Predicción Plazo de Realización Obras Pública

Barbera Alan, Kot Ailen, Tokashiki Micaela

Universidad Tecnológica Nacional, Facultad Regional Buenos Aires

Introducción y objetivos

- El siguiente trabajo busca realizar un **análisis y clasificación de datos**, con el fin de efectuar predicciones precisas para futuras toma de decisiones
- El objetivo del mismo es estimar el plazo de construcción de distintos tipos de **obras públicas** en el territorio de CABA.
- En base a esos tiempos se podrá efectuar una **planificación y programación** de futuras obras a realizar
- Se partirá de características representativas de las mismas, como lo son, la **ubicación, tipo de trabajo, monto**, etc.

Métodos

- Para poder predecir el plazo de realización de la obra a partir de conocer el monto del contrato, el tipo de obra y su ubicación, utilizaremos estrategias de **aprendizaje supervisado** (labels conocidos) a partir de modelos clasificadores. Se llevan a cabo las siguientes etapas

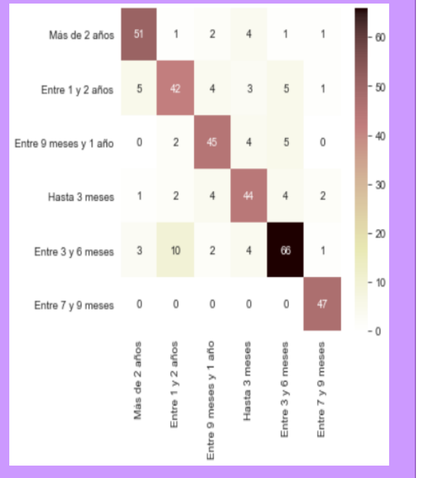
1. Se divide el dataset en train y test, para que el modelo aprenda los mejores parámetros. Se utilizó el método de **Standard Scaler**, donde al valor de la muestra le resta el valor de la media y lo divide por su desvío estándar para obtener una muestra normalizada.
2. Para obtener los mejores hiperparámetros para lograr una mejor precisión se utilizará el método de **Cross Validation (CV)** y **Grid Search (GS)**
3. Se usaran los siguientes modelos clasificadores:
   Support Vector Classifier(SVC); K-Nearest Neighbour Classifier (KNC)
   Random Forest Classifier; Logistic Regression

NOTA: Al tratarse de un dataset desbalanceado también se utilizó el método de 'resample' para agregar muestras similares en aquellas clases con poca información

Resultados

- Por medio de las herramientas anteriormente mencionadas, evaluamos cual es el de accuracy mas elevado

En meses		En año	
SVC	0,090468	SVC	0,712173
KNN	0,134087	KNN	0,80091
Random forest	0,150242	Random forest	0,847554
LR	0,108239	LR	0,63595
Trimestral sin Resample		**Trimestral Con Resample**	
SVC	0,321075	SVC	0,60929
KNN	0,340877	KNN	0,806011
Random forest	0,356436	Random forest	0,797814
LR	0,384724		



Datasets

- El Dataset utilizado fue extraído del portal de datos abiertos Buenos Aires Data, en la rama de **urbanismo y territorio.**
- El mismo cuenta con un total de **1117 muestras** y **36 features** (características iniciales), de cuales para este proyecto desacataremos

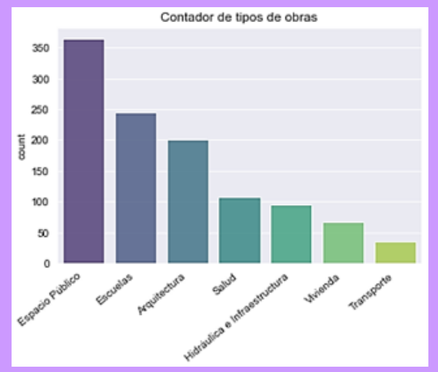
- Etapa: (texto), indica el estado de evolución de la obra
- Tipo: (texto), indica el tipo de obra
- Monto contrato:(número entero): monto en pesos por el cual se firmó el contrato
- Comuna: (número entero), indica el número de comuna en donde se encuentra localizada la obra}
- Plazo meses: (número entero)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

Análisis de datos

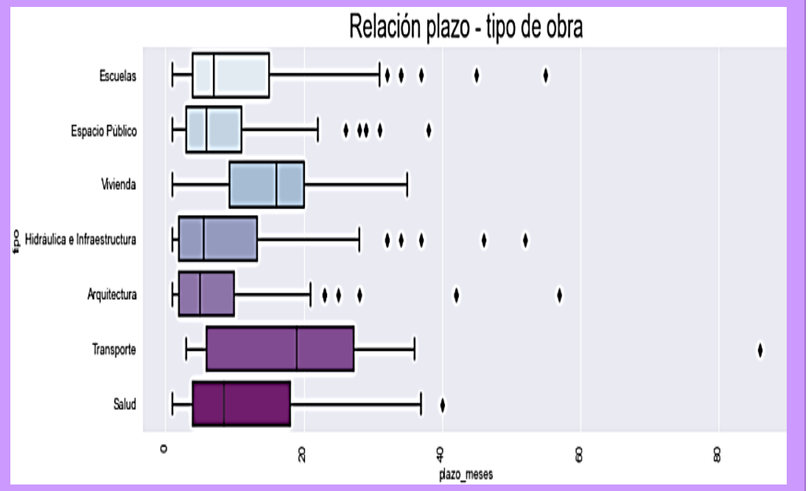
- Solo se toman en cuenta para el análisis las obras finalizadas

etapa	tipo	Porcentaje
Finalizada	1052	94.180842
En ejecución	61	5.461056
En licitación	4	0.358102

- Observamos cómo se distribuyen las muestras entre los distintos tipos de obras



- El plazo en meses resulta una característica crítica ya que será el target de nuestro modelo de aprendizaje supervisado
- Se grafica la relación de plazo con el tipo de obra y se observa que no tiene relación alguna



Conclusión

- Se comprobó los efectos de un dataset desbalanceado frente a la aplicación de un modelo de machine learning (clasificación por meses) así como una posible alternativa de solución (clasificación trimestral)
- Consideramos que la mejor alternativa de las 3 analizadas es la clasificación por trimestres. Si bien el accuracy en períodos anuales fue superior, se pierde nivel de detalle. Siendo más útil poder estimar y obtener información para la toma de decisiones si el rango de realización es de 3 meses.
- Recomendación para trabajar a futuro, sería en especificar aún más en detalles de la obra al momento de recolectar información, ejemplo, si el período es menor a 1 mes, indicar las semanas.