4 Preparación de los datos

1. Creamos un workflow llamado “Preparación de los datos”
2. Nodo “File Reader” para la lectura del dataset adult.data
3. Realizamos una primera exploración a los datos con los nodos “Statistics” “Histogram (Interactive)”
4. Eliminamos outliers con un “Row filter” en la variable edad.
5. Filtramos con el nodo “Column Filter” los campos:
    a. education_num → Es equivalente a education.
    b. Fnlwgt → Personas que cumplen las características del censo.
6. Con “Missing Value” filtramos todos los registros que contengan algún valor vacío.
7. Seguimos explorando los datos. En “Histogram (Interactive)” podemos comprobar que para la variable workclass, se puede reducir las categorías para que el modelo sea más sencillo. Agregamos las que son funcionalmente parecidas y además, individualmente son poco significativas. Utilizaremos el nodo “Rule Engine” e introduciremos las siguientes reglas:
    a. Local-gov → Other-gov
    b. State-gov → Other-gov
    c. Self-emp-inc → Self-employed
    d. Self-emp-not-inc → Self-employed
    e. Without-pay → Not-working
    f. Never-worked → Not-working
8. Ahora reduciremos las categorías para education
    a. 12th → Not-HS-Graduate
    b. 11th → Not-HS-Graduate
    c. 10th → Not-HS-Graduate
    d. 9th → Not-HS-Graduate
    e. 7th-8th → Not-HS-Graduate
    f. 5th-6th → Not-HS-Graduate
    g. 1th-4th → Not-HS-Graduate
    h. Preschool → Not-HS-Graduate
    i. Assoc-acdm → Associates
    j. Assoc-voc → Associates
    k. Some-college → HS-Grad

    - En KNIME:

    $education$ = "12th" => "Not_HS_Graduate"
    $education$ = "11th" => "Not_HS_Graduate"
    $education$ = "10th" => "Not_HS_Graduate"
    $education$ = "9th" => "Not_HS_Graduate"
    $education$ = "7th-8th" => "Not_HS_Graduate"
    $education$ = "5th-6th" => "Not_HS_Graduate"

$education$ = "1st-4th" => "Not_HS_Graduate"
$education$ = "Preschool" => "Not_HS_Graduate"
$education$ = "Assoc-acdm" => "Associates"
$education$ = "Assoc-voc" => "Associates"
$education$ = "Some-college" => "HS-Grad"
TRUE => $education$

9. Trabajaremos también con la variable marital-status
    a. Married-AF-spouse → Married
    b. Married-civ-spouse → Married
    c. Married-spouse-absent → Not-Married
    d. Separated → Not-Married
    e. Divorced → Not-Married

- En KNIME:

$marital-status$ = "Married-AF-spouse" => "Married"
$marital-status$ = "Married-civ-spouse" => "Married"
$marital-status$ = "Married-spouse-absent" => "Not-Married"
$marital-status$ = "Separated" => "Not-Married"
$marital-status$ = "Divorced" => "Not-Married"
TRUE => $marital-status$

10. Dividimos los datos para entrenamiento y test mediante el nodo “Partitioning”, elegimos el método Draw radomly
11. “Normalizer” con Z-Score Normalization (Gaussian), sobre las dos particiones.
12. Escribimos en csv
    a. training_set
    b. test_set