UNIVERSIDAD DE GRANADA
E.T.S. de Ingenierías Informática y de Telecomunicación

Departamento de Ciencias de la Computación e Inteligencia Artificial

Inteligencia de Negocio

Guión de Prácticas

Práctica 3: Competición en Kaggle

Curso 2017-2018

Cuarto Curso del Grado en Ingeniería Informática

Práctica 3

Competición en Kaggle

1. Objetivos y Evaluación

En esta tercera práctica de la asignatura Inteligencia de Negocio veremos el uso de métodos avanzados para aprendizaje supervisado en regresión sobre una competición real en la plataforma Kaggle (`https://www.kaggle.com/`). El estudiante adquirirá destrezas para mejorar la capacidad predictiva del modelo mientras se familiariza con una de las plataformas de competición en Ciencias de Datos más extendida hoy día.

La evaluación de la práctica se dará en función de la posición final (relativa al conjunto de estudiantes participantes) que ocupe el resultado propuesto por el estudiante con el siguiente criterio:

puesto	1°	2°	3°	4°	5°	...	último
puntuación	2	2	1,75	1,5	1,25	...	0,5

Las posiciones serán linealmente proporcionales entre los 1,25 puntos del 5° y los 0,5 puntos del último. Para ser evaluado no bastará con subir los resultados a Kaggle, se deberá también adjuntar un documento que describa el proceso seguido por el estudiante para resolver la práctica y demostrar mediante la actividad registrada en Kaggle que ha habido un esfuerzo por mejorar los resultados. En otro caso, el alumno no obtendrá ninguna puntuación en esta práctica.

Sobre la puntuación obtenida en base a la posición, se aplicará un factor corrector $[0{,}5, 1{,}5]$ (es decir, se podrá reducir o aumentar hasta un 50 %) en función de la calidad de la documentación presentada y las soluciones abordadas.

2. Descripción del Problema y Tareas

La competición será la *House Prices: Advanced Regression Techniques* disponible en `https://www.kaggle.com/c/house-prices-advanced-regression-techniques`. El objetivo es predecir el precio final de una vivienda a partir de 79 variables que describen diferentes aspectos

de ella con información sobre su ubicación, barrio, construcción, equipamiento, etc. Se trata de datos reales recogidos entre 2006 y 2010. Es un problema de regresión donde la variable dependiente a predecir es continua. La métrica de evaluación consiste en la raíz del error cuadrático medio sobre el logaritmo del valor real y el logaritmo del valor predicho (RMSLE).

La competición es de un nivel *"Getting Started"*, siendo por tanto un entorno ideal para aprender en ciencia de datos. Existen centenares de tutoriales y *scripts* que pueden servir de ayuda en la sección *Kernels* de la competición. Por ejemplo:

- Análisis exploratorio de los datos:
  `https://www.kaggle.com/pmarcelino/comprehensive-data-exploration-with-python`
- Punto de partida para una primera solución (disponible en la web de la asignatura):
  `https://www.kaggle.com/neviadomski/how-to-get-to-top-25-with-simple-model-sklearn`

En esta competición se permite el uso de cualquier software, algoritmo o lenguaje que el alumno considere útil. Está terminantemente prohibido usar el precio de la vivienda en los datos de *test* para entrenar, configurar o mejorar el modelo predictivo. Cualquier indicio de esta conducta supondrá la anulación de la práctica.

3. Documentación

La documentación explicará las estrategias seguidas y el progreso que se ha ido desarrollando durante la competición. Deberán razonarse brevemente los diferentes pasos tomados apoyándose en visualización de datos u otras técnicas de análisis para comprender las características del problema. Se recomienda añadir también extractos de los *scripts* para explicar el trabajo realizado. Será obligatorio incluir una tabla que contenga tantas filas como soluciones se han subido a Kaggle incluyendo columnas que resuman cada experimento conteniendo, al menos:

- la fecha y hora de subida a Kaggle,
- la posición que ocupó en ese momento,
- el *score* obtenido en Kaggle al subir la predicción en *test*,
- la RMSLE sobre el conjunto de datos de entrenamiento,
- breve descripción del preprocesado realizado,
- breve descripción de el/los algoritmo(s) de regresión empleado(s) y
- configuración de parámetros de esos algoritmos.

La ausencia de esta tabla o una descripción incompleta de la misma supondrá la anulación de la práctica.

De cada subida realizada a Kaggle se conservará el fichero `.csv` y el *script* en Python o similar usado para ese experimento. Se nombrarán de forma clara y enumerada para poder identificar con claridad a qué experimento de la tabla corresponde. Este material se entregará junto a la documentación.

El alumno deberá definir como "Team Name" en Kaggle su nombre de pila y primer apellido terminando con (UGR). Por ejemplo: `Jorge Casillas (UGR)`.

4. Entrega

La competición en Kaggle finaliza el jueves **4 de enero de 2018** a las 23:59. Cualquier subida a Kaggle posterior a esa fecha supondrá la anulación de la práctica.

Tras acabar la competición, el estudiante deberá también entregar antes del 10 de enero de 2018 a las 23:59 una documentación que explique las tareas realizadas y todas las soluciones `.csv` subidas a Kaggle junto son los *scripts* utilizados.

Este material se entregará a través de la web de la asignatura en `https://decsai.ugr.es` en un único archivo `zip`. Por ejemplo, la estudiante "María Teresa del Castillo Gómez" subirá el archivo `P3-delCastillo-Gómez-MaríaTeresa.zip`. La documentación, contenida en ese mismo archivo `zip`, tendrá el mismo nombre pero con extensión `pdf`.