UNIVERSIDAD DE GRANADA

UNIVERSIDAD DE GRANADA

Departamento de Ciencias de la Computación e Inteligencia Artificial

Inteligencia de Negocio

Guión de Prácticas

Práctica 1: Análisis Predictivo Mediante Clasificación.

Curso 2024-2025

Práctica 1

Análisis Predictivo Mediante Clasificación

1.1. 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 se adquirirán capacidades para abordar problemas reales en donde la minería de datos puede aportar valor en forma de conocimiento para ayudar en la toma de decisiones. Se trabajará con un conjunto de datos real sobre el 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 2.5 puntos**. Se valorará 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.

1.2. Problemas Abordados

En esta práctica trabajaremos con tres 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, 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 será el disponible en la web de la asignatura, ya que puede diferir del original**:

- **Predicción de aprobación de créditos**: Este conjunto de datos recopila información para predecir los resultados de la aprobación de préstamos en función de los detalles individuales del solicitante, distintas métricas financieras y los factores específicos del préstamo. Contiene un amplio conjunto de 32.581 instancias, que ofrecen diversas características que influyen en las decisiones de aprobación de préstamos.. Puede consultar https://www.kaggle.com/datasets/itshappy/ps4e9-original-data-loan-approval-prediction/data para más detalles.
- **Predicción de una segunda cita**: Este conjunto de datos recopila información obtenida de participantes en eventos experimentales de citas rápidas. Durante los eventos,

los asistentes tenían una *primera cita* de cuatro minutos con cada otro participante del sexo opuesto. Al final de los cuatro minutos, se preguntaba a los participantes si les gustaría volver a ver a su cita. También se les pidió que puntuaran a su cita en seis atributos: atractivo, sinceridad, inteligencia, diversión, ambición e intereses compartidos. El conjunto de datos también incluye datos de cuestionarios recogidos de los participantes en distintos momentos del proceso. Estos campos incluyen: datos demográficos, hábitos de citas, autopercepción de atributos clave, creencias sobre lo que los demás consideran valioso en una pareja e información sobre el estilo de vida. Se puede consultar https://www.openml.org/search?type=data&sort=runs&id=40536&status=active para obtener el listado detallado de atributos y más información.

- **Predicción del tipo de enfermedad eritemato-escamosa**: El diagnóstico diferencial de las enfermedades eritematoescamosas es un verdadero problema en dermatología. Todas comparten las características clínicas del eritema y la descamación, con muy pocas diferencias. Las enfermedades de este grupo son la psoriasis, la dermatitis seborreica, el liquen plano, la pitiriasis rosada, la dermatitis crónica y la pitiriasis rubra pilaris. Normalmente es necesaria una biopsia para el diagnóstico, pero por desgracia estas enfermedades también comparten muchas características histopatológicas. Otra dificultad para el diagnóstico diferencial es que una enfermedad puede mostrar los rasgos de otra enfermedad en la fase inicial y presentar los rasgos característicos en las fases siguientes. En primer lugar se evaluó a los pacientes clínicamente con 12 rasgos. Posteriormente, se tomaron muestras de piel para la evaluación de 22 características histopatológicas. Los valores de las características histopatológicas se determinan mediante un análisis de las muestras al microscopio. Puede consultar https://archive.ics.uci.edu/dataset/33/dermatology para más detalles.

Los datos se encuentran en el fichero **datasets_practica1.zip**, disponible en PRADO.

1.3. Tareas a Realizar

La práctica consiste principalmente en que cada estudiante 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* o *predictores*) que favorecen una determinada clase. El trabajo se realizará sobre la plataforma KNIME (https://www.knime.com), incluyendo cualquiera de sus extensiones disponibles (en especial, la de Weka que ofreced una 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, Naïve-Bayes, k-NN, etc.). En los casos en los que el algoritmo requiera un preprocesamiento, se deberá de añadir antes de aplicar el algoritmo.

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

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

4. 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 incluya los resultados medios de todos los algoritmos analizados. El análisis no podrá reducirse a una simple lectura de los resultados obtenidos. Se 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? ¿Por qué se ha aplicado ese preprocesamiento?

5. Para uno de los problemas se probarán configuraciones alternativas de los parámetros de los algoritmos empleados justificando los resultados obtenidos. Por ejemplo, ¿puedo evitar o paliar el sobre-aprendizaje 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 variaciones estudiadas. Si el análisis es suficientemente completo, no es necesario estudiar todos los algoritmos analizados, se pueden escoger solo algunos de ellos.

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, ...). Deberán justificarse las acciones tomadas y analizar por qué 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 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)

así como visualizar los resultados de predicción de los modelos sobre diferentes casos de entrada (*What-If Analysis*).

1.4. Esquema de la Documentación

La documentación entregada deberá ajustarse al siguiente esquema, para cada problema (debe respetarse la numeración y nombre de las secciones).

1. **Introducción**: se hablará sobre el problema abordado, 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. **Procesado de datos**: Cuando corresponda se describirá el tipo de procesado (ej: balanceo, forma de abordar los valores nulos) que se vayan a estudiar. El procesamiento puede ser común a todos los algoritmos, o específico para determinados algoritmos. Se incluirán capturas de pantalla de KNIME con los flujos de trabajo usados para los distintos procesamientos.

3. **Resultados obtenidos**: incluirá un apartado x por cada algoritmo estudiado. En cada apartado se añadirán capturas de pantalla de KNIME que expliquen el flujo de trabajo empleado, y una tabla con las métricas para dicho algoritmo.

4. **Configuración de algoritmos**: Cuando proceda se incluirá un apartado mostrando para cada algoritmo los parámetros estudiados, incluyendo una tabla con los resultados, y se realizará el análisis como se describe en la tarea 5.

5. **Análisis de resultados**: incluirá la tabla resumen de todos los algoritmos analizados así como su interpretación y análisis mencionados en el punto 4 del apartado anterior. Las tablas resumen del análisis no serán capturas de pantalla, sino tablas creadas en el procesador de texto empleado. Se podrán añadir gráficas y visualizaciones (por ejemplo, boxplots), que apoyen el análisis como sugiere la tarea 3.

6. **Interpretación de los datos**: como se describe en la tarea 7, se identificarán los atributos que identifican mejor cada clase. Además, se analizarán los modelos interpretables (ej: visualizándolos) para sustentar la interpretación de resultados.

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

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

No se aceptarán otras secciones distintas de éstas. Además, la primera página de la documentación incluirá una portada con el nombre completo del alumno/de la alumna, 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`.

1.5. Entrega

La fecha límite de entrega será el miércoles **10 de noviembre** de 2024 hasta las **23:59**. La entrega se realizará a través de la web de la asignatura en PRADO.En ningún caso se aceptarán entregas a través de enlaces como Dropbox, Google Drive, WeTransfer o similares.

Se entregarán, al menos, los siguientes ficheros:

Documentación En formato **pdf**. El nombre del archive se compondrá con **P1** y los apellidos y nombre sin espacios: **P1-apellido1-apellido2-nombre.pdf**, sin acentos. Es decir, la alumna "María Teresa del Castillo Gómez" subirá el archivo **P1-delCastillo-Gomez-MariaTeresa.pdf**.

Proyecto(s) de KNIME Uno o varios proyectos con extensión knwf resultado de exportarlo desde KNIME. Se subirá ya ejecutado.