Universidad del Valle
Facultad de Ingeniería
Escuela de Ingeniería de Sistemas y Computación
Inteligencia artificial
Informe sobre *Machine learning*

Para la experimentación con técnicas de *machine learning* se usará un conjunto de datos que tiene información de 1067 vehículos. Dicha información se utiliza para obtener modelos que permitan predecir las emisiones de dióxido de carbono (CO_2) dadas ciertas características. Cada vehículo se describe utilizando los cinco atributos que se muestran en la tabla. Las variables independientes son ENGINESIZE, CYLINDERS, FUELCONSUMPTION_CITY, y FUELCONSUMPTION_HWY. La variable dependiente es el atributo *CO2EMISSIONS* cuyos valores pueden ser 0 ó 1, donde 0 significa que las emisiones de dióxido de carbono son Bajas y 1 indica que son Altas. En este taller se obtendrán modelos que intentan predecir la variable dependiente *CO2EMISSIONS*.

#	Atributo	Descripción
1	ENGINESIZE	Tamaño del motor en litros
2	CYLINDERS	Cantidad de cilindros que posee el motor
3	FUELCONSUMPTION_CITY	Consumo de combustible del vehículo en zona urbana (L/100 km)
4	FUELCONSUMPTION_HWY	Consumo de combustible del vehículo en zona extraurbana (L/100 km)
5	*CO2EMISSIONS*	Emisiones de CO_2 del vehículo, donde 0 significa que las emisiones son Bajas y 1 significa que son Altas

En la siguiente tabla se muestra una de las instancias del conjunto de datos. En este caso es un vehículo con un motor de 3.7 litros (atributo 1), 6 cilindros (atributo 2), un consumo de combustible en zona urbana de 13.4 (atributo 3), y un consumo de combustible en zona extraurbana de 9.5 (atributo 4). Para este vehículo se conoce que tiene una emisión Alta de dióxido de carbono (atributo 5).

Atributo	1	2	3	4	*5*
Valor	3.7	6	13.4	9.5	1

El objetivo de este informe es crear dos notebooks. Uno donde se utilice la técnica de redes neuronales y otro para la técnica de árboles de decisión. Inicialmente se deben probar diferentes topologías de redes neuronales y modificar los hiperparámetros de tal manera que se puedan obtener modelos que permitan predecir si las emisiones de dióxido de carbono son Bajas o Altas. Para esto, debe entregar un notebook donde se realicen las siguientes tareas:

1. Leer el archivo *CO2 emissions.csv*.
2. Seleccionar aleatoriamente el 80% del conjunto de datos para entrenar y el 20% restante para las pruebas
3. Utilizar una estrategia para normalizar los datos.
4. Construir 5 redes neuronales variando la función de activación, el solver, y la cantidad de capas ocultas y de neuronas por cada capa oculta. Como funciones de activación puede seleccionar ‘identity’, ‘logistic’, ‘tanh’, o ‘relu’. Por su parte, para los solvers puede seleccionar ‘lbfgs’, ‘sgd’, o ‘adam’. En todas las pruebas

debe usar un random_state=123. Incluya en el notebook una tabla a manera de resumen con el *accuracy* obtenido en cada caso. Además, debe mostrar las cinco matrices de confusión.

5. Indique en el notebook usando una celda de tipo texto los hiperparámetros que por el momento le permiten obtener la red con mayor *accuracy*.
6. Seleccione uno de los hiperparámetros disponibles en la documentación (https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html) que sea diferente al solver, a la función de activación, y al random_state. Realice dos variaciones en el hiperparámetro seleccionado manteniendo los otros hiperparámetros del punto anterior. Indique el *accuracy* obtenido al modificar el hiperparámetro seleccionado y analice si la red mejora, empeora, o mantiene su exactitud. Incluya en el notebook dicho análisis.

En el segundo notebook se deben realizar las siguientes tareas:

1. Leer el archivo *CO2 emissions.csv*.
2. Seleccionar aleatoriamente el 80% del conjunto de datos para entrenar y el 20% restante para las pruebas
3. Utilizar una estrategia para normalizar los datos.
4. Configurar los hiperparámetros del árbol de decisión de la siguiente manera: criterion='gini', splitter='best', y random_state=123. Obtener 10 árboles de decisión que resultan de modificar el hiperparámetro max_depth desde 1 hasta 10 de 1 en 1.
5. Incluya en el notebook una tabla con el *accuracy* para los 10 árboles del punto anterior.
6. Repita el mismo procedimiento del punto 4 usando como hiperparámetros criterion='entropy', splitter='best', random_state=123, y variando el hiperparámetro max_depth desde 1 hasta 10 de 1 en 1.
7. Incluya en el notebook una tabla con el *accuracy* para los 10 árboles del punto anterior.
8. Repita el mismo procedimiento del punto 4 usando como hiperparámetros criterion='entropy', splitter='random', random_state=123, y variando el hiperparámetro max_depth desde 1 hasta 10 de 1 en 1.
9. Incluya en el notebook una tabla con el *accuracy* para los 10 árboles del punto anterior.
10. Indique en el notebook los hiperparámetros que por el momento le permiten obtener el árbol con mayor *accuracy*.
11. Seleccione uno de los hiperparámetros disponibles en la documentación (https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html) que sea diferente al criterion, splitter, max_depth, y random_state. Realice dos variaciones en el hiperparámetro seleccionado manteniendo los otros hiperparámetros del punto anterior. Indique el *accuracy* obtenido al modificar el hiperparámetro seleccionado y analice si el árbol de decisión mejora, empeora, o mantiene su exactitud.