Inteligencia Artificial

1er Cuatrimestre 2024

Trabajo Práctico:

Aprendizaje supervisado

Facultad de Ciencias de la Administración

Fecha límite de entrega: Viernes 17/05/2024 a las 23:59.

Condiciones de entrega: el trabajo práctico deberá ser realizado en forma individual. Se deberá subir en la sección del Campus Virtual correspondiente un único archivo comprimido con formato `zip, rar, tar.gz` u otro. El mismo contendrá los archivos `.py` separados por cada punto/consigna. En cada archivo `.py` se debe presentar el código solución a la consigna y algunas líneas adicionales de código que sirvan para testear la solución presentada. Además, pueden incluir un `.pdf` que presente las respuestas, suposiciones y aclaraciones pertinentes de cada punto.

1. **Librería NumPy.**

   a) Cargar un dataset desde un archivo CSV usando la función loadtxt de NumPy.

   b) Dividir el dataset en un conjunto de entrenamiento y un conjunto de prueba usando la función train_test_split de NumPy.

   c) Normalizar los datos en el conjunto de entrenamiento usando la función StandardScaler de Scikit-learn.

   d) Aplicar la misma normalización al conjunto de prueba usando la función transform del objeto StandardScaler.

2. **Librería Pandas.**

   a) Carga el archivo *car.csv* en un DataFrame y mostrar las primeras 5 filas.

   b) Mostrar la información básica del DataFrame

   c) Mostrar las estadísticas básicas del DataFrame

   d) Seleccionar las columnas del DataFrame y mostrarlas por pantalla.

   e) Filtra las filas del DataFrame que cumplan con la condición persons=2.

   f) Agrupa el DataFrame por una columna específica y muestra la media de otra columna.

   g) Convierte la variable objetivo a una variable binaria usando Label Encoding.

Arboles de decisión.

3. **Dataset car.csv.**
   Se debe crear un modelo de árbol de decisión confiable que sea capaz de ayudar a una empresa a encontrar automóviles que los clientes probablemente comprarán. Se debe construir un modelo de árbol de decisión que clasifique los automóviles como aceptables o no aceptables. El dataset se encuentra disponible en el campus junto a este práctico `car.csv` y se compone de seis características diferentes: compra, mantenimiento, puertas, personas, maletero y seguridad. La variable objetivo clasifica la aceptabilidad de un automóvil determinado. Puede tomar 0 o 1, siendo 1 aceptable.

   (a) Utilizar el 70 % de los datos para entrenamiento y el 30 % restante para testeo.

   (b) Evaluar la precisión del modelo utilizando **Accuracy**.

   (c) A partir de la métrica obtenida: ¿Qué podemos decir del modelo creado?

4. **Dataset wine.**
   Crear un modelo de árbol de decisión a partir del dataset **wine** disponible en la librería sklearn.

   (a) Utilizar el 65 % de los datos para entrenamiento y el 35 % restante para testeo.

   (b) Obtener las métricas **Accuracy**, **Precision**, **Recall** y **F1**.

(c) Dar conclusiones sobre el modelo a partir de las métricas obtenidas.

Regresión Logística.

5. **Dataset Breast Cancer** Utilizar el dataset de cáncer de mama disponible en sklearn para predecir la presencia de cáncer maligno utilizando regresión logística.
   (a) Utilizar el 20 % de los datos para testeo.
   (b) Evaluar el rendimiento del modelo utilizando las métricas **Accuracy**, **Precision** y **Recall**.
   (c) Obtener la matriz de confusión del modelo.

6. Dataset **ClientesEnLinea.csv**
   Crear un modelo de Regresión Logística utilizando el dataset *ClientesEnLinea.csv* que cuenta con información de clientes que compran o no ciertos productos en línea para ello contamos con información sobre el género, la edad y el salario estimado, clasificando a los clientes con 0 y 1 si no compró o si compró respectivamente.
   - Utilizar como métrica comparativa el promedio de una validación cruzada K-fold con 5 folds para entrenamiento y testeo.

   (a) ¿Cómo se comporta el modelo si consideramos todos los predictores?
   (b) ¿Qué sucede cuando solo consideramos como predictores Sexo y Edad?

Regresión lineal

7. Dataset **articulos_ml.csv**
   A partir del dataset **articulos_ml.csv** que se encuentra disponible en el campus y contiene diversas URLs a artículos sobre Machine Learning. Se debe construir un modelo de regresión lineal para predecir cuantas veces será compartido un artículo en redes sociales basándonos en la cantidad de palabras del artículo.
   (a) Mostrar las columnas disponibles en el dataset **articulos_ml.csv**.
   (b) Crear gráficos para visualizar la relación entre las variables del dataset.
   (c) Filtrar los artículos que tengan menos de 3500 palabras y una cantidad de compartidos menor a 80,000 para analizar un conjunto más específico de datos.
   (d) Utilizar los datos filtrados para generar un modelo de regresión lineal y graficar la relación entre las palabras del artículo y la cantidad de veces que son compartidos.
   (e) Utilizar el modelo generado para predecir la cantidad de veces que serán compartidos artículos de 2000, 5000 y 10000 palabras.
   (f) Mostrar los coeficientes del modelo.
   (g) Evaluar el modelo aplicando las métricas **Error Cuadrático Medio** y **Coeficiente de Determinación (R2)**.

Referencias

[1] *https://scikit-learn.org/stable/modules/tree.html*

[2] *https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html*

[3] *https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html*

[4] D. Poole, A. Mackworth., *Artificial Intelligence. Foundations of Computational Agents. 2010. Cambridge University Press.*

[5] S. Russell, P. Norvig., *Artificial Intelligence: A Modern Approach. 3rd Edition. 2010. Prentice Hall.*