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

Departamento de Ciencias de la Computación e Inteligencia Artificial

Inteligencia de Negocio

Guion de Prácticas

Práctica 1: Análisis Predictivo Mediante Clasificación

Curso 2021-2022

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

Práctica 1

Análisis Predictivo Mediante Clasificación

1. Objetivos y Evaluación

En esta primera práctica de la asignatura Inteligencia de Negocio veremos el uso de algoritmos de aprendizaje supervisado de clasificación como herramienta para realizar análisis predictivo en una empresa u organización. En ella el alumno adquirirá capacidades para abordar problemas reales donde la ciencia de datos puede aportar valor en forma de conocimiento para ayudar en la toma de decisiones. Se trabajará con varios conjuntos de datos reales sobre los que se emplearán diferentes algoritmos de clasificación, para su comparación, y a la luz del conocimiento descubierto se podrán concluir estrategias para resolver cada problema. Para ello, se deberán crear informes de resultados y análisis lo suficientemente profundos para resultar de utilidad.

La práctica se calificará hasta un **máximo de 2 puntos**. Se valorará el acierto en los recursos de análisis gráficos empleados, la complejidad de los experimentos realizados, la interpretación del conocimiento extraído, la organización y redacción del informe, etc.

2. Problemas Abordados

En esta práctica trabajaremos con cuatro problemas cuyos conjuntos de datos están disponibles en la web de la asignatura. Estos problemas combinan distintas propiedades, existiendo casos de clasificación binaria y multiclase, clases balanceadas o no, atributos nominales y numéricos, existencia o no de valores perdidos, etc.

A continuación se enumeran estos problemas, indicando la fuente en cada caso para una mejor comprensión del problema a estudiar, aunque el conjunto de datos concreto a utilizar (que puede diferir del original en algún caso) será el disponible en la web de la asignatura:

- Heart Failure Prediction: `https://www.kaggle.com/fedesoriano/heart-failure-prediction`

- Mobile Price Classification: `https://www.kaggle.com/iabhishekofficial/mobile-price-classification`
- Bank Marketing: `https://archive.ics.uci.edu/ml/datasets/bank+marketing`
- Tanzania Water Pump: `https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table`

3. Tareas a Realizar

La práctica consiste principalmente en que el alumno estudie el comportamiento de distintos algoritmos de clasificación mediante el diseño experimental apropiado y el análisis comparado de resultados. Además, también deberá extraer conclusiones a partir del conocimiento aprendido mediante estos algoritmos para comprender las relaciones entre las variables (también llamadas *características*, *atributos* o *predictores*) que favorecen una determinada clase. El trabajo se realizará sobre la plataforma KNIME (`http://www.knime.com`), incluyendo cualquiera de sus extensiones disponibles (en especial, la de Weka que ofrece una gran variedad de algoritmos adicionales).

Concretamente, se deberán resolver adecuadamente las siguientes tareas:

1. Se considerarán al menos cinco algoritmos de clasificación distintos. Se valorará la selección justificada de estos algoritmos en función de las características del conjunto de datos así como la elección de variedad de tipos de representación (árboles, reglas, redes neuronales, etc.).

2. Toda la experimentación se realizará con validación cruzada de 5 particiones. Para sustentar el análisis comparativo se emplearán tablas de errores, matrices de confusión y curvas ROC. Además de la precisión, se añadirán, al menos, las medidas de rendimiento TPR, TNR, Valor-F_1, G-mean y AUC así como medidas de complejidad del modelo (número de hojas, número de reglas, número de nodos, etc.).

3. Todos los análisis de resultados serán comparativos, de forma que se estudien los pros y contras de cada representación y/o de cada algoritmo. La documentación deberá incluir al menos una tabla resumen que contenga los resultados medios de todos los algoritmos analizados. El análisis no podrá reducirse a una simple lectura de los resultados obtenidos. El alumno deberá formular y argumentar hipótesis sobre las razones de cada resultado. En este problema, ¿por qué el algoritmo X funciona mejor que el Y? ¿Por qué la representación X presenta ciertas ventajas respecto a la Y?

4. Se probarán configuraciones alternativas de los parámetros de dos de los algoritmos empleados justificando los resultados obtenidos. Por ejemplo, ¿puedo evitar o paliar el sobreaprendizaje ajustando los parámetros? ¿Puedo obtener modelos más fácilmente interpretables sin sacrificar excesiva precisión? Para realizar este análisis, se incluirán tablas

comparativas con los resultados del algoritmo con parámetros o configuración por defecto y con las distintas alternativas estudiadas, procurando suficiente variedad de valores de los parámetros para un análisis correcto.

5. Se deberán analizar los datos con diferentes gráficas para comprender su naturaleza e influencia en el proceso de clasificación.

6. A la luz de este análisis, se deberá estudiar un procesado básico de los datos que mejore la predicción (por ejemplo, eliminar alguna característica por razón justificada, agrupar los valores posibles de una característica, eliminar ciertas instancias del conjunto de entrenamiento que se consideren erróneas, convertir una característica categórica en varias binarias, imputar valores perdidos, equilibrar el balanceo de clases, descomponer problemas multiclase...). Según las particularidades de cada problema, serán de utilidad unas u otras estrategias. Deberán justificarse las acciones tomadas y analizar porqué determinado procesado funciona mejor en un determinado tipo de algoritmo. Si no se consigue mejorar la predicción, se podrá al menos describir los procesados que se han probado y los resultados obtenidos. De nuevo, se requiere una tabla resumen que muestre los resultados antes y después de los diferentes procesados de datos.

7. Basado en todo lo anterior, se deberán extraer conclusiones en cada problema sobre los factores que determinan cada clase. Para llegar a estas conclusiones, se pueden analizar los modelos legibles generados (por ejemplo, árboles de decisión, conjuntos de reglas o regresiones lineales), analizar la importancia de cada característica en el proceso de clasificación (especialmente útil en *ensemble learning*, por ejemplo) y visualizar los resultados de predicción de los modelos sobre diferentes casos de entrada (*What-If Analysis*).

4. Esquema de la Documentación

La documentación entregada deberá ajustarse al siguiente esquema (debe respetarse la numeración y nombre de las secciones):

1. **Introducción**: se hablará sobre los problemas abordados, las particularidades de cada caso (dimensiones, tipos de variables, desbalanceo de clase, existencia de valores perdidos, etc.) y todas las consideraciones generales que se deseen indicar.

2. **Resultados obtenidos**: incluirá un apartado 2.x por cada algoritmo estudiado. En cada apartado se añadirán capturas de pantalla de KNIME que expliquen el flujo de trabajo empleado y una única tabla con los resultados obtenidos por el algoritmo en todos los problemas como se describe en la tarea 2.

3. **Análisis de resultados**: incluirá la tabla resumen de todos los algoritmos analizados así como su interpretación y análisis mencionados en la tarea 3. Se podrán añadir gráficas y visualizaciones (por ejemplo, boxplots) que apoyen el análisis.

4. **Configuración de algoritmos**: se incluirá un apartado para cada algoritmo cuya configuración y parámetros hayan sido estudiados. En cada apartado, se incluirá una tabla con los resultados y se realizará el correspondiente análisis como se describe en la tarea 4.

5. **Procesado de datos**: se describirá el procesado realizado, la tabla de resultados y su análisis como se describe en la tarea 6. Se incluirán capturas de pantalla de KNIME con los flujos de trabajo usados para los distintos procesados.

6. **Interpretación de resultados**: como se describe en la tarea 7. Se incluirán las representaciones de modelos y visualizaciones de casos necesarias para sustentar la interpretación de resultados en cada problema.

7. **Contenido adicional**: cualquier tarea adicional a las descritas en este guion puede presentarse en esta sección.

8. **Bibliografía**: referencias y material consultado para la realización de la práctica.

Las tablas de resultados no deberán ser capturas de pantalla, sino tablas creadas en el procesador de texto empleado. No se aceptarán otras secciones distintas de estas. Además, la primera página de la documentación incluirá una portada con el nombre completo del alumno, grupo de prácticas y dirección email. También se incluirá una segunda página con el índice del documento donde las diferentes secciones y páginas estarán enlazadas en el `pdf`.

5. Entrega

La fecha límite de entrega es el miércoles **3 de noviembre** de 2021 hasta las **23:59**. La entrega se realizará a través de la web de la asignatura en `https://prado.ugr.es/`. En ningún caso se aceptan entregas a través de enlaces como Dropbox, Google Drive, WeTransfer o similares.

Se entregarán, al menos, los siguientes ficheros:

1. Documentación: En formato `pdf`. El nombre del archivo se compondrá con `P1` y los apellidos y nombre del alumno sin espacios: `P1-apellido1-apellido2-nombre.pdf`. Es decir, la alumna "María Teresa del Castillo Gómez" subirá el archivo `P1-delCastillo-Gómez-MaríaTeresa.pdf`.

2. Proyecto(s) de KNIME: Uno o varios proyectos con extensión `knwf` resultado de exportarlo desde KNIME. Para evitar un excesivo tamaño de archivo, el proyecto KNIME se puede entregar sin ejecutar (marcando la casilla "Reset Workflow(s) before export" al exportarlo) para que no se almacenen todos los resultados intermedios y por tanto se reduzca drásticamente su tamaño.