Tarea 1. Machine Learning.

Profesor: Julio Erasmo Godoy Del Campo (jugodoy@inf.udec.cl)

Ayudante: Daniel Moreno Cartagena (dmoreno2016@inf.udec.cl)

Fecha de entrega: **Por determinar**

¿Cómo debe subir la tarea?: La tarea debe ser subida a la plataforma de Canvas en un archivo con el formato **NombreApellido_tarea1.zip**. Este debe contener el código solicitado junto con los datos utilizados.

Nota: Recuerde que este es un proceso iterativo, no lineal, por ende, es muy probable que tenga que volver a pasos anteriores para modificar ciertas cosas.

Problema 1: Clasificación de estudiantes

El set de datos contiene información respecto a los estudiantes que han ingresado a la facultad de ingeniría. Se le pide que implemente un modelo que permita aprender a clasificar a que cluster pertenece cada alumno de acuerdo a la información que se entrega. Recuerde que por convención la última columna del set de datos representa la etiqueta que usted quiere que el modelo aprenda. También tenga en cuenta que las clases están desbalanceadas y que se espera que usted utilice 60% de los datos para entrenar, el 20% restante para validar y el 20% para testear. Además, presente una matriz de confusión y el f1-score sobre los datos de validación y test.

Limpieza y exploración de datos

Si dentro de esta etapa toma ciertas decisiones sobre el conjunto de datos, déjelas por escrita.

1.1 Grafique la distribución de clases.
1.2 Identifique valores faltantes.
1.3 Preprocese los datos para entrenar un algoritmo de Machine Learning.
1.4 Divida el conjunto de datos en entrenamiento, validación y test.

Modelamiento

Nota 1: Recuerde que el dataset está desbalanceado, el cual es un problema típico en la práctica.

Se espera que entrene el modelo con los datos de entrenamiento y busqué los mejores hiperparametros sobre el conjunto de validación. Recuerde utilizar las métricas vistas en clases.

1.5 Entrene el o los algoritmos sobre el conjunto de entrenamiento.
1.6 Valide sus resultados.

Modelo en producción

Luego de asegurarse de que encontró los mejores hiperparametros para el modelo, debe ponerlo en producción. Para esto se utilizarán los datos de test (es muy importante que estos datos no hayan

sido utilizados antes), ya que esto permite evaluar el desempeño de su modelo sobre datos que no se tienen disponibles en la práctica.

1.7 Pruebe el algoritmo sobre datos con los que no ha trabajado antes (datos de test).
1.8 Concluya los resultados obtenidos, tanto para el conjunto de validación, como para el conjunto de test.

Problema 2: Prediciendo la demanda de taxis en NYC

En este problema, construiremos un modelo de regresión para predecir el número de taxis solicitados en la ciudad de Nueva York. Estos modelos suelen ser útiles, por ejemplo, para monitorear el tráfico en la ciudad.

Los datos para este problema se encuentran en el archivo "nyc_taxi.csv". La primera columna indica la hora del día en minutos, mientras que la segunda columna indica la cantidad de taxis que están recogiendo pasajeros en ese momento.

Se necesita un modelo que reciba la hora en minutos como predictor y prediga la demanda promedio de taxis para ese momento del día. Los modelos deben ser ajustados en el conjunto de entrenamiento, validados sobre el conjunto de validación y evaluados sobre el conjunto de test. Además, la métrica de evaluación debe ser el R^2.

2.1 Cargue los datos y describa la distribución de los atributos.
2.2 Separe el conjunto de entrenamiento.
2.3 Grafique el conjunto de entrenamiento utilizando un scatter plot.
2.4 Brevemente explique el patrón de comportamiento de los datos a lo largo del día.
2.5 En el grafico debería ver un agujero entre los minutos 500 y 550 donde la demanda es aproximadamente 20-30. Explique el fenómeno.
2.6 Entrene un modelo de regresión.
2.7 Reporte y explique el valor de R^2.
2.8 Que significa un R^2 = 0? y si fuese negativo?