UNIVERSIDAD NACIONAL DE COLOMBIA - SEDE MEDELLÍN

FACULTAD DE MINAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN Y DE LA DECISIÓN

Eq.21

- **Inteligencia Artificial,** Código: **3007855**
- **Introducción a la Inteligencia Artificial,** Código: **3010476**

Semestre**: 01/2024**

Prof. Demetrio Arturo Ovalle Carranza, Ph.D. (e-mail: dovalle@unal.edu.co)

Monitor: Daniel Metaute Medina (**e-mail: dmetaute@unal.edu.co)**

ENUNCIADO DEL MICRO-PROYECTO3 (15%)

APRENDIZAJE SUPERVISADO:
ÁRBOLES DE DECISIÓN – CLASIFICACIÓN BAYESIANA

Fecha enunciado: jueves 15 de agosto 2024

Fecha de entrega: lunes 26 de agosto 2024

Estudiantes: Equipo 21
Hinara Pastora Sánchez Mata
Juan José Tobón
Anna Ospina Bedoya

OBJETIVO –Predecir el riesgo de que una empresa sea fraudulenta (0 si no lo es, 1 si lo es).

DATASET: 2 archivos (Audit_risk_summary.txt, audit_risk.csv)

Se sugiere mezclar los datos de forma aleatoria antes de realizar cualquier otro tipo de preprocesamiento.

Indicaciones de entrega: sólo debe entregarse el archivo descargado de Google Colaboratory con el siguiente formato de nombre: "Equipo#_Microproyecto3" *(ejemplo: Equipo21_Microproyecto21)*. Los trabajos serán probados con el conjunto de datos original, por lo tanto, todas las modificaciones hechas al dataset deben realizarse desde el código. Los integrantes deben aparecer en la libreta (.ipynb) para que se les pueda asignar la calificación obtenida.

Usando algoritmos similares a los presentados en las clases y en las monitorías, realizar programas en Python que utilicen las técnicas supervisadas DT (Árboles de Decisión) y BC (Clasificación Bayesiana) para clasificación de datos.

1) Realice una pequeña descripción de los datos, de las características (features) que los describen y contextualice el problema determinando los valores objetivo (target) que se desean predecir. Explique si tuvo que realizar algún tipo de preprocesamiento en los datos y cómo lo hizo; sea lo más descriptivo posible. Una vez finalizado el pre-procesamiento analizar la distribución de la variable objetivo (use un gráfico de barras).
2) Para la técnica DT justifique la característica escogida para el nodo raíz y la de sus nodos hijo. Pruebe las métricas entropía e índice gini, ¿Cuál es más adecuada para su conjunto de datos? Explique.
3) Explique qué porcentaje de los datos se usaron para entrenamiento y cuántos para prueba. Justifique su respuesta.

4) Explique qué variables usó para la clasificación y por qué, incluya en su análisis el gráfico de correlación de Pearson y el resultado de aplicar SelectKBest.
5) Muestre la exactitud (accuracy) obtenida para diferentes profundidades de árboles y justifique cuál de estos niveles es el más apropiado para su trabajo.
6)Explique: ¿Cuál fue el criterio utilizado para determinar la cantidad mínima de muestras por nodo y el mínimo de muestras en cada hoja? ¿Cuál fue la exactitud utilizando diferentes valores? Adicionalmente, explique si fue necesario balancear las clases (variable objetivo) en el conjunto de entrenamiento (ej. Atributo peso).
Finalmente, muestre el árbol de decisión generado.
7) Presente la matriz de confusión del DT en entrenamiento (train) y en validación (test) con su debida interpretación.
8) Calcule la métrica de exactitud (ver método score en la libreta) tanto para el entrenamiento como para la validación.
9) Calcule las métricas de precisión, sensibilidad (recall) y F1-Score, tanto para el entrenamiento como para la validación.
10) Utilice este mismo dataset para correr la técnica de Clasificación Bayesiana y compare los resultados obtenidos a través del cálculo de la métrica de exactitud obtenida en los modelos.
11) Análisis de resultados y conclusiones.
12) Buscar y entregar 2 datasets, uno para realizar aprendizaje no supervisado (sin etiquetas) y otro con etiquetas para realizar aprendizaje supervisado. Los datasets deben tener al menos 5 variables y 500 registros. Los datasets deben ser diferentes a los que se encuentran en las listas que aparecen en los siguientes links. También, deben incluir, en lo posible, el archivo de datos que describen las variables (atributos).

NOTA: Recuerde explicar qué está haciendo en cada parte del código.