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

UNIVERSIDAD DE GRANADA

Departamento de Ciencias de la Computación e Inteligencia Artificial

Sistemas Inteligentes para la Gestión en la Empresa

Guión de Prácticas

Práctica 1: Competición en Kaggle sobre Clasificación Binaria

Curso 2016-2017

Máster Profesional en Ingeniería Informática

Práctica 1

Competición en Kaggle sobre Clasificación Binaria

1. Objetivos y Evaluación

En esta primera práctica de la asignatura Sistemas Inteligentes para la Gestión en la Empresa veremos el uso de algoritmos de aprendizaje supervisado en clasificación, una tarea muy habitual al abordar problemas reales en *business analytics*. Se hará uso de la plataforma Kaggle (`https://www.kaggle.com/`) para evaluar los resultados. Los estudiantes adquirirán destrezas en análisis de datos y uso de algoritmos de predicción, y se familiarizarán con una de las plataformas de competición en Ciencias de Datos más extendida hoy día.

La práctica se desarrollará de forma individual. La evaluación se realizará en función de: (1) la posición final que ocupe el resultado propuesto por el estudiante (posición relativa respecto al conjunto de estudiantes); (2) la calidad de la memoria presentada. 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. La práctica calificará para el **50 %** de la puntuación de prácticas; esto es, 3 puntos sobre 6.

2. Descripción del Problema y Reglas de la Competición

Se trabajará sobre la competición "Titanic: Machine Learning from Disaster":

`https://www.kaggle.com/c/titanic`

El problema consiste en predecir si un pasajero del Titanic sobrevive o no en función de una serie de variables relativas a la edad, gánero, familia o tipo de pasaje. Se trata de una competición "Getting Started" que ayuda al alumno a iniciarse en el campo de la Ciencia de Datos y en el uso de la plataforma Kaggle. Existen numerosos tutoriales de diferente complejidad y para distintos lenguajes de progamación.

Como primer paso para la realización de la práctica, cada estudiante deberá registrarse en Kaggle. La práctica se debe resolver individualmente, por lo que en este caso los equipos tendrán un solo miembro. El nombre del usuario y del equipo usado en Kaggle deberá comunicarse al profesor de prácticas para poder realizar su seguimiento.

3. Entrega

Se podrá competir en Kaggle hasta el miércoles **19 de abril de 2017**. No se aceptará ninguna práctica cuya solución más reciente subida a Kaggle supere esa fecha. Una vez finalizada la competición, deberá realizar la siguiente entrega antes del lunes **24 de abril de 2017 a las 23:59** a través de la web de la asignatura en `https://decsai.ugr.es` en un único archivo `zip` con el nombre (sin espacios): `P1-apellido1-apellido2-nombre.zip`. Es decir, la estudiante "María Teresa del Castillo Gómez" subiría el archivo: `P1-delCastillo-Gómez-MaríaTeresa.zip`. El documento `pdf` con la memoria tendrá el mismo nombre.

Este archivo deberá documentar en detalle el trabajo realizado, aportando tablas, gráficas y cualquier material de apoyo. Se deberán incluir, al menos, los siguientes apartados:

1. Portada: Incluirá el nombre del estudiante, nombre del equipo usado en Kaggle, ranking global del equipo en la competición, puntuación del equipo.

2. Exploración de datos: Descripción y discusión de las técnicas utilizadas para estudiar la estructura y la semántica de los datos, incluyendo visualización, y los hallazgos preliminares, así como discusión y justificación de decisiones iniciales sobre el proceso de análisis que se llevará a cabo.

3. Preprocesamiento de datos: Descripción y discusión de las técnicas de preprocesamiento utilizadas y análisis crítico de su utilidad en el problema.

   - Integración y detección de conflictos e inconsistencias en los datos: valores perdidos, valores fuera de rango, ruido, etc.
   - Transformaciones: normalización, agregación, generación de características adicionales, etc.
   - Reducción de datos: técnicas utilizadas para selección de características, selección de ejemplos, discretización, agrupación de valores, etc.

   Se valorará el uso de técnicas para procesamiento de datos en clases no balanceadas y la evaluación de su utilidad en la obtención de mejores soluciones para el problema.

4. Técnicas de clasificación: Discusión de las técnicas de clasificación empleadas y justificación de su elección.

5. Presentación y discusión de resultados: Descripción y discusión de las soluciones obtenidas, incidiendo en la interpretación de los resultados. Análisis comparativo en caso de utilizar diferentes técnicas y/o parámetros de configuración en diferentes aproximaciones.

6. Conclusiones y trabajo futuro: Breve resumen de las técnicas aplicadas y de los resultados obtenidos, así como ideas de trabajo futuro para continuar mejorando las soluciones desarrolladas.

7. Listado de soluciones: Tabla de soluciones, incluyendo una fila por cada solución subida a Kaggle durante la competición. El número de filas deberá coincidir con el número de intentos reflejado en la web de la competición. En cada fila se aportará, al menos, la siguiente información separada por columnas:

   a) número de solución,

   b) descripción breve del preprocesamiento de datos aplicado,

   c) enumeración de los algoritmos y software empleados,

   d) resultado de porcentaje de acierto sobre conjunto de ejemplos etiquetados (extraído a partir del conjunto de entrenamiento, que puede ser el valor medio si se aplica validación cruzada u otro método similar),

   e) resultado de porcentaje de acierto sobre conjunto de ejemplos no etiquetados (*public score* obtenido en Kaggle),

   f) posición ocupada en el ranking de Kaggle en el momento de subir la solución (calculada seleccionando solo esta solución para *final score*) y fecha/hora de la subida.

   En esta tabla se debe resaltar la mejor solución obtenida; normalmente será la más reciente.

8. Bibliografía: Bibliografía utilizada. En particular, se deben incluir los tutoriales, blogs y similares sobre la competición Titanic en los que se basan las soluciones aportadas.