Módulo: Machine Learning

Autores: Inmaculada Gutiérrez, Daniel Gómez, Juan Antonio Guevara

MACHINE LEARNING

Trabajo de Clasificación binaria

Este trabajo está orientado a predecir una variable **binaria** a través de diferentes algoritmos de clasificación.

Normas para la realización del trabajo

a) El trabajo se hará de forma individual.
b) La entrega del trabajo consta de dos partes.
    i) Memoria del trabajo realizado, entregada en **pdf** a través del Campus Virtual. La tarea se evaluará exclusivamente con el contenido de la memoria, que debe ser **autoexplicativa**, incluyendo fragmentos del código empleado si fuera necesario. **No se admitirán memorias enviadas en formato *.ipynb*** o descargadas de Jupyter (con errores, log y similar)**.**
    ii) Código en Python, para poder reproducir los resultados, pero no será evaluado).
c) El trabajo deberá estar **explicado** (no basta con responder a las cuestiones), indicando, si es necesario, el **código utilizado** en el pdf. El trabajo explicado debe ser reproducible, no basta con exponer los resultados. Se valora la claridad de exposición en el informe y la estructura. Puede contener Anexos de datos y gráficos o no, todo según vuestro criterio. Se establece un límite de **20 páginas**.
d) Se debe usar la matriz de datos proporcionada: *datos_tarea25.csv*
e) Es importante tener en cuenta que cada técnica exige una depuración/preparación de los datos específicos. No es necesario desarrollar este proceso en cada una de las preguntas,

pero es **imprescindible explicar para cada técnica, qué parte de la depuración/preparación es necesaria**.

f) La base de datos está relacionada con la información del siguiente repositorio *https://www.kaggle.com/datasets/deepcontractor/car-price-prediction-challenge*.

g) Hay que construir la variable objetivo a partir de *Color*, tratando de predecir si un coche debe pintarse de color blanco o no.

Enunciado

Una empresa dedicada a la venta de coches usados se enfrenta al desafío de determinar el color óptimo para repintar vehículos que llegan en condiciones deficientes. Tras evaluar las opciones, decide limitarse a los colores blanco y negro, por ser los más comunes en el mercado. Para decidir el color de repintado de cada coche, la empresa planea desarrollar un modelo predictivo que, basándose en las características de los vehículos en el mercado de segunda mano, determine si originalmente eran blancos o negros. La base de datos disponible incluye las siguientes variables independientes

Precio de venta, Cantidad de Impuestos a pagar, Fabricante, Año de fabricación, Categoría, Interior de cuero, Tipo de combustible, Volumen del motor, Kilometraje, Cilindros, Tipo de caja de cambios, Ruedas motrices, Lugar del volante, Número de Airbags

Y de la variable dependiente *Color*. La decisión final es si el coche debe pintarse de ***blanco*** o no.

Para llegar a esta base de datos se ha utilizado la base de datos de Kaggle "*Car Price Prediction Challenge*" (*https://www.kaggle.com/datasets/deepcontractor/car-price-prediction-challenge*) con las siguientes modificaciones

1) Se han eliminado las filas color diferente al *Blanco* o *Negro*
2) Se han eliminado las variables con muchas categorías (Identificador, Modelo y Número de puertas)
3) Se han eliminado las categorías con menos de 625 observaciones en alguna variable (*Fabricante, Categoría, Tipo de combustible, Tipo de caja de cambios* y *Ruedas motrices*).

Cuestiones a responder

1) Analiza y depura la base de datos proporcionada, justificando todos los procesos de *feature engineering* y su necesidad en posteriores procesos de predicción.

2) Obtener el mejor modelo de regresión logística. A partir de las variables input involucradas en la mejor regresión logística, encontrar la mejor red neuronal en términos de Accuracy. Justificar las parametrizaciones probadas, el *threshold* establecido en función de la evolución de las métricas, y los parámetros escogidos.

3) Aplicar un árbol decisión, y selecciona las variables más importantes. Usando estas variables como input, encontrar el mejor modelo de red neuronal en términos de Recall. Justificar las parametrizaciones probadas, el *threshold* establecido en función de la evolución de las métricas, y los parámetros escogidos.

4) Desarrolla un proceso comparativo completo de los modelos y procesos seleccionados en los apartados 2) y 3), justificando y argumentando las decisiones y resultados.