos con el EDA aquellos tipos que tienen un porcentaje bajo (como por ejemplo Transporte con un 3.2%, es probable que la predicción resulte con un bajo accuracy)





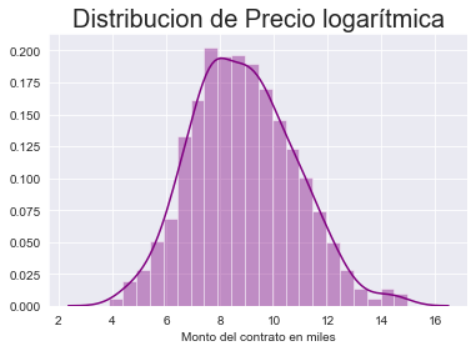
* Monto del contrato

En total hay 70 samples que no poseen el detalle del monto del contrato, de las cuales, 60 de ellas corresponden a obras ya finalizadas.

Se contempló la opción de reemplazar los valores ausentes por valores como la media de la feature en función del tipo de obra, para evitar una disminución de datos, sin embargo, se observó una gran dispersión en los montos del contrato con un desvío estándar elevado, debido a la gran amplitud entre el valor mínimo y máximo que puede obtener cada obra, por ello se decidió eliminar los valores nulos.

Para facilitar la visualización de este dato, también resulta conveniente generar una nueva variable en la que se guarde el monto del contrato expresado en miles.

Visualizando la distribución del monto de contrato (normalizado logarítmicamente) se observa una figura similar a la Campana de Gauss

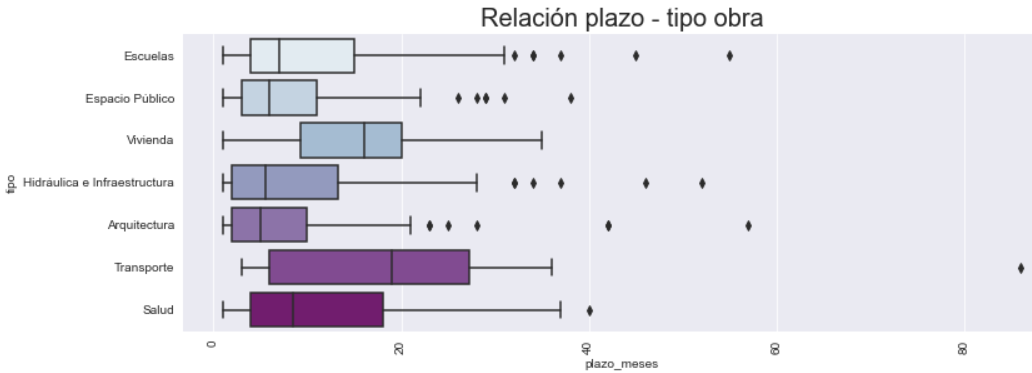


* Comuna

La comuna hace referencia a la ubicación geográfica de la obra, por ello, es un dato que no puede ser estimado o aproximado, al igual que con el monto del contrato se define eliminar aquellas samples cuyo valor de comuna sea nulo. Una alternativa en esta instancia sería realizar un modelo de clustering para asignar un número de comuna, sin embargo, esta opción fue descartada porque la misma también está sujeta a cierto error e incluirla en este momento del análisis implicaría un error aún mayor en la predicción del objetivo del trabajo.

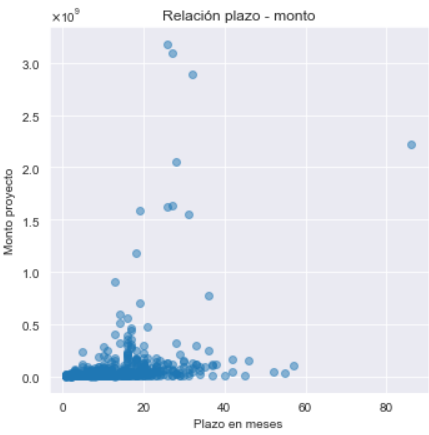
* Plazo en meses

Esta feature resulta crítica ya que será el target de nuestro modelo clasificador. Nuevamente, la política es de eliminar aquellas muestras cuyo valor sea nulo o un plazo igual a 0 meses.



El dato, originalmente está expresado en números, indicando la cantidad de meses en los que fue realizada la obra. Este punto será tratado en detalle al aplicar los modelos de machine learning.

Para concluir con el EDA, observamos la relación entre el plazo en meses y el monto de la obra



Así como el plazo y el tipo de obra

***4. Pre procesamiento y modelos de ML***

***5. Resultados / Métricas de performance***

***6. Conclusiones / Trabajo futuro***

***7. Referencias***