Trabajo 1: Preprocesamiento y evaluación de clasificadores

Máster Interuniversitario en Big Data: Tecnologías de Análisis de Datos Masivos

Minería de datos

Curso 2023/2024

El alumno deberá resolver las cuestiones que se plantean. Una vez resueltas, deberá crear una memoria en formato Jupyter Notebook con los comentarios en formato Markdown con el código incrustado, la salida en formato HTML, en la que quede reflejado el proceso de resolución seguido. Además de la corrección de las soluciones propuestas, se valorará la presentación de la memoria y la justificación a las decisiones tomadas.

Los archivos a entregar se comprimirán en un fichero .ZIP con nombre ApellidosNombre.zip. El nombre de los alumnos que componen el grupo deben aparecer claramente identificados al principio de la memoria.

Predecir los accidentes cerebrovasculares

Según la Organización Mundial de la Salud (OMS), el eventos cerebrovascular son la segunda causa principal de muerte a nivel mundial, responsables de aproximadamente el 11% del total de muertes. El conjunto de datos que se proporciona en esta práctica se utilizará para predecir si un paciente probablemente sufrirá un accidente cerebrovascular basado en parámetros de entrada como el género, la edad, diversas enfermedades y el estado de fumador.

Sobre el problema anterior, realizar los siguientes puntos:

- Realizar un análisis descriptivo de los datos y aplicación de las técnicas necesarias para curarlos y limpiarlos.
- Realizar un proceso de selección de variables y la generación de los conjuntos de datos con las variables seleccionadas.
- Entrenar 4 clasificadores (Random forest, regresión logística, SVC y KNN).
- En el entrenamiento de cada clasificador se debe realizar una búsqueda para determinar qué combinación de hiperparámetros presenta la mejor eficacia.
- Evaluación y comparación de los clasificadores para determinar cuál es el mejor.

Además, el documento se estructurará de la siguiente forma:

- Librerías utilizadas
- Carga de datos
    - Información de los datos
    - Información de cada atributo
- Preprocesado
    - Gestión de datos ausentes
    - Transformación de datos (si es necesario)
    - Normalización de datos (si es necesario)
    - Selección de variables
- Modelado
    - Random Forest
    - Regresión logística

    - SVC
    - KNN
- Discusión de resultados
- Conclusiones

Normas generales:

- Se debe garantizar la reproducibilidad de los resultados.
- Se debe aportar toda la evidencia gráfica posible junto con los comentarios pertinentes.
- Se debe comentar el *proceso completo, las conclusiones y todas las decisiones tomadas.*