31/10/2023

UNIDAD 3 – ACTIVIDAD 2: ÁRBOLES DE DECISIÓN Y VALIDACIÓN.

Grupo B

Descripción breve

Se busca que el estudiante pueda utilizar un modelo de ML e identifique su rendimiento.

M.I.A. Wilfrido Cortes Orozco
wilfrido.co@morelia.tecnm.mx

Objetivos de la actividad.

Esta actividad busca cumplir los siguientes objetivos:

- Que el estudiante pueda crear un modelo de árbol de decisión.
- Que el estudiante pueda distinguir entre problemas de clasificación y regresión.
- Que el estudiante comprenda las métricas para cada tipo de problema y sepa identificar el comportamiento del modelo.

Descripción de la actividad.

El set de datos para esta actividad se llama “diabetes.csv”, es un conjunto de datos con información acerca de pacientes que tienen (o no) diabetes. Este dataset ya se ha utilizado previamente.

Hay que realizar un análisis exploratorio de los datos breve:

1. Cargar el set de datos en un dataframe de pandas.
2. Mostrar información de las columnas (tipos), hay que mencionar la cantidad de columnas y filas.
3. Mostrar información estadística de las columnas.
4. Para esta actividad se usarán 2 columnas como objetivo: “Outcome” y “Age”. Hay que mostrar sus distribuciones (se puede hacer con una gráfica de distribución o con un boxplot).
5. Identificar la correlación de las columnas con respecto a “Outcome”, ¿qué columnas son las que presentan una correlación más fuerte?, ¿se puede usar un modelo de regresión lineal?
6. Identificar la correlación de las columnas con respecto a “Age”, ¿qué columnas son las que presentan una correlación más fuerte?, ¿se puede usar un modelo de regresión lineal?
7. Mostrar gráficas de dispersión:
    a. Mostrar la gráfica de dispersión de la columna con mayor correlación con Age vs Age y colorizar en base a su valor de Outcome. (variableX=columnaMayorCorrelación, variableY=Age, hue=Outcome)
    b. Mostrar la gráfica de dispersión de la columna con mayor correlación con Outcome vs Age y colorizar en base a su valor de Outcome. (variableX=columnaMayorCorrelación, variableY=Age, hue=Outcome)
8. Describir la correlación vista en ambas gráficas.
9. ¿Qué tipo de problema (regresión o clasificación) sería la predicción de la columna “Age”?
10. ¿Qué tipo de problema (regresión o clasificación) sería la predicción de la columna “Outcome”?
11. Se debe separar el dataset en muestras para entrenamiento y muestras para pruebas, la proporción tiene que ser 80/20.
12. Se debe importar, instanciar, entrenar y evaluar un árbol de decisión para predecir la etiqueta de “Age”. ***El tipo de árbol y las métricas de evaluación dependen del tipo de problema seleccionado en la pregunta 9. Si se trata de regresión, entonces hay que evaluar con el error cuadrático medio (MSE); si se trata de clasificación, entonces hay que evaluar con accuracy_score. La selección de hiperpárametros queda a decisión del estudiante (criterion, max_depth, min_samples_split, min_samples_leaf, …)***

13. Se debe importar, instanciar, entrenar y evaluar un árbol de decisión para predecir la etiqueta de “Outcome”. ***El tipo de árbol y las métricas de evaluación dependen del tipo de problema seleccionado en la pregunta 10. Si se trata de regresión, entonces hay que evaluar con el error cuadrático medio (MSE); si se trata de clasificación, entonces hay que evaluar con accuracy_score. La selección de hiperpárametros queda a decisión del estudiante (criterion, max_depth, min_samples_split, min_samples_leaf, …)***
14. Si hay un problema de regresión, hay que usar la validación cruzada (vista en clase) para determinar si el modelo está sobreajustado o si es adecuado.