Instituto Tecnológico de Costa Rica, ITCR
Programa de Ciencia de Datos
Aprendizaje Automático
Quiz 01

Profesora: María Auxiliadora Mora

Entrega: Subir al TEC-Digital un archivo comprimido que contenga: Cuaderno de Jupyter y los archivos de datos, el original y el resultante.

Modo de trabajo: Individual

Objetivo: En el presente trabajo se repasarán aspectos básicos del preprocesamiento de datos, selección de características y análisis de regresión con el fin de poner en práctica el conocimiento adquirido en clase.

Datos: Para realizar el trabajo se utilizará un conjunto de datos publicado por la Facultad de Ciencias de la Información y la Computación Donald Bren de la Universidad de California en Irvine. El conjunto de datos será utilizado para entrenar modelos **para predecir el costo de vehículos**. Una descripción detallada de los datos está disponible en https://archive.ics.uci.edu/ml/datasets/Automobile.

Ejercicios

Utilice el conjunto de datos y prepare un cuaderno Jupyter para realizar actividades de preprocesamiento de los datos, selección de características y análisis de regresión. Utilizando celdas de texto y de código, realice las siguientes acciones. Para cada requerimiento debe insertarse primero una celda de texto en la que se explica qué se va a realizar y los resultados obtenidos y una celda de código en Python en la cual se realiza el procesamiento. Recuerde que puede usar la biblioteca Scikit-Learn.

1. Cargar los datos y visualizar su contenido, utilice los gráficos y funciones que considere aportan para entender cómo se comportan las variables. (5 puntos)

2. Para los atributos categóricos cuente el número y despliegue la lista de categorías únicas por atributo e indique si el atributo corresponde a un dato nominal u ordinal. (5 puntos)

3. Codifique todos los atributos categóricos, seleccione el método de codificación y explique porqué selecciona cada método. (5 puntos)

4. Realice un análisis de datos faltantes en todo el conjunto de datos y aplique imputación para habilitar esos registros. (8 puntos)

5. Escale los datos: El seleccionar el mecanismo de escalado de datos apropiado depende de los investigadores y del conjunto de datos (por ejemplo, de la distribución de los datos, de si existen valores negativos y positivos, de si existen datos atípicos, entre otros). Varios de estos conceptos son parte del módulo de estadística (parte del Programa de Ciencia de Datos), sin embargo, es importante tener en cuenta que existen varios algoritmos y el impacto de estos en los resultados del proceso de modelización. Estudie algunas de opciones de escalado de datos descritas en [1] y [2]. Con el conocimiento que tenemos actualmente seleccione una de las opciones y explique porqué selecciona el algoritmo de acuerdo al conjunto de datos en uso. El objetivo de este punto del ejercicio es que los estudiantes conozcan los algoritmos y la importancia de estos y que evalúen los datos con los que están trabajando (no se espera que estudien la distribución de los datos). (5 puntos)

6. Aplique los métodos de selección de características Información mutua (3 puntos)

7. Realice un análisis de regresión utilizando Regresión polinómica (3 puntos).

8. Evalúe, compare los resultados de los tres algoritmos de regresión utilizando la métrica R^2 y genere al menos cuatro conclusiones sobre el ejercicio. (5 puntos)

9. Incluya referencias en formato APA (1 punto).

Referencias

El manual de Scikit-learn tiene muy buena explicación de los algoritmos

[1] Scikit-learn (2023). Manual. Sección 6.3. Preprocessing data. Recuperado de https://scikit-learn.org/stable/modules/preprocessing.html

Ejemplos aplicados a un conjunto de datos simple están disponibles en:

[2] Analyticsvidhya (2020). Feature Transformation and Scaling Techniques to Boost Your Model Performance. Recuperado de https://www.analyticsvidhya.com/blog/2020/07/types-of-feature-transformation-and-scaling/