Práctica 2

Implementación de k-nearest neighbors.

Guillermo Arcal García

UNIVERSIDAD
DE BURGOS

Universidad de Burgos
Minería de datos

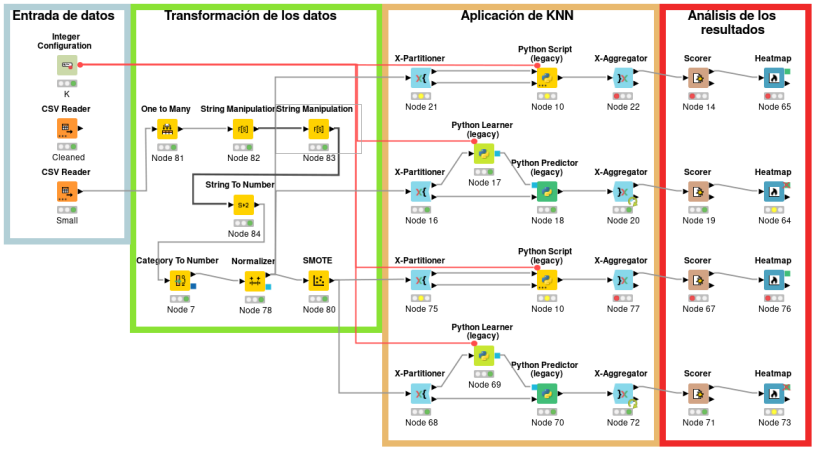
27 de marzo de 2023

Índice

1. Introducción

El objetivo de esta práctica es la implementación del algoritmo K-nearest neighbors

2. Partes del flujo



2.1. Entrada de datos

En la entrada de datos se le el archivo CSV que se desee mediante un bloque “CSV reader”. Además del archivo anterior también se encuentra un bloque ‘Integer configuration” para establecer el valor de K que utilizarán todos los bloques que implementen KNN.

2.2. Transformación de los datos

Ya que el conjunto de datos original contiene clases categóricas es necesario convertir estas a valores numéricos.

Para convertir el atributo “race” que no está ordenado se ha utilizado el bloque “One to Many”, posteriormente para los atributos “GenHealth” y “AgeCategory” se han usado sendos bloques “String Manipulation”, finalmente se han convertido los números en formato texto que dan los bloques anteriores a valores numéricos con un bloque “String to number”, finalmente para el resto de atributos púramente categóricos se ha usado el bloque “String To Number”.

Una vez se tengan valores numéricos es conveniente normalizar los datos para una mejor clasificación de los mismos, esto se consigue mediante el bloque “Normalizer” que normaliza todos los datos numéricos a un valor entre 0 y 1

Ya que el conjunto de datos está desequilibrado y esto no es conveniente se ha considerado analizar los algoritmos también tras equilibrar los datos. Se ha equilibrado el dataset mediante oversampling utilizando el bloque “SMOTE” el cual introduce en los datos copias de los indivíduos de las clases minoritarias.

2.3. Aplicación de KNN

En el workflow se ha hecho de forma simultánea el caso con el dataset normal y su versión equilibrada mediante oversampling pero la forma de aplicar KNN es igual por lo que solo se van a explicar los bloques una vez.

Se han probado dos versiones de KNN:

2.3.1. Implementación propia

Esta versión creada por mí está implementada en el archivo knn.py

2.3.2. Implementación de scikit-learn

Esta versión están implementada en los bloques “Python learner” y “Python predict”

Para ambas versiones de KNN se separan los datos de entrenamiento y test mediante un bloque “X-partitioner” y finalmente se ejecutan 10 veces mediante un bloque “X-Agreggator”.

2.4. Análisis de resultados

En esta sección se toman los resultados de la parte anterior y interpretan los resultados.

Para ello se utilizan bloques “Scorer” que generan las matrices de confusión y bloques “Heatmap” para representarlas de manera mas visual.

3. Resultados

A continuación se van a ver los diferentes resultados de las matrices de confusión para varios valores de K

Es importante puntualizar que debido a la naturaleza del algoritmo K-Nearest neighbors, dados los mismos datos y el mismo valor de K, toda implementación correcta del algoritmo dará el mismo resultado por lo que solo se mostrarán dos matrices de confusión por cada valor de K, la matriz de confusión del dataset normal y la del dataset equilibrado. Además, solo se mostrará la precisión en el caso del dataset equilibrado ya que no es precisa en el dataset normal.

3.1. K = 1

3.1.1. Dataset normal

	NO	YES
NO	2734	189
YES	224	50

3.1.2. Dataset Equilibrado

	NO	YES
NO	2615	308
YES	46	2877

La precisión es de 0.939

3.2. K = 2

3.2.1. Dataset normal

	NO	YES
NO	2894	29
YES	263	11

3.2.2. Dataset Equilibrado

	NO	YES
NO	2674	249
YES	114	2809

La precisión es de 0.937

3.3. K = 3

3.3.1. Dataset normal

	NO	YES
NO	2846	77
YES	239	35

3.3.2. Dataset Equilibrado

	NO	YES
NO	2469	454
YES	45	2878

La precisión es de 0.914

3.4. K = 5

3.4.1. Dataset normal

	NO	YES
NO	2879	247
YES	44	27

3.4.2. Dataset Equilibrado

	NO	YES
NO	2375	548
YES	64	2375

La precisión es de 0.895

3.5. K = 10

3.5.1. Dataset normal

	NO	YES
NO	2917	6
YES	226	8

3.5.2. Dataset Equilibrado

	NO	YES
NO	2296	627
YES	137	2786

La precisión es de 0.869
Cómo se puede ver la precisión disminuye cuánto mas se aumenta la K