WEKA

GR31

Manuel Pérez Vélez,
Julia Sánchez Márquez

WEKA

The workbench for machine learning

WEKA

¿Qué es?

Es un **software** desarrollado por la Universidad de Waikato, en Nueva Zelanda, que se utiliza para realizar **minería de datos** y **aprendizaje automático**.

Está diseñado para facilitar el análisis de datos y la construcción de modelos predictivos de forma **sencilla**, incluso para personas que no son expertas en programación

Su nombre proviene de **Waikato Environment for Knowledge Analysis**

WEKA

¿Qué ofrece?

Una **interfaz gráfica** que permite aplicar algoritmos de machine learning directamente a conjuntos de datos, sin necesidad de escribir código.

También incluye herramientas para **preprocesamiento de datos**, **selección de atributos**, **visualización**, y **evaluación de modelos**.

Por ser **de código abierto** y **gratuito**, es muy popular en ambientes académicos, investigaciones y proyectos de ciencia de datos

WEKA

¿Qué archivos soporta?

- ARFF (.arff)
  - Es el formato nativo de Weka (**Attribute-Relation File Format**).
  - Es un archivo de texto que contiene:
    - Una sección para **definir los atributos** (nombre y tipo de dato).
    - Y otra sección para **listar los datos**

Es el ideal para que Weka entienda claramente los datos y los tipos

```
@relation filtrado-weka.filters.unsupervised.instance.RemoveWithValues-S0.0-C24-L2

@attribute Marital_status numeric
@attribute Application_mode numeric
@attribute Application_order numeric
@attribute Course numeric
@attribute Daytime/evening_attendance numeric
@attribute Previous_qualification numeric
@attribute Previous_qualification_(grade) numeric
@attribute Nacionality numeric
@attribute Mother_s_qualification numeric
@attribute Father_s_qualification numeric
@attribute Mother_s_occupation numeric
@attribute Father_s_occupation numeric
@attribute Admission_grade numeric
@attribute Displaced numeric
@attribute Educational_special_needs numeric
@attribute Debtor numeric
@attribute Gender numeric
@attribute Scholarship_holder numeric
@attribute Age_at_enrollment numeric
@attribute International numeric
@attribute Unemployment_rate numeric
@attribute Inflation_rate numeric
@attribute GDP numeric
@attribute TARGET {Dropout,Graduate}

@data
1,17,5,171,1,1,122,1,19,12,5,9,127.3,1,0,0,1,0,20,0,10.8,1.4,1.74,Dropout
1,15,1,9254,1,1,160,1,1,3,3,3,142.5,1,0,0,1,0,19,0,13.9,-0.3,0.79,Graduate
1,1,5,9070,1,1,122,1,37,37,9,9,124.8,1,0,0,1,0,19,0,10.8,1.4,1.74,Dropout
1,17,2,9773,1,1,122,1,38,37,5,3,119.6,1,0,0,0,0,20,0,9.4,-0.8,-3.12,Graduate
```

WEKA

- ARFF (.arff)

```
@relation filtrado-weka.filters.unsupervised.instance.RemoveWithValues-S0.0-C24-L2

@attribute Marital_status numeric
@attribute Application_mode numeric
@attribute Application_order numeric
@attribute Course numeric
@attribute Daytime/evening_attendance numeric
@attribute Previous_qualification numeric
@attribute Previous_qualification_(grade) numeric
@attribute Nacionality numeric
@attribute Mother_s_qualification numeric
@attribute Father_s_qualification numeric
@attribute Mother_s_occupation numeric
@attribute Father_s_occupation numeric
@attribute Admission_grade numeric
@attribute Displaced numeric
@attribute Educational_special_needs numeric
@attribute Debtor numeric
@attribute Gender numeric
@attribute Scholarship_holder numeric
@attribute Age_at_enrollment numeric
@attribute International numeric
@attribute Unemployment_rate numeric
@attribute Inflation_rate numeric
@attribute GDP numeric
@attribute TARGET {Dropout,Graduate}

@data
1,17,5,171,1,1,122,1,19,12,5,9,127.3,1,0,0,1,0,20,0,10.8,1.4,1.74,Dropout
1,15,1,9254,1,1,160,1,1,3,3,3,142.5,1,0,0,1,0,19,0,13.9,-0.3,0.79,Graduate
1,1,5,9070,1,1,122,1,37,37,9,9,124.8,1,0,0,1,0,19,0,10.8,1.4,1.74,Dropout
1,17,2,9773,1,1,122,1,38,37,5,3,119.6,1,0,0,0,0,20,0,9.4,-0.8,-3.12,Graduate
```

ibute-

ributos

atos

y los tipos

¿Q
arch
sopo

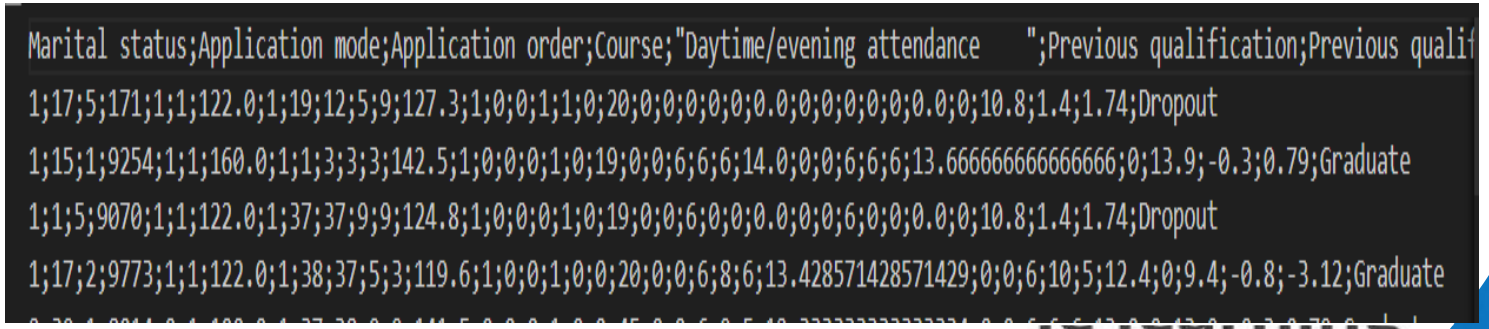
WEKA

¿Qué archivos soporta?

➢ CSV (.csv)

- **Comma-Separated Values**: los datos están separados por comas o puntos y comas.
- No tiene una sección especial para definir tipos de datos, pero Weka puede **adivinarlos automáticamente** o tú puedes especificarlo al importar.

Muy usado porque es **fácil de exportar** desde Excel



WEKA

¿Qué formato hemos usado?

```
@ATTRIBUTE TARGET {Dropout,Enrolled,Graduate}
```

```
@ATTRIBUTE Curricular_units_2nd_sem_(approved) NUMERIC
@ATTRIBUTE Curricular_units_2nd_sem_(grade) NUMERIC
@ATTRIBUTE Curricular_units_2nd_sem_(without_evaluations) NUMERIC
```

```python
def infer_arff_type(series):
    """Determina si una columna es numérica o de texto."""
    if pd.api.types.is_numeric_dtype(series):
        return 'NUMERIC'
    else:
        return 'STRING'


def generate_arff(df_input, output_path, relation_name):
    arff_lines = []
    arff_lines.append(f"@RELATION {relation_name}\n")

    # 1) Definición de atributos
    for col in df_input.columns:
        if col.upper() == 'TARGET':
            unique_values = sorted(df_input[col].dropna().unique())
            # Los valores no llevan comillas
            values_str = ",".join(str(v) for v in unique_values)
            arff_lines.append(f"@ATTRIBUTE TARGET {{{values_str}}}")
        else:
            col_name = sanitize_attribute_name(col)
            arff_type = infer_arff_type(df_input[col])
            arff_lines.append(f"@ATTRIBUTE {col_name} {arff_type}")
    arff_lines.append("")  # línea en blanco antes de @DATA
    arff_lines.append("@DATA")
```

WEKA

Principales tareas

- **Clasificación** (predecir categorías).
- **Regresión** (predecir valores numéricos).
- **Agrupamiento** (clustering).
- **Reducción de características** (selección y extracción de atributos).
- **Minería de reglas de asociación** (descubrir relaciones entre variables).

En nuestro caso de la que más hemos hecho uso es de la **clasificación**

WEKA

Algoritmos que tiene implementados

En nuestro caso solo nos centraremos en los algoritmos de **clasificación**

Estos sirven para predecir una categoría o clase.

- **Árboles de decisión**:
  - J48
  - Random Tree
  - Random Forest (muchos árboles juntos)
- **Redes bayesianas**:
  - Naive Bayes
  - BayesNet
- **Máquinas de soporte vectorial** (SVM):
  - SMO (Sequential Minimal Optimization)
- **k-Vecinos más cercanos**:
  - IBk (implementa k-NN)
- **Reglas de decisión**:
  - JRip
  - PART
- **Redes neuronales**:
  - Multilayer Perceptron
- **Meta-classifier:**
  - LogitBoost (mejores resultados y el usado)

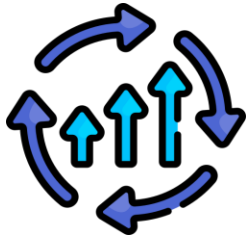
WEKA

Algoritmo usado en nuestro caso

LogitBoost + REPTree

LogitBoost + DecisionStump

LogitBoost



LogitBoost mejora la precisión de los modelos de predicción a lo largo de las iteraciones.

Cada **modelo** corregirá los errores de los modelos anteriores.

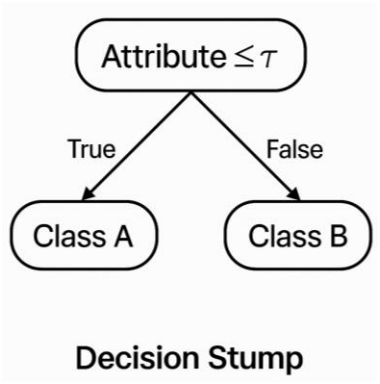
Forman un **clasificador final** que tiene una mayor precisión que cualquier modelo individual.

WEKA

Algoritmo usado en nuestro caso

LogitBoost + REPTree

LogitBoost + DecisionStump



Decision Stump

Es un árbol con una sola decisión. Hace una partición en base a un único atributo con un umbral o valor categórico

Modelo **extremadamente simple**. Con más atributos útiles, se necesita **menos** "complejidad interna" en cada árbol y **más** "cooperación" entre muchos modelos simples

Por sí solo es poco efectivo para clasificaciones complejas

WEKA

Algoritmo usado en nuestro caso

LogitBoost + REPTree

LogitBoost + REPTree

REPTree (Reduced Error Pruning Tree) es un algoritmo de **árbol de decisión** rápido y eficiente que se utiliza tanto para **clasificación** como para **regresión**. Se usa mucho en Weka y es conocido por su **velocidad** y **capacidad de generalización**.

Incorpora **poda** para evitar el sobreajuste, eliminando ramas que no aportan a la predicción final.

Destaca en tareas pequeñas y directas

WEKA

Algoritmo usado en nuestro caso

LogitBoost + REPTree

REPTree vs DecisionStump

REPTree

`Archivo_completo_sin_enrolled.arff`

8 expertos, cada uno muy inteligente, pero si se equivocan un poco, pueden insistir demasiado en su opinión.

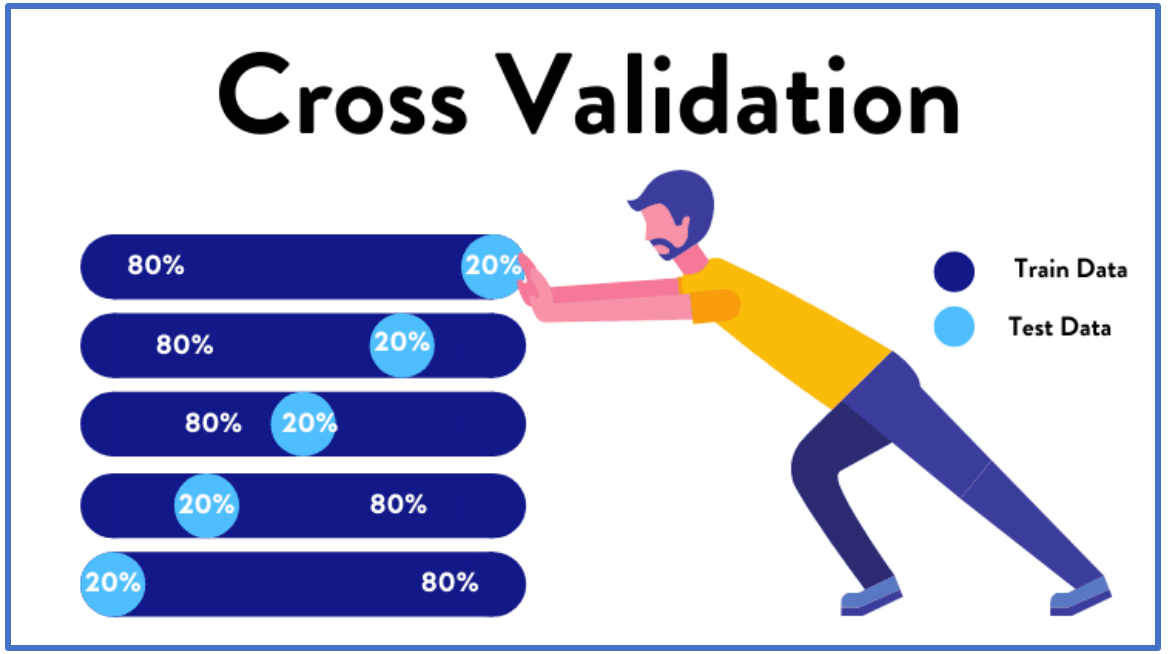
`Archivo_completo_sin_enrolled.arff`

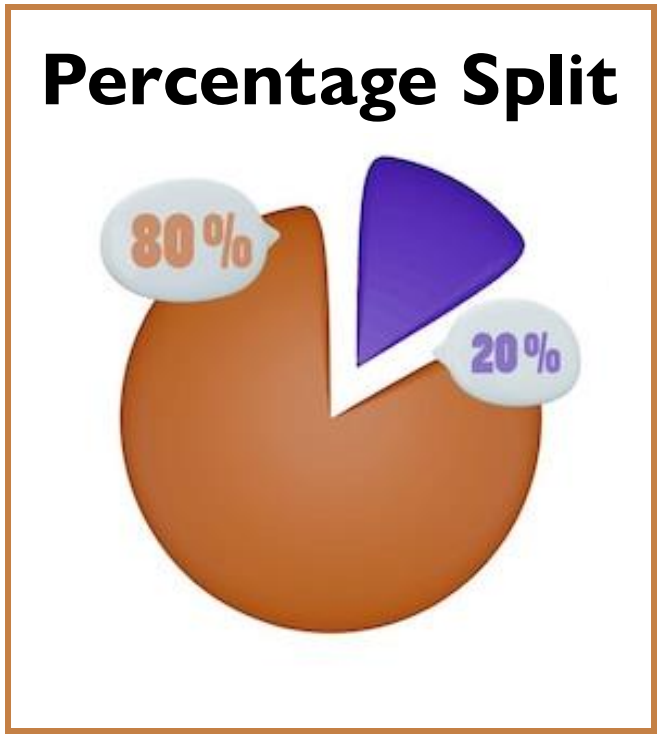
100 personas con una idea muy **simple**, que poco a poco, corrigen los errores y se van acercando juntas a la mejor solución.

DecisionStump

WEKA

¿Métricas usadas?





WEKA

¿Qué archivos hemos usado?

Archivo_filtrado_sin_enrolled.arff
Archivo_completo_sin_enrolled.arff
Archivo_filtrado_con_enrolled.arff
Archivo_completo_con_enrolled.arff

`@ATTRIBUTES DROP`

Nacionalidad

Internacional

ciales

Desempleo

Inflación

WEKA

¿Qué archivo hemos usado?

Archivo_filtrado_sin_enrolled.arff

Archivo_completo_sin_enrolled.arff

Conjunto entrenamiento

Archivo_filtrado_solo_enrolled.arff

Archivo_completo_solo_enrolled.arff

Conjunto prueba

WEKA

Resultados Modelo Final

Archivo Filtrado Sin Enrolled

```
=== Summary ===

Correctly Classified Instances        3162               87.1074 %
Incorrectly Classified Instances       468               12.8926 %
Kappa statistic                          0.7323
Mean absolute error                      0.0972
Root mean squared error                  0.2674
Relative absolute error                 30.3351 %
Root relative squared error             66.8165 %
Total Number of Instances             3630
```

¡¡¡87.10!!!

```
=== Confusion Matrix ===

    a    b    c   <-- classified as
 1232    0  224 |    a = Dropout
    0    0    0 |    b = Enrolled
  244    0 1930 |    c = Graduate
```

Archivo Completo Sin Enrolled

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3288               90.5785 %
Incorrectly Classified Instances       342                9.4215 %
Kappa statistic                          0.7994
Mean absolute error                      0.0888
Root mean squared error                  0.2181
Relative absolute error                 27.9341 %
Root relative squared error             54.7256 %
Total Number of Instances             3630

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,840    0,052    0,913      0,840   0,875      0,801    0,952     0,949     Dropout
               ?        0,000    ?          ?       ?          ?        ?         ?         Enrolled
               0,948    0,160    0,902      0,948   0,925      0,801    0,952     0,957     Graduate
Weighted Avg.  0,906    0,118    0,906      0,906   0,905      0,801    0,952     0,954

=== Confusion Matrix ===

    a    b    c   <-- classified as
 1193    0  228 |    a = Dropout
    0    0    0 |    b = Enrolled
  114    0 2095 |    c = Graduate
```

90.58%

WEKA

Algoritmo usado en nuestro caso

LogitBoost + REPTree

REPTree vs DecisionStump

Archivo completo sin enrolled.arff

Métrica	LogitBoost + REPTree (8 iter.)	LogitBoost + DecisionStump (100 iter.)
Precisión Global	89.31%	90.58%

nuy
uivocan un
masiado en

Archivo_completo_sin_enrolled.arff

100 personas con una idea muy **simple**, que poco a poco, corrigen los errores y se van acercando juntas a la mejor solución.

DecisionStump

WEKA

Algoritmo usado en nuestro caso

LogitBoost + REPTree

REPTree vs DecisionStump

Archivo completo sin enrolled.arff

Métrica	LogitBoost + REPTree (8 iter.)	LogitBoost + DecisionStump (100 iter.)
Precisión Global	89.31%	90.58%

87.1%

nuy
uivocan un
masiado en

Archivo_completo_sin_enrolled.arff

100 personas con una idea muy **simple**, que poco a poco, corrigen los errores y se van acercando juntas a la mejor solución.

DecisionStump

WEKA

Conclusiones

Conjunto de prueba

Archivo_filtrado_solo_enrolled.arff

Predicciones **muy** cercanas a 1


```
=== Predictions on test set ===

    inst#     actual  predicted error prediction
        1        1:? 3:Graduate       0.983
        2        1:? 3:Graduate       0.73
        3        1:?  1:Dropout       0.511
        4        1:? 3:Graduate       1
        5        1:? 3:Graduate       0.99
        6        1:?  1:Dropout       0.694
        7        1:? 3:Graduate       0.929
        8        1:?  1:Dropout       0.691
        9        1:? 3:Graduate       0.629
       10        1:? 3:Graduate       0.938
       11        1:? 3:Graduate       0.812
       12        1:?  1:Dropout       0.625
       13        1:? 3:Graduate       0.913
       14        1:? 3:Graduate       0.708
       15        1:?  1:Dropout       0.985
       16        1:?  1:Dropout       0.998
       17        1:?  1:Dropout       0.989
       18        1:?  1:Dropout       0.994
       19        1:? 3:Graduate       0.648
       20        1:?  1:Dropout       1
       21        1:?  1:Dropout       0.62
       22        1:? 3:Graduate       0.93
       23        1:?  1:Dropout       0.919
       24        1:? 3:Graduate       0.998
       25        1:? 3:Graduate       0.993
       26        1:? 3:Graduate       0.996
       27        1:? 3:Graduate       0.986
       28        1:? 3:Graduate       0.938
       29        1:? 3:Graduate       0.994
       30        1:?  1:Dropout       0.769
       31        1:?  1:Dropout       0.977
       32        1:? 3:Graduate       0.808
       33        1:?  1:Dropout       0.972
       34        1:?  1:Dropout       0.857
       35        1:? 3:Graduate       0.867
       36        1:? 3:Graduate       0.941
       37        1:? 3:Graduate       0.638
       38        1:? 3:Graduate       0.599
       39        1:? 3:Graduate       0.991
       40        1:? 3:Graduate       0.999
```

WEKA

Conclusiones

Perfiles Estudiantiles

Buenas notas de admisión
Padres con mayor **nivel de cualificación**
Edad ingreso en rangos **moderados**
Becarios
No **desplazamiento**
No **necesidades especiales**

Deudas
Entornos socioeconómicos **desfavorecidos**
Mayores al ingresar
No reciben **becas**
Historial académico **débil**

WEKA

Conclusiones

Técnicas de prevención