75.06/95.58 Organización de Datos
Primer Cuatrimestre de 2018
Trabajo Práctico 2: Enunciado

El segundo TP es una competencia de Machine Learning en donde cada grupo debe intentar determinar, para cada usuario presentado, cuál es la probabilidad de que se postule a un cierto aviso laboral.

La competencia se desarrolla en la plataforma de Kaggle, se proveen una serie de archivos en:

https://drive.google.com/file/d/1K4uRag5nmGtfuvzyJV9RL_73lzsh_iTO/view?usp=sharing
https://drive.google.com/file/d/1Pudf2TrUn_hfd8Dks4UTTJLf9ZdnGUd_/view?usp=sharing
https://drive.google.com/file/d/1ic7saV_7q-vpaUBkrta83nHRDn2d5qMb/view?usp=sharing
https://drive.google.com/file/d/1K7E7gxh6O24BHCGShXKM9cy4cRas7WeF/view?usp=sharing
(el ultimo es el detalle de los 338 avisos que originalmente faltaban)

que deben ser usados para entrenar un modelo de Machine Learning y un archivo "test_final_100k.csv" que contiene 100.000 rows con (id, idaviso, idpostulante y para el cual debe agregarse una nueva columna indicando la probabilidad de que exista una postulación para ese id. El formato del archivo resultante será entonces (id, probabilidad). Por favor notar que el archivo a submitir tiene el id, y la probabilidad, pero no incluye idaviso y idpostulante (estos estan determinados por el id).

El link a la competencia es https://www.kaggle.com/t/3917603da7044a8ba47cfc606a94e235

Los grupos deberán probar distintos algoritmos de Machine Learning para predecir cuál es la probabilidad de un postulante, a presentarse a un cierto aviso laboral publicado por una empresa. A medida que los grupos realicen pruebas deben realizar el correspondiente submit en Kaggle para evaluar el resultado de los mismos.

Al finalizar la competencia el grupo que mejor resultado tenga obtendrá 10 puntos para cada uno de sus integrantes que podrán ser usados en el examen por promoción o segundo recuperatorio.

Requisitos para la entrega del TP2:

- El TP debe programarse en Python o R.
- Debe entregarse una carpeta con el informe de algoritmos probados, algoritmo final utilizado, transformaciones realizadas a los datos, feature engineering, etc.

- El grupo debe presentar el TP en una computadora en la fecha indicada por la cátedra, el TP debe correr en un lapso de tiempo razonable (inferior a 1 hora) y generar un submission válido que iguale el mejor resultado obtenido por el grupo en Kaggle.

El TP2 se va a evaluar en función del siguiente criterio:

- Cantidad de trabajo (esfuerzo) del grupo: ¿Probaron muchos algoritmos? ¿Hicieron un buen trabajo de pre-procesamiento de los datos y feature engineering?
- Resultado obtenido en Kaggle (obviamente cuanto mejor resultado mejor nota)
- Presentación final del informe, calidad de la redacción, uso de información obtenida en el TP1, conclusiones presentadas.
- Performance de la solución final.