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 de 700 personas, algunas de las cuales sufrieron derrame cerebral. Cada persona se describe utilizando los 11 atributos que se presentan en la tabla. Estas variables incluyen el sexo, la edad, si el paciente sufre de hipertensión o enfermedades cardíacas, entre otra información. La etiqueta de clase o variable dependiente es el atributo *stroke* cuyos valores pueden ser 1 ó 0 indicando si la persona sufrió un derrame cerebral, o no. En este taller se obtendrán modelos que intentan predecir la variable dependiente *stroke*. En la siguiente tabla se listan los atributos y se presenta su descripción.

#	Atributo	Descripción	Posibles valores
1	gender	Sexo del paciente	Male Female
2	age	Edad del paciente	Valores reales positivos
3	hypertension	Indica si el paciente sufre de hipertensión	0 – No 1 – Sí
4	heart_disease	Indica si el paciente sufre de enfermedades cardíacas	0 – No 1 – Sí
5	ever_married	Indica si el paciente está o ha estado casado	No Yes
6	work_type	Tipo de trabajo del paciente	children Govt_job Never_worked Private Self-employed
7	Residence_type	Tipo de residencia del paciente	Rural Urban
8	avg_glucose_level	Nivel de glucosa promedio en la sangre	Valores reales positivos
9	bmi	Índice de masa corporal	Valores reales positivos
10	smoking_status	Nivel en el que el paciente fuma	formerly smoked never smoked smokes Unknown
11	stroke	Indica si el paciente sufrió de un derrame cerebral, o no	0 – No 1 – Sí

En la siguiente tabla se muestra como ejemplo la información de un paciente. Se trata de un hombre (atributo 1), de 80 años (atributo 2), que no sufre de hipertensión (atributo 3), pero sí de enfermedades cardíacas (atributo 4), ha estado casado (atributo 5), trabajó en el sector privado (atributo 6), vive en una residencia de tipo rural (atributo 7), su nivel de glucosa es de 105.92 (atributo 8), su índice de

masa corporal es 32.5 (atributo 9), y nunca ha fumado (atributo 10). Esta persona tuvo un derrame cerebral (atributo 11).

Atributo	1	2	3	4	5	6	7	8	9	10	11
Valor	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	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 una persona sufrirá de un derrame cerebral, o no. Para esto, debe entregar un notebook donde se realicen las siguientes tareas:

1. Leer el archivo stroke.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 y llenar los datos faltantes
4. Construir 5 redes neuronales variando en la topología de la red la cantidad de capas ocultas y de neuronas por cada capa oculta. Puede también variar los hiperparámetros solver y la función de activación. 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 y también las matrices de confusión
5. Indique en el notebook usando una celda de tipo *Markdown* 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 stroke.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 y llenar los datos faltantes
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 5 hasta 50 con incrementos de 5
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 5 hasta 50 con incrementos de 5
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 5 hasta 50 con incrementos de 5
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.