

# Análisis predictivo con KNIME

---

SISTEMAS INTELIGENTES PARA LA GESTIÓN DE LA EMPRESA  
CURSO 2016-2017

# Plataforma KNIME

---

KNIME es una herramienta software para manipulación, análisis y visualización de datos

## Características

- Basada en programación visual
- *Open source*
- Ecosistema de módulos (+1500)
  - Análisis descriptivo
  - Predicción
  - Análisis de redes y grafos
  - Minería de textos
- Puede utilizarse en combinación con otras plataformas: Weka, Java, Scala, R, Python, Hadoop/Spark, etc.

<https://www.knime.org>

# El entorno de programación KNIME

The screenshot displays the KNIME Analytics Platform interface with a workflow titled "Example Workflow". The workflow consists of the following nodes:

- File Reader**: Read iris.csv
- Color Manager**: Assign colors
- Statistics**: Calculates statistic measures: mean, max, min, variance, median, etc.
- Partitioning**: Split data 60/40
- Decision Tree Learner**: Train model
- Decision Tree Predictor**: Apply model
- Scatter Plot**: View test data
- Scorer**: Compute confusion matrix
- Interactive Table**: Explore test data

A yellow box highlights a text description of the workflow:

This Example Workflow uses a File Reader node to import the iris dataset (included). It then assigns visual properties with a Color Manager node and computes some basic statistics with a Statistics node. The data is split into training and testing fractions with a Partitioning node. The Decision Tree Learner generates a predictive model in PMML from the training fraction which is then applied to the test fraction using the Decision Tree Predictor. Model performance is evaluated with a Scorer node, which is applied after the Decision Tree Predictor. Finally, errors can be explored interactively, by using an Interactive Table node to highlight certain classes of errors which can then be visualized using a Scatter Plot node.

The Node Description panel on the right shows details for the **Decision Tree Learner** node:

This node induces a classification decision tree in main memory. The target attribute must be nominal. The other attributes used for decision making can be either nominal or numerical. Numeric splits are always binary (two outcomes), dividing the domain in two partitions at a given split point. Nominal splits can be either binary (two outcomes) or they can have as many outcomes as nominal values. In the case of a binary split the nominal values are divided into two subsets. The algorithm provides two quality measures for split calculation; the gini index and the gain ratio. Further, there exist a post pruning method to reduce the tree size and increase prediction accuracy. The pruning method is based on the minimum description length principle. The algorithm can be run in multiple threads, and thus, exploit multiple processors or cores. Most of the techniques used in this decision tree implementation can be found in "C4.5 Programs for machine learning", by J.R. Quinlan and in "SPRINT: A Scalable Parallel Classifier for Data Mining", by J. Shafer, R. Agrawal, M. Mehta (<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.104.152&rep=rep1&type=pdf>) If the optional PMML import is connected and contains preprocessing operations in the TransformationDictionary those are added to the learned model.

**Dialog Options**

**Class column**  
To select the target attribute. Only nominal attributes are allowed

**Quality measure**

The Console panel at the bottom shows the following output:

```
KNIME Console
*****
*** Welcome to the KNIME Analytics Platform v3.2.1.v281608190927 ***
*** Copyright by KNIME GmbH, Konstanz, Germany ***
*****
Log file is located at: /Users/jgomez/Documents/OneDrive/Trabajo/docencia/2016-2017/Sistemas Inteligentes para la Gestion en la Empresa/practicas/sesion 1/workspace/
WARN Decision Tree Predictor 0:4 DataColumnSpec already contains a color handler, ignoring color handler from second spec.
WARN Scatter Plot 0:8 Some columns are ignored: too many/missing nominal values.
```

Updating Software: (20%)

# Ejemplo: Predicción con *Iris Dataset*

Dataset clásico en aprendizaje automático  
– R. Fisher (1936) *The use of multiple measurements in taxonomic problems*

## Descripción

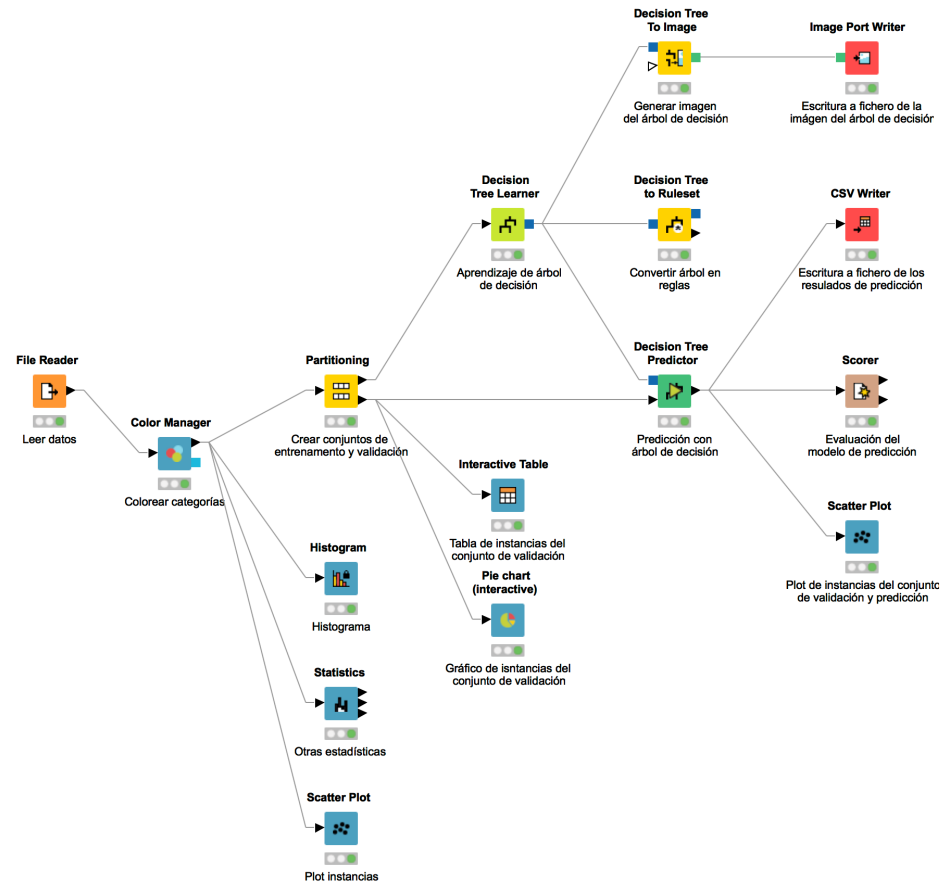
- Tres especies de flores: iris setosa, iris virginica, iris versicolor)
- 50 muestras de cada especie
- 4 características: longitud y anchura de pétalos y sépalos

## Objetivo

- Obtener la clase de una flor dados los valores de las 4 características

sepal length	sepal width	petal length	petal width	class
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5.0	3.6	1.4	0.2	Iris-setosa
7.0	3.2	4.7	1.4	Iris-versicolor
6.4	3.2	4.5	1.5	Iris-versicolor
6.9	3.1	4.9	1.5	Iris-versicolor
5.5	2.3	4.0	1.3	Iris-versicolor
6.5	2.8	4.6	1.5	Iris-versicolor
6.7	2.5	5.8	1.8	Iris-virginica
7.2	3.6	6.1	2.5	Iris-virginica
6.5	3.2	5.1	2.0	Iris-virginica
6.4	2.7	5.3	1.9	Iris-virginica
6.8	3.0	5.5	2.1	Iris-virginica
...				

# Ejemplo: Predicción con *Iris Dataset*



# Ejemplo: Predicción de crimen

*Challenge* organizado por el National Institute of Justice con datos reales sobre eventos criminales en Portland [\[link\]](#)

## Descripción

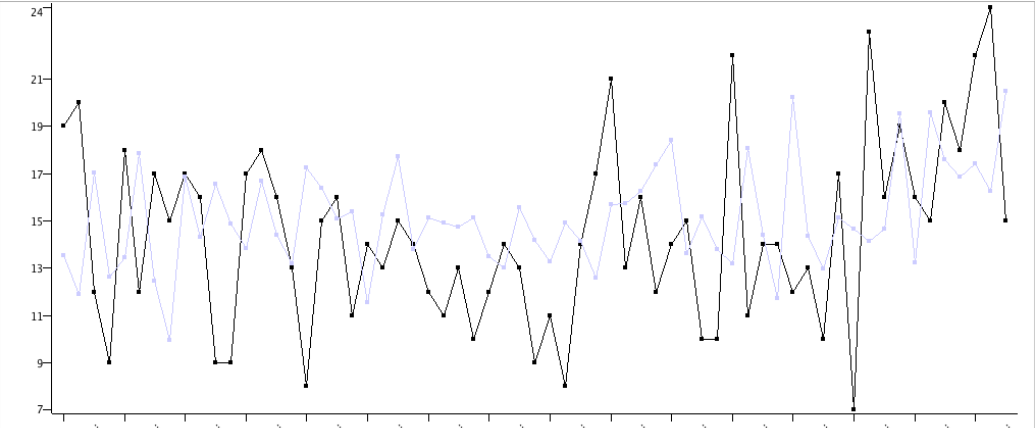
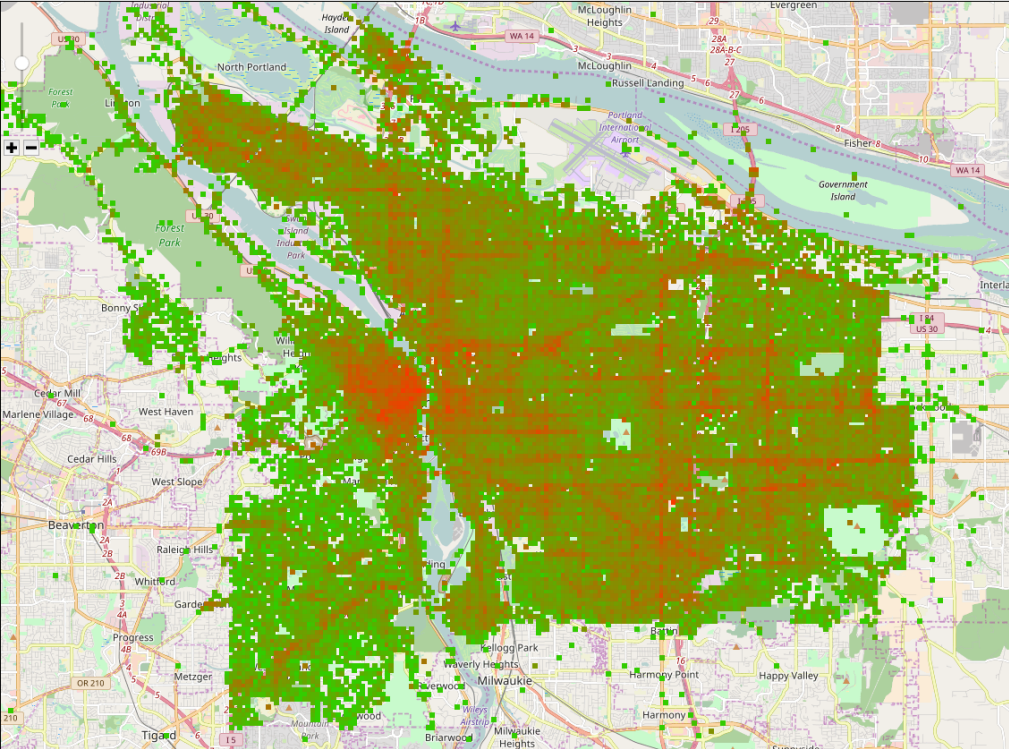
- Datos de eventos criminales de 2013 en adelante, geolocalizados
- 8 características: Categoría, Grupo de Llamada, Identificación Final del Caso, Descripción, Fecha, Coordenadas y Localización Censal
- Dependencias entre características

## Objetivo

- Predecir zonas de concentración de crímenes

CATEGORY	CALL_GROUPS	final_case	CASE_DESC	occ_date	x_coordinate	y_coordinate	census_tract
STREET CRIMES	DISORDER	DISTP	DISTURBANCE - PRIORITY	1/12/16	7627009	710228	4102
STREET CRIMES	DISORDER	DISTP	DISTURBANCE - PRIORITY	1/12/16	7627109	710045	4102
STREET CRIMES	DISORDER	DISTP	DISTURBANCE - PRIORITY	1/12/16	7644761	690250	2303
STREET CRIMES	DISORDER	DISTP	DISTURBANCE - PRIORITY	1/12/16	7649826	680465	1101
STREET CRIMES	DISORDER	DISTP	DISTURBANCE - PRIORITY	1/12/16	7664343	696396	7500
STREET CRIMES	DISORDER	DISTP	DISTURBANCE - PRIORITY	1/12/16	7666986	663517	8600
STREET CRIMES	DISORDER	DISTP	DISTURBANCE - PRIORITY	1/12/16	7669275	668437	602

# Ejemplo: Predicción de crimen



# Documentación adicional

---

## **KNIME Learning Hub**

<https://www.knime.org/learning-hub>

## **KNIME Online Self-Training**

<https://www.knime.org/knime-online-self-training>

## **O'Reilly “Introduction to Data Analytics with KNIME”**

<https://www.safaribooksonline.com/library/view/introduction-to-data/9781491967546/>

## **KNIME Beginner’s Luck (Rosaria Filipo)**

<https://www.knime.org/knimepress/beginners-luck>

## **KNIME Node Guide**

<https://www.knime.org/nodeguide>

## **Coursera “Machine Learning with Big Data”**

<https://es.coursera.org/learn/big-data-machine-learning>