Ciencia de Datos y herramientas Big Data para la Investigación

Trabajo Final del Curso

Prof. Dr. Gualberto Asencio Cortés

Nota. Comprima los entregables de cada ejercicio (un workflow de Orange por cada ejercicio) en un único archivo ZIP y entregue éste en la actividad habilitada de entrega en el aula virtual.

Análisis y Preprocesado de Datos

Ejercicio 1

Cree un workflow en Orange en el que realice los siguientes pasos:

1. Cargue un dataset con clase categórica, el que desee, a partir del nodo "Datasets" de Orange.
2. Analice y visualice la distribución de los datos, utilizando las técnicas estudiadas más apropiadas para el conjunto de datos, tanto univariantes como multivariantes.
3. Discretice una variable que sea numérica creando cinco intervalos de igual anchura.
4. Normalice los valores de todas las variables numéricas al intervalo [0, 1].
5. Realice un ranking de relevancia de atributos con respecto a la clase.

El entregable de este ejercicio será un único archivo de workflow de Orange.

Aprendizaje no supervisado

Ejercicio 2

Cree un workflow en Orange en el que realice los siguientes pasos:

1. Cargue un dataset con clase categórica, el que desee, pero distinto al del ejercicio 1, a partir del nodo "Datasets" de Orange.
2. Analice y visualice la distribución de los datos, utilizando las técnicas estudiadas más apropiadas para el conjunto de datos, tanto univariantes como multivariantes.
3. Realice un agrupamiento (clustering) de las instancias usando el algoritmo k-Means y tomando el número de clusters óptimo según la métrica Silhouette. Muestre en un gráfico las instancias coloreadas según su clúster.

El entregable de este ejercicio será un único archivo de workflow de Orange.

Aprendizaje supervisado

Ejercicio 3

Cree un workflow en Orange en el que realice los siguientes pasos:

1. Cargue un dataset con clase categórica, el que desee, pero distinto a los de los ejercicios 1 y 2, a partir del nodo "Datasets" de Orange.
2. Realice una validación cruzada de 10 bolsas (sin estratificar) con los algoritmos kNN (con k=1 y k=5), Tree y NaiveBayes. ¿Cuál es el algoritmo con el mejor valor de la métrica F1?
3. Divida el conjunto de datos en dos partes: una con el 70% aleatorio de las filas (training) y otra con el resto (test).
4. Entrene el algoritmo que mejor valor de F1 haya obtenido en el paso (2) con la parte de training del paso anterior.
5. Obtenga una tabla de predicciones del modelo entrenado del paso anterior aplicado sobre la parte de test del paso (3).

El entregable de este ejercicio será un único archivo de workflow de Orange.

Ejercicio 4

Cree un workflow en Orange en el que realice los siguientes pasos:

1. Cargue un dataset con clase numérica (regresión), el que desee, a partir del nodo "Datasets" de Orange.
2. Repita los mismos pasos del Ejercicio 3 pero usando los algoritmos Linear Regression, Tree y Neural Network.

El entregable de este ejercicio será un único archivo de workflow de Orange.