UNIVERSIDAD DE GRANADA

UNIVERSIDAD DE GRANADA

Departamento de Ciencias de la Computación e Inteligencia Artificial

Inteligencia de Negocio

Guión de Prácticas

Práctica 3: Competición en Kaggle.

Curso 2020-2021

Cuarto Curso del Grado en Ingeniería Informática

Sobre esta práctica

1. Objetivos y Evaluación

En esta tercera y última práctica de la asignatura Inteligencia de Negocio veremos el uso de métodos avanzados para aprendizaje supervisado en clasificación sobre una competición disponible en *Kaggle*, https://www.kaggle.com, creada *ex-proceso* para esta práctica. El/La estudiante adquirirá destrezas para mejorar la capacidad predictiva del modelo mientras se familiariza con una de las plataformas más populares de competición en ciencias de datos.

La práctica se calificará hasta un *máximo de 3 puntos*. La posición en la competición dará desde 2 puntos hasta el último puesto, que obtiene 0,75. Para ser evaluado no bastará con subir los resultados a Kaggle, si no que se deberá de adjuntar también un mínimo documento que describa el proceso seguido por el/la estudiante para resolver la práctica, y demostrar que ha habido un esfuerzo por obtener buenos resultados. En otro caso, el/la estudiante no obtendrá ninguna puntuación en esta práctica. Se aplica un factor de corrección de hasta un 50 % para mejorar o empeorar la puntuación en función de la calidad de la documentación y trabajo realizado.

2. Descripción del Problema y Tareas

La competición es especialmente creada para la práctica, y está disponible en https://www.kaggle.com/c/ugrin2020-vehiculo-usado-multiclase/ aunque para poder participar hay que entrar de forma identificada mediante el siguiente enlace https://www.kaggle.com/t/d9f6a48447ae4ba88c4fecda442a8b52. En este problema existen datos de una serie de coches vendidos, y se han clasificado los coches en 5 categorías de precio. Se desea predecir la categoría del precio del coche, por lo que es un problema de clasificación multiclase.

El conjunto de entrenamiento consta de alrededor de 6000 instancias, y 14 atributos (de los cuales, *id* toma valores únicos y solo sirve para identificar cada ejemplo) con datos categóricos y enteros. Se trata de predecir la variable ordinal *Precio_cat*, que representa el grado de coste que supone. Hay cinco valores: 1, representa a los más baratos; 2, representa aquellos baratos pero menos; 3, representa a los que están en precio promedio; 4, representa a los que son más caros que el promedio, y 5, que representa a los coches más caros. Para medir el rendimiento de nuestros algoritmos, la competición usará la medida de precisión (*accuracy*), aunque se podrá utilizar otras medidas o criterios para identificar los algoritmos más promedores en la memoria.

2.1. Descripción de los atributos

A continuación se detalla el significado de los distintos atributos:

Nombre del tipo de coche.

Ciudad de la venta del coche.

Año del coche.

Kilometros recorridos del coche.

Combustible del coche (*Petrol*, *Diesel*, *Electric*, *LPG* y *CNG*).

Tipo de marcha del coche (*Manual* y *Automatic*).

Mano si es primera mano (*First*), Segunda (*Second*), Tercera (*Third*) y cuarta o más (*Fourth & Above*).

Consumo en kilómetros por litro (kmpl).

Motor CC medido en centrímetros cúbicos (CC).

Potencia del motor medido (bhc).

Asientos número de plazas máximas del coche.

Descuento descuento realizado por oferta especial (en porcentaje).

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 el experimentos subidos, al menos:

- La fecha y hora de subida a Kaggle.
- La posición que ocupó en ese momento.
- El *score* sobre el conjunto de datos de entrenamiento (aplicando validación cruzada sobre dicho conjunto).
- El *score* obtenido al subir la predicción en *test*.
- Breve descripción del preprocesado realizado.
- Breve descripción de el/los algoritmos(s) de clasificación empleado(s).
- Configuración de parámetros de esos algoritmos.

La ausencia de la tabla o una descripción demasiado incompleta supondrá la anulación de la práctica. Adicionalmente, la segunda página de la documentación (después de la portada y antes del índice) contendrá una captura de pantalla de *Leaderboard* (mostrando la fila con los valores de *Team Name* y *Score* del estudiante), disponible en la web de *Kaggle* en el apartado de *Leaderboard*.

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 facilidad a qué experimento de la tabla corresponde. Este material se entregará junto a la documentación. El alumno deberá definir como equipo en su nombre de pila y primer apellido terminando con el DNI.

4. Competición

Para trabajar se dispone de varios ficheros descargables en la página web de la competición:

train.csv Fichero con todos los atributos, incluyendo el objetivo a predecir. Es el conjunto de datos que se puede usar para realizar el aprendizaje automática. Además, se deberá de evaluar usando validación cruzada para identificar los algoritmos más prometedores, y realizar el proceso de *tuning* que se considere. Se deberá de aplicar las técnicas de pre-procesamiento que se consideren interesantes.

test.csv Fichero con las instancias a predecir. Posee el mismo formato que **train.csv**, a excepción del atributo *Precio_cat*, que evidentemente no aparece.

sample.csv Fichero con el formato del fichero a someter, que contiene únicamente el id y los valores de *Precio_cat*.

De cara a facilitar una correcta etiquetación y normalización (que debería de aplicarse por igual a los datos de ambos ficheros) se ofrece un fichero para cada atributo, ej: *combustible.csv*, *nombre.csv*, ...

Una vez identificado el algoritmo adecuado, y con el preprocesamiento deseado, se deberá de predecir las instancias del fichero *test.csv*, y se deberá de guardar los valores predichos (con el mismo formato que el fichero *sample.csv*) y subirlo a Kaggle en la competición. Una vez sometido el fichero con los resultados predichos, Kaggle calculará la medida de precisión, y lo mostrará de forma ordenará según dicha medida.
Límites: Sólo se permitirá 3 subidas por día, y un total de 4 ó 5. Aunque el sistema Kaggle no impone límite total, se penalizará mucho superar dicho número en la calificación. Por tanto, hay que tenerlo en cuenta.

5. Entrega

La fecha límite de la competición será el miércoles **1 de Enero** de 2020 hasta las **23:59**, y para la entrega de la memoria hasta el día **4 de Enero** de 2020 hasta las **23:50** por medio de

PRADO.

Para participar en la competición hay que darse de alta en Kaggle `https://www.kaggle.com` identificándose como equipo (*team*) con su nombre completo y DNI. En la tarea de PRADO hay que subir un único fichero `zip` se incluirá la documentación que exploque las tareas realizadas y todas las soluciones .csv subidas a *Kaggle* junto con los *scripts* de Python empleados. El nombre del archivo `zip` será el siguiente (sin espacios): `P3-apellido1-apellido2-nombre.zip`. La documentación tendrá el mismo nombre pero con extensión `pdf`. Es decir, la alumna "María Teresa del Castillo Gómez" subirá el archivo `P3-delCastillo-Gómez-MaríaTeresa.zip` que contendrá, entre otros, el archivo `P3-delCastillo-Gómez-MaríaTeresa.pdf`.