75.06/95.58 Organización de Datos
Primer Cuatrimestre de 2017
Trabajo Práctico 2: Enunciado

El segundo TP es una competencia de Machine Learning en donde cada grupo debe intentar predecir la duración de los viajes en base a los datos de los mismos.

La competencia se desarrolla en la plataforma de Kaggle, se provee un archivo "train.csv"que debe ser usado para entrenar un modelo de Machine Learning y un archivo "test.csv" que tiene los datos de los viajes a predecir. Adicionalmente pueden usarse los datos de las estaciones, del status de cada estación minuto a minuto y la información meteorológica.

ADVERTENCIA MUY IMPORTANTE: Dado que los datos son públicos los resultados de la competencia ya se saben, son parte de los datos del TP1, es fundamental que ningún grupo suba a Kaggle submissions que usan esta información ya que distorsiona el score de la competencia y una vez subido un submission es IMPOSIBLE ELIMINARLO. Los submissions deben generarse en base a un modelo de machine learning y nunca en base a los resultados que ya se conocen. TL;DR: NO SUBIR A KAGGLE SUBMISSIONS QUE HACEN TRAMPA, ES IMPOSIBLE BORRARLOS!!!

Los grupos deberán probar distintos algoritmos de Machine Learning para predecir la duración de los viajes en base a los datos de los mismos. 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.

ADVERTENCIA IMPORTANTE #2: Bajo ningún concepto debe interpretarse que es necesario finalizar el TP1 para poder comenzar el TP2, quienes incurran en este error se encontrarán que el tiempo necesario para desarrollar el TP2 es insuficiente. Es fundamental, imprescindible y vital comenzar el desarrollo del TP2 en forma paralela al TP1 para evitar problemas en el cumpimiento de las fechas de entrega.