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

Para la experimentación con técnicas de machine learning se empleará un conjunto de datos compuesto por 1044 personas que participaron en un estudio orientado a predecir si aprueban un curso de matemáticas. La información recopilada incluye variables como edad, sexo, ocupación de la madre y del padre, tiempo de estudio semanal, acceso a internet y otros datos relevantes. Cada estudiante está descrito mediante los 17 atributos presentados en la siguiente tabla. La variable dependiente es *approved*, que toma el valor 1 cuando el estudiante aprueba el curso y 0 en caso contrario. En este informe se desarrollan modelos para predecir dicha variable a partir de los demás atributos, abordando el problema como una tarea de clasificación.

#	Atributo	Descripción	Tipo de variable	Posibles valores
1	sex	Sexo de la persona	Categórica	F M
2	age	Edad de la persona	Numérica	15 - 22
3	famsize	Tamaño de la familia	Categórica	LE3 si la familia tiene 3 o menos personas GT3 si la familia tiene más de 3 personas
4	Pstatus	Estado de convivencia de los padres	Categórica	T (together) si los padres viven juntos A (apart) si los padres están separados
5	Medu	Educación de la madre	Numérica	0 – ninguna 1 - educación primaria 2 - de 5º a 9º grado 3 - educación secundaria 4 - educación superior
6	Fedu	Educación del padre	Numérica	0 – ninguna 1 - educación primaria 2 - de 5º a 9º grado 3 - educación secundaria 4 - educación superior
7	Mjob	Trabajo de la madre	Categórica	teacher – docente health - relacionado con salud services - servicios civiles, por ejemplo, administrativo o policía at_home - dedicada al hogar other - otro
8	Fjob	Trabajo del padre	Categórica	teacher – docente

				health - relacionado con salud services - servicios civiles at_home - dedicado al hogar other - otro
9	traveltime	Tiempo de traslado de la casa a la institución educativa	Numérica	1 – menos de 15 minutos 2 – de 15 a 30 minutos 3 – de 30 minutos a 1 hora 4 – más de 1 hora
10	studytime	Tiempo de estudio semanal	Numérica	1 - menos de 2 horas 2 - de 2 a 5 horas 3 - de 5 a 10 horas 4 - más de 10 horas
11	failures	Cantidad de veces que ha perdido el mismo curso	Numérica	1 2 3 4, si ha perdido el curso 4 o más veces
12	internet	Indica si tiene acceso a internet en la casa	Categórica	yes no
13	romantic	Indica si se encuentra en una relación romántica	Categórica	yes no
14	goout	Indica si sale con amigos	Numérica	Valor entero de 1 a 5, donde 1 representa 'muy bajo' y 5 representa 'muy alto'
15	Walc	Consumo de alcohol los fines de semana	Numérica	Valor entero de 1 a 5, donde 1 representa 'muy bajo' y 5 representa 'muy alto'
16	health	Estado de salud actual	Numérica	Valor entero de 1 a 5, donde 1 representa 'muy malo' y 5 representa 'muy bueno'
17	approved	Indica si el estudiante ganó el curso, o no	Variable dependiente a predecir	0 = No aprueba el curso 1 = Sí aprueba el curso

El objetivo de este informe es crear dos notebooks. En el primer notebook se usará la técnica de redes neuronales probando diferentes topologías y modificando los hiperparámetros. Para esto, debe entregar un notebook donde se realicen las siguientes tareas:

1. Leer el archivo *student_performance.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 numéricos y una forma de codificar los atributos categóricos.
4. Construir 5 redes neuronales variando la cantidad de capas ocultas y de neuronas por cada capa oculta, el solver, y la función de activación.
5. Incluya en el notebook una tabla con el *accuracy* obtenido sobre el conjunto de prueba para las 5 redes neuronales del punto anterior.
6. Indique en el notebook, usando una celda de tipo texto, los hiperparámetros que por el momento le permiten obtener la red con mayor *accuracy.*
7. 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 y a la función de activación. 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 *student_performance.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 numéricos y una forma de codificar los atributos categóricos.
4. Obtener 5 árboles de decisión que resultan de modificar el hiperparámetro *max_depth* desde 2 hasta 10 con incrementos de 2, usando el criterio *gini.*
5. Incluya en el notebook una tabla con el *accuracy* obtenido sobre el conjunto de prueba para los 5 árboles del punto anterior.
6. Obtener 5 árboles de decisión que resultan de modificar el hiperparámetro *max_depth* desde 2 hasta 10 con incrementos de 2, usando el criterio *entropy.*
7. Incluya en el notebook una tabla con el *accuracy* obtenido sobre el conjunto de prueba para los 5 árboles del punto anterior.
8. Indique en el notebook los hiperparámetros (*max_depth y criterion*) que por el momento le permiten obtener el árbol con mayor *accuracy.*
9. Seleccione uno de los hiperparámetros disponibles en la documentación (https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html) que sea diferente a *criterion* y a *max_depth*. 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.