TEMA 2. ADQUISICIÓN Y VISUALIZACIÓN DE DATOS

Contenidos

I. Introducción
II. Adquisición de datos (extracción)
III. Transformación de datos
IV. Carga de datos
V. Visualización de datos
VI. Actividades

I. Introducción

Los datos en el entrenamiento son cruciales.

- Si son de baja calidad o incompletos, el aprendizaje será inadecuado.

Se necesitan datos de entrenamiento para resolver cualquier problema.

- En el diagnóstico de enfermedades, los datos de entrenamiento son las variables médicas junto al diagnóstico del médico.

<u>Preguntas comunes previas al aprendizaje:</u>

- ¿Los datos del problema son representativos?
  Con esos datos, ¿estamos modelando bien el problema?
- ¿Necesitamos obtener la clase?
  Depende de la tarea requerida. En caso de que sea necesario, para obtener la clase:
  - Personas expertas: diagnóstico de enfermedades
  - Variable medible: precio de las casas

I. Introducción

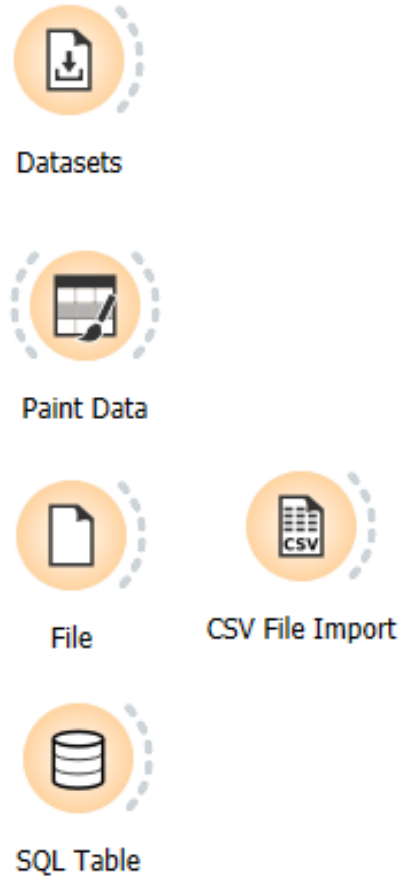
Posibles problemas con el conjunto de datos de entrenamiento:

- **Ruido**: Instancias con etiqueta errónea, datos procedentes de sensores no fiables.
- **Valores perdidos**: El valor de unos cuantos atributos para algunas instancias son desconocidos.
- **Dimensión**: Muchos sistemas aprenden muy lentamente cuando el número de atributos del conjunto de entrenamiento es alto.
- **Una mala selección de atributos**: atributos irrelevantes sin influencia sobre el conocimiento que intentamos inferir, atributos redundantes,...
- **Proporción entre número de atributos e instancias**: el conocimiento descubierto no es fiable cuando usamos muchos atributos y pocas instancias.

II. Adquisición de datos (extracción)

- **Extracción de datos**: técnicas para adquirir datos desde diferentes orígenes y almacenarlos en estructuras intermedias para su transformación y carga posteriores.

- **Orígenes de datos**:
  - Repositorios de datasets (**En Orange: nodo Datasets**)
  - Datos sintéticos (artificiales) (**En Orange: nodo Paint Data**)
  - Ficheros (**En Orange: nodo File o CSV File Import** : TXT, CSV, Excel, TAB, …)
  - Bases de datos (**En Orange: Conector con PostgreSQL**)
  - APIs (web services)
  - Web scrapping



II. Adquisición de datos (extracción)

Repositorios Open Data:

- UCI (Machine Learning Repository): https://archive.ics.uci.edu/ml/datasets.php
- Competición Kaggle: https://www.kaggle.com
- BUFA (Bilkent University Function Approximation Repository): http://funapp.cs.bilkent.edu.tr/DataSets/
- NCBI (National Center for Biotechnology Information): https://www.ncbi.nlm.nih.gov/gds
- Datos Abiertos Universidad de Granada: http://opendata.ugr.es/
- Datos Abiertos Espaciales del Ayuntamiento de Sevilla: http://datosabiertos.sevilla.org/
- Datos Abiertos Ayuntamiento de Barcelona: http://opendata-ajuntament.barcelona.cat/es/
- Proyecto del Genoma del Cáncer Pediátrico de la Universidad de Washington, del Hospital Infantil St. Jude, datos completos del genoma del cáncer humano: https://www.stjude.org/research/pediatric-cancer-genome-project.html

III. Transformación de datos

Técnicas para manipular, transformar, limpiar e integrar datos previamente extraídos con el objetivo de producir datos limpios que cargar posteriormente.

- **Información datos** (**en Orange: nodo Data Info y Data Table**)

- **Tipos de transformaciones de datos**:
  - Cambiar formato o tipo de datos en campos
  - Reestructurar valores de campos (fusionar o dividir campos)
  - Filtrado de filas y/o columnas (según criterios matemáticos o por muestreo aleatorio)
  - Reorganización de filas y columnas (pivotar tablas)
  - Resumen, agrupaciones y estadísticas
  - Reunión (integración)
  - Creación de nuevos campos calculados
  - Limpieza: Normalización de valores numéricos, Tratamiento de valores ausentes, Detección y corrección de outliers, Discretización de atributos continuos, ….

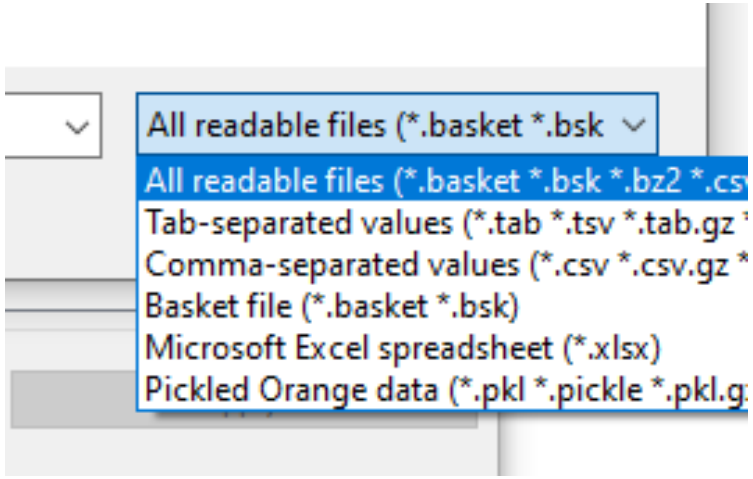
III. Transformación de datos

Nodos útiles en Orange:

IV. Carga de datos

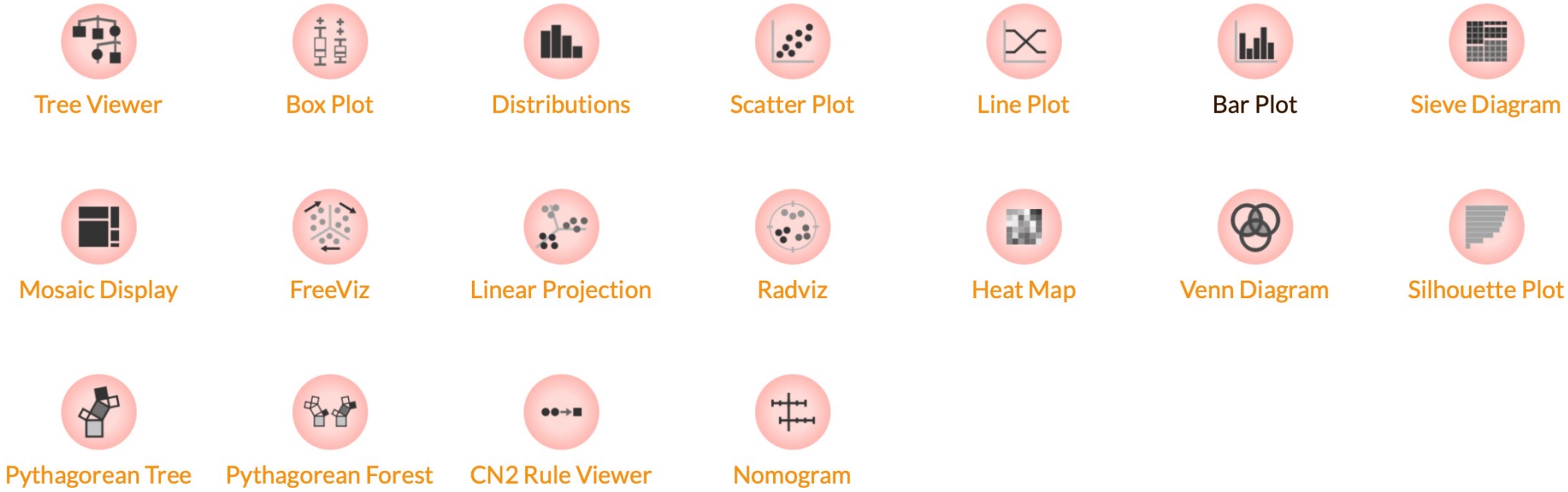
Después de las transformaciones, los datos se deben guardar (**en Orange: nodo Save Data**) para posteriormente poder introducirlos en el modelo de aprendizaje sin tener que volver a aplicarle todas las transformaciones.

En Orange los datos se pueden importar de un fichero en varios formatos: Excel (.xlsx), CSV (.csv), archivos de texto plano (.txt), URLS, textos separados por tabulación (.tab). Los datos también se pueden leer de una base de datos SQL.



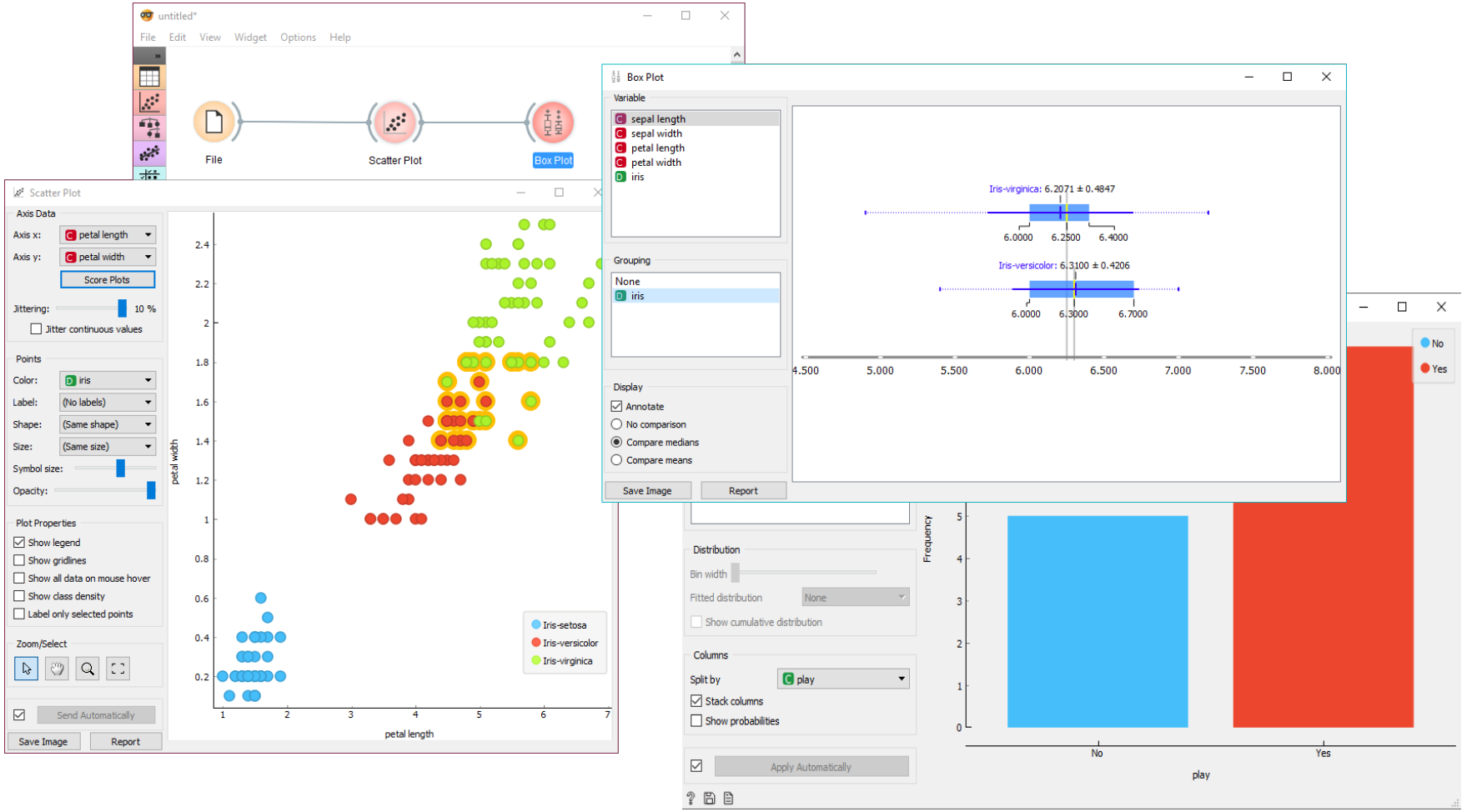
V. Visualización de datos

Visualize



https://orangedatamining.com/widget-catalog

VI. Visualización de datos



VII. Actividad 1: Adquisición y visualización de datos

Trabajaremos con un dataset muy conocido en análisis de datos llamado ”Iris”.

El dataset se compone de 150 observaciones de flores de la planta iris.

Existen tres tipos de clases de flores iris: **virginica, setosa y versicolor**.

Hay 50 observaciones de cada una. Las variables o atributos que se miden de cada flor son:

1. El tipo de flor como variable categórica.
2. El largo y el ancho del pétalo en cm como variables numéricas.
3. El largo y el ancho del sépalo en cm como variables numéricas.

Figura: Virginica - Setosa - Versicolor

VII. Actividad 1: Adquisición y visualización de datos

1. Arrastre el nodo **Datasets** al canvas en la pestaña "Data".
2. Cargue el dataset **Iris**.
3. Indique de qué tipos son los atributos con el nodo **Data Info**.
4. Identifique toda la información: Nº de ejemplos, Nº Atributos, distribución de valores en las clases, características de cada atributo, … **Feature Statistics**
5. Use el nodo **Data Table** para seleccionar solamente las instancias de la clase “Iris-versicolor” e “Iris-setosa” y repita el paso 4.

VII. Actividad 2: Adquisición y visualización de datos

Trabajaremos con el dataset de “tennis”.

El dataset se compone de 14 observaciones de datos meteorológicos y si se jugó o no al tenis.

Las variables o atributos que se miden de cada instancia:

1. Pronóstico con valores categóricos.
2. Temperatura como variable numérica.
3. Humedad como variable numérica.
4. Viento como variable categórica.
5. **Si se jugó o no** como variable categórica.

VII. Actividad 2: Adquisición y visualización de datos

1. Arrastre el nodo **CSV File Import** al canvas en la pestaña "Data".
2. Busque la ubicación del archivo .csv "**tennis.csv**" y cambie las opciones de importación del fichero para que se ajusten al mismo ("Cell delimiter", "Quote character", "Number separators"). Ajuste el tipo de columna con "Column type".
3. Definamos cuál es la clase de nuestro dataset con **Select Columns**
4. ¿Cuántas filas y columnas hay? ¿Cuántos valores tiene la clase categórica? Use el nodo **Data Table**
5. Identifique toda la información: Nº de ejemplos, Nº Atributos, distribución de valores en las clases, características de cada atributo, … **Feature Statistics**