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 DrivenData

Curso 2021-2022

Grado en Ingeniería Informática
Grado en Ingeniería Informática y Matemáticas
Grado en Ingeniería Informática y Administración y Dirección de Empresas

Práctica 3

Competición en DrivenData

1. Objetivos y Evaluación

En esta tercera práctica de la asignatura Inteligencia de Negocio veremos el uso de métodos avanzados para aprendizaje supervisado y preprocesado en clasificación sobre una competición real disponible en DrivenData (`https://www.drivendata.org/`). 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 que está ganando gran interés por dirigir la resolución de problemas al bien social.

La práctica se calificará hasta un *máximo de 3 puntos*. La evaluación se dará en parte en función de la posición final (relativa al conjunto de estudiantes participantes) que ocupe el resultado propuesto por el estudiante, con una asignación linealmente proporcional entre 2 puntos para la primera posición y 0,5 puntos para el último. Para ser evaluado, no bastará con subir los resultados a DrivenData, 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 DrivenData 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 *Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines* disponible en `https://www.drivendata.org/competitions/66/flu-shot-learning/page/211/`.

El objetivo es predecir si una persona se vacuna contra la gripe H1N1 y la gripe estacional utilizando información sobre sus antecedentes, opiniones y comportamientos de salud. En este reto, se analiza la vacunación, una medida clave de salud pública utilizada para luchar contra las enfermedades infecciosas. Las vacunas proporcionan inmunización a los individuos, y una

inmunización suficiente en una comunidad puede reducir aún más la propagación de enfermedades a través de la "inmunidad de rebaño". En octubre de 2009 se hizo pública una vacuna contra el virus de la gripe H1N1. A finales de 2009 y principios de 2010, Estados Unidos realizó la Encuesta Nacional sobre la Gripe H1N1 2009. Esta encuesta telefónica preguntaba a los encuestados si habían recibido las vacunas contra la gripe H1N1 y la gripe estacional, junto con preguntas sobre ellos mismos. Estas preguntas adicionales abarcaban sus antecedentes sociales, económicos y demográficos, sus opiniones sobre los riesgos de la enfermedad y la eficacia de la vacuna, y sus comportamientos para mitigar la transmisión. Una mejor comprensión de cómo estas características se asocian con los patrones de vacunación personal puede proporcionar orientación para los futuros esfuerzos de salud pública.

El conjunto de entrenamiento consta de 26.707 instancias y 36 atributos (de los cuales, `respondent_idbuilding` toma valores únicos y solo sirve para identificar cada ejemplo) categóricos, enteros y binarios.

Para este concurso, hay dos variables objetivo:

- `h1n1_vaccine` - Si el encuestado recibió la vacuna contra la gripe H1N1.
- `seasonal_vaccine` - Si el encuestado recibió la vacuna contra la gripe estacional.

Ambas son variables binarias: 0 = No; 1 = Sí. Algunos encuestados no recibieron ninguna de las dos vacunas, otros recibieron sólo una, y algunos recibieron ambas. Esto se formula como un problema multietiqueta (y no multiclase).

El rendimiento se evaluará según el área bajo la curva ROC (AUC) para cada una de las dos variables objetivo. La media de estas dos puntuaciones será la puntuación global. Un valor más alto indica un mayor rendimiento. En Python, se puede calcular esto utilizando `sklearn.metrics.roc_auc_score` para esta configuración multietiqueta con el parámetro por defecto `average='macro'`.

En esta competición se permite el uso de cualquier *software*, algoritmo o lenguaje que el alumno considere útil. Está terminantemente prohibido usar la clase en los datos de *test*, en caso de conocerse, para entrenar, configurar o mejorar el modelo predictivo. También se prohíbe que los alumnos compartan entre ellos soluciones de la competición. Cualquier indicio de estas conductas 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 DrivenData incluyendo columnas que resuman cada experimento conteniendo, al menos:

- la fecha y hora de subida a DrivenData,
- la posición que ocupó en ese momento,
- el *score* sobre el conjunto de datos de entrenamiento,
- el *score* obtenido en DrivenData al subir la predicción en *test*,
- breve descripción del preprocesado realizado,
- breve descripción de el/los algoritmo(s) de clasificación/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.

Adicionalmente, la segunda página de la documentación (después de la portada y antes del índice) contendrá una **captura de pantalla de la tabla** *Submissions* (que contiene las columnas *Score*, *Submitted by* y *Timestamp*) disponible en la web de DrivenData para cada usuario.

De cada subida realizada a DrivenData 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 usuario en DrivenData su nombre de pila y primer apellido terminando con `_UGR_IN`. Por ejemplo: `Jorge_Casillas_UGR_IN`. La posición que se valorará para la práctica será la mejor de la cuenta `xxx_UGR_IN`, no se aceptará ninguna otra. Se recuerda al alumnado que DrivenData incorpora mecanismos para detectar múltiples inscripciones en la competición de la misma persona y bloquea las cuentas cuando sucede, impidiendo así seguir compitiendo.

4. Entrega

La competición en DrivenData finaliza el miércoles **5 de enero de 2022** a las 23:59. Cualquier subida a DrivenData posterior a esa fecha supondrá la anulación de la práctica. Tras acabar la competición, el estudiante deberá también entregar antes del viernes **7 de enero de 2022** a las 23:59 una documentación que explique las tareas realizadas y todas las soluciones `.csv` subidas a DrivenData junto con los *scripts* utilizados.

Este material se entregará a través de la web de la asignatura en `https://prado.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`.