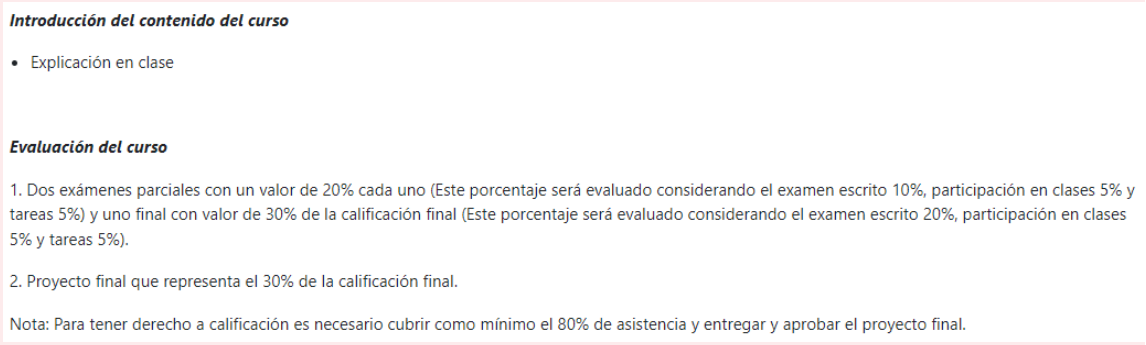
Minería de Datos

▼ Primer Parcial

▼ 08-08-2022

Docente: Miguel Ángel Meza de Luna



Introducción del contenido del curso

- Explicación en clase

Evaluación del curso

1. Dos exámenes parciales con un valor de 20% cada uno (Este porcentaje será evaluado considerando el examen escrito 10%, participación en clases 5% y tareas 5%) y uno final con valor de 30% de la calificación final (Este porcentaje será evaluado considerando el examen escrito 20%, participación en clases 5% y tareas 5%).

2. Proyecto final que representa el 30% de la calificación final.

Nota: Para tener derecho a calificación es necesario cubrir como mínimo el 80% de asistencia y entregar y aprobar el proyecto final.

Formas de trabajo:

- Asistencia cuenta mucho. Faltar perjudica la participación
- Las tareas probablemente serán semanales
- Probablemente examen práctico contrarreloj (2 o 3 horas) o para llevar
- El proyecto final que se entreguen avances. No obligatorios

Propuestas:

- Quitar las tareas y ponerle ese porcentaje a participación (queda tareas 0%, 10%) (Aceptable)
- Participación 0%, tareas 10%.

El proyecto requerirá autocapacitación trabajando en Python. Usaremos PowerBI, Excel, Python, R, otras.

Se queda la propuesta de participación 0%, tareas 10%.

▼ 09-08-2022

Muchos datos, poca información

Revisión de la presentación de introducción

- **Video Documental “Big Data: el valor de nuestra información”**
  - Revolución de los datos
  - Actos cotidianos están observados y proporciona información
  - No está bien pensar “no tengo nada que esconder”
  - No es igual compartir dónde estoy con cercanos que con desconocidos
  - Big data o datos masivos
  - Es necesario velocidad de procesamiento y tecnología para recopilar los datos
  - Transformación: Sociedad industrial a sociedad del conocimiento
  - La cantidad de información generada cada vez es mayor
  - Cifras muy potentes de búsquedas en Google y tweets realizados
  - Internet of things
  - La domótica. Calefacción en casas, parking, semáforos, reciclaje.
  - Minería de datos requiere también estadística y matemáticas
  - Recogida, integración, almacenamiento, validación, análisis y explotación
  - Datos → Conocimiento
  - Reconocimiento de patrones → Predicción
  - ¿Qué precauciones debemos tomar como usuarios sobre nuestra información?

- **Video Documental “Big Data & AI for bad guys”**
  - La IA puede usarse para bien como para mal
  - A veces a la gente no le importan sus datos de Gmail pero los de FB o Twitter sí
  - Ejemplo de Netflix para recomendar películas simplemente al consumir la aplicación

Ejemplo de uso de fake news para cambiar la opinión de una persona pese a quela gente piensa que “verifica la información”

Data + Fake News = Alert

La ubicación habla mucho, ¿por dónde estás? ¿con qué te transportas? y muchísimas muchísimas cosas para obtener información

Es peligroso los Data Brokers que venden los datos personales a terceros

A veces al darle Aceptar en “Tu privacidad nos importa” no nos damos cuenta en a todo lo que aceptamos para que los sitios webs se lleven nuestros datos y se vendan a otras empresas

Political Consulting Services son empresas que buscan regular la transmisión de datos con fines malos

Weaponizing features (ejemplo de lectura de documentos Word para trackear las computadoras de una empresa con los metadatos de los documentos)

Ejemplo de Grinder (tinder para homosexuales) haciendo Data Leakage.

Have I been pwned? Es una web que te dice si tus datos se han filtrado a la web. Habían fallos de seguridad que han pasado 5mil millones. Comenta la forma de weaponizar sobre las opciones de recuperaciones

Dirty Business Cards para hacer clustering

Clustering targets → Spear Fake News

Uso de IA para que ellos creen las fake news

Deep Fakes, Face Swaps

Generative Adversarial Networks usa dos IAs. Una que hace los Deep Fakes y las envía a las otra a ver si logra engañarla

Ahora se puede suplantar en tiempo real a una persona

▼ **10-08-2022**

No hubo clase

▼ **11-08-2022**

Investigación de salarios por Software Gurú en área de análisis de datos

Tarea 1: En la presentación, las últimas dos diapositivas vienen dos videos. Hay que verlos y tener elementos de interés

▼ **12-08-2022**

Descargar Orange Data Mining

Añadimos el widget File, lo conectamos a Data Table y File también lo conectamos a Scatter Plot

▼ **16-08-2022**

Review de los videos de tarea

▼ **17-08-2022**

Prácticas en Orange

▼ **18-08-2022**

Presentación de PowerPoint de KDD

▼ **19-08-2022**

No hubo clase

▼ **Tarea**

▼ 1. Welcome to Orange

Instalación de Orange

Al abrir Orange sale la pantalla de bienvenida

Al cerrarla, comienzas en Orange con una página en blanco

Orange trabaja con Widgets

Muchos de los análisis de datos empiezan con el widget File con el que se cargan los datos

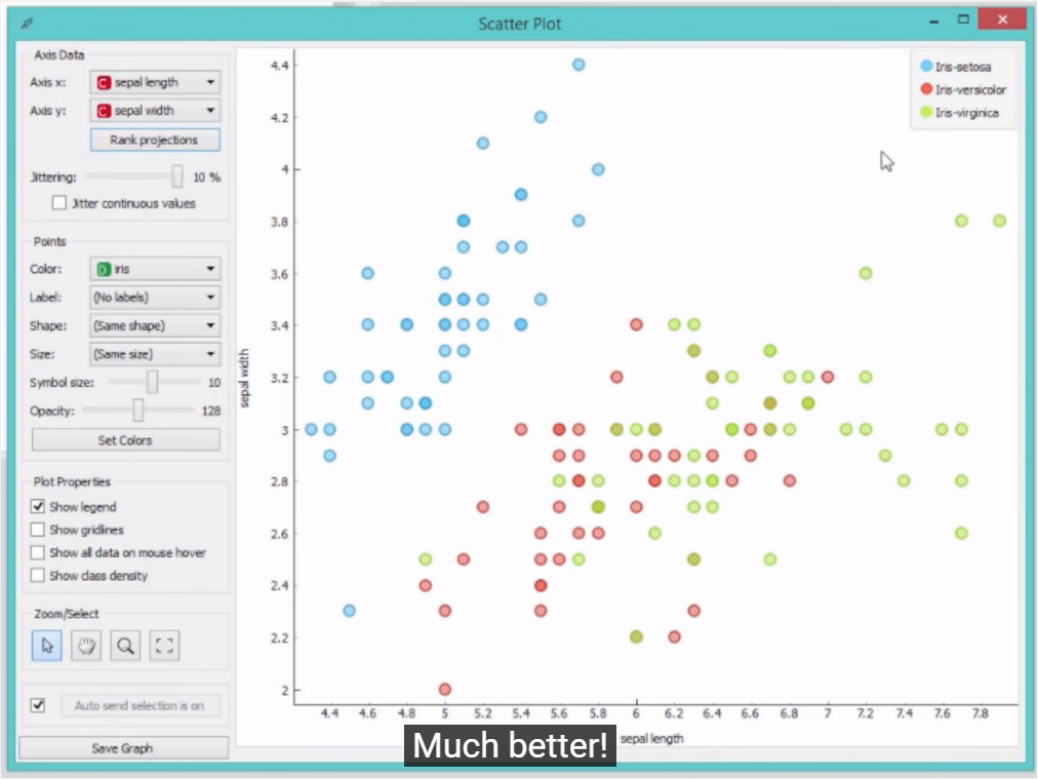
Al clickear aparece en la página

Al darle doble click al widget en la página se puede cargar un dataset

Lo siguiente:

Permite ver los datos del archivo cargado

Puede también graficarse los datos usando un Scatter Plot:

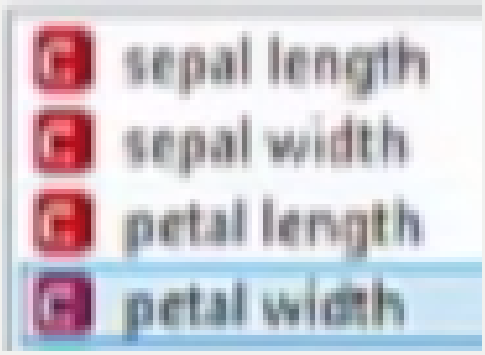


▼ 2. Data workflows

Se puede crear un flujo de análisis de datos de la siguiente forma:

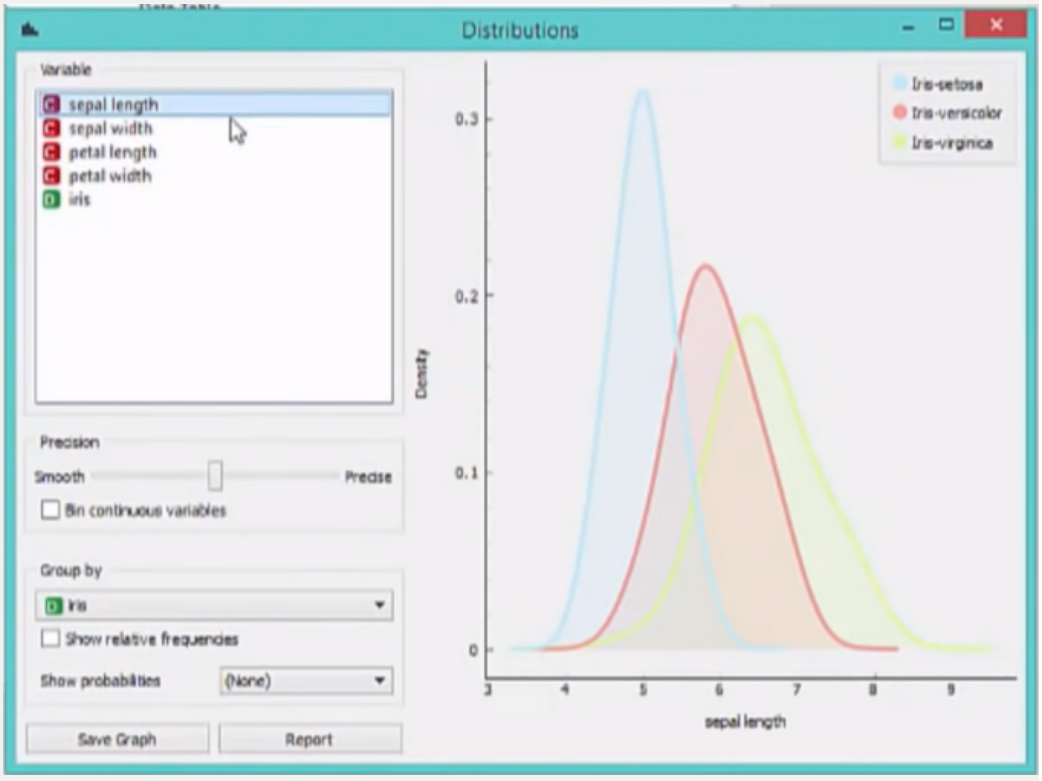
Se trabaja con el conjunto de Iris de Fischer

Tiene cuatro campos:

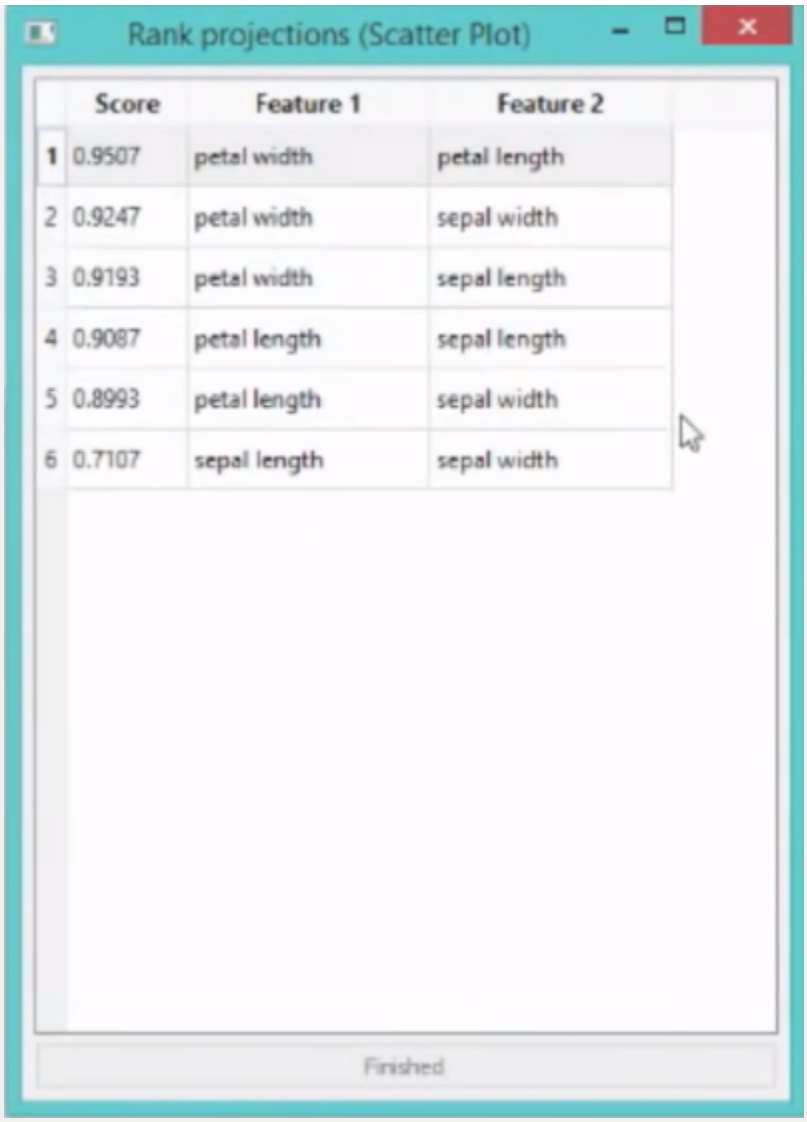


Hay tres tipos de flor: iris setosa, iris virgínica, iris versicolor

Si se conecta File a Distributions, puede verse el siguiente gráfico:

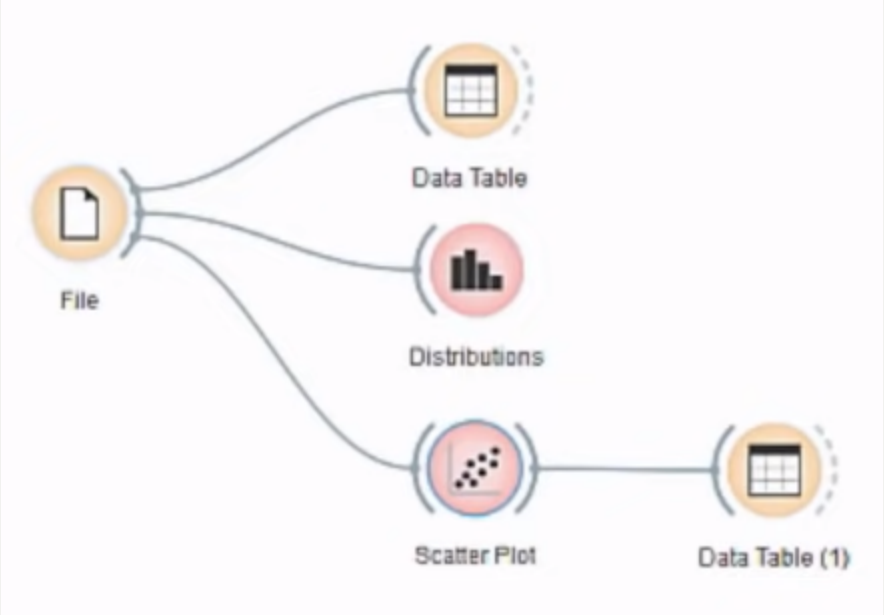


Al hacer un Scatter Plot puede clickearse en Rank Projections y observar el número que indica que a mayor sea, más define la separación entre los conjuntos:



El Scatter Plot envía automáticamente la info al output del widget

El workflow fue el siguiente:



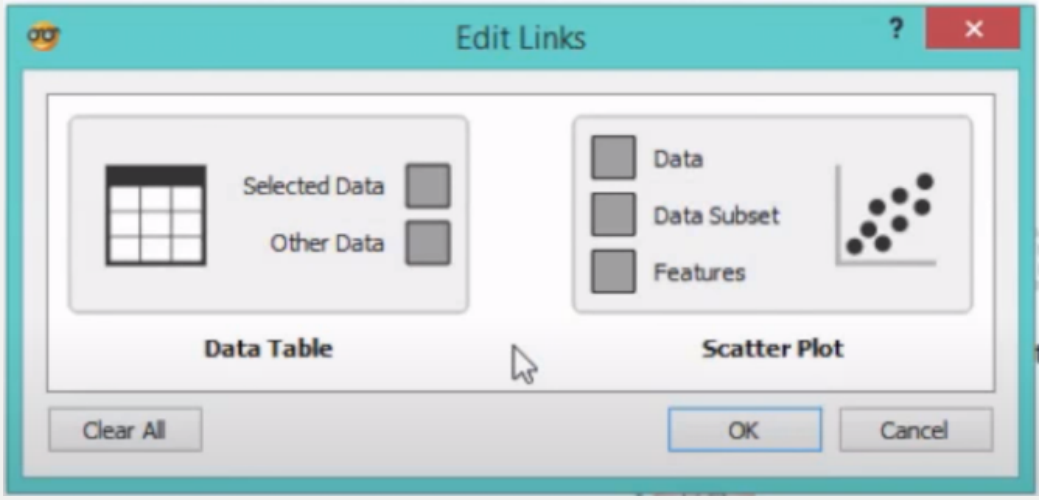
▼ 3. Widgets y canales

Hay múltiples formas de añadir un widget al workflow:

1. Clickear en el widget en el panel de widgets
2. Clickear y arrastrar a la página
3. Click derecho en la página y elegir el widget del menú
4. Arrastrar un canal de un output y soltar el mouse en un lugar vacío

Uno puede elegir qué datos pueden enviarse de un output a otro input

Al hacer doble click en la conexión de un widget a otro y puede verse qué información se está enviando



Así se puede definir cuál es la comunicación entre Widgets

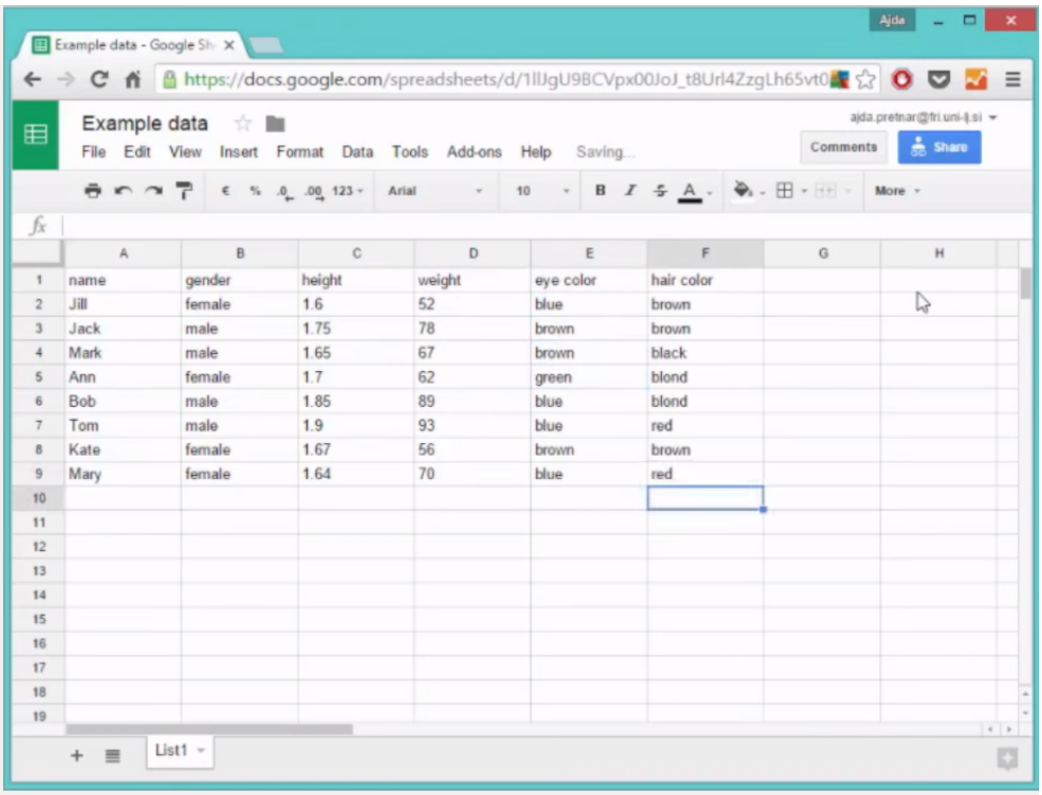
Con click derecho puede eliminarse una conexión entre Widgets

▼ 4. Loading your data

Orange puede leer Excel, tab y archivos separados por comas

Los datos normalmente son una tabla donde las instancias son renglones y las columnas atributos

Ejemplo con:



	A	B	C	D	E	F	G	H
1	name	gender	height	weight	eye color	hair color		
2	Jill	female	1.6	52	blue	brown		
3	Jack	male	1.75	78	brown	brown		
4	Mark	male	1.65	67	brown	black		
5	Ann	female	1.7	62	green	blond		
6	Bob	male	1.85	89	blue	blond		
7	Tom	male	1.9	93	blue	red		
8	Kate	female	1.67	56	brown	brown		
9	Mary	female	1.64	70	blue	red		
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								

El archivo se carga en el widget de File

Con el widget Select Columns se pueden corregir algunas cosas malinterpretadas por Orange al cargar automáticamente un archivo

Se pueden guardar los datos a la computadora con el widget Save. Es muy recomendable que se guarde con el formato nativo tab ya que prepara anotaciones para los atributos y otras cosas

Se pueden añadir dos renglones después de los nombres de atributos

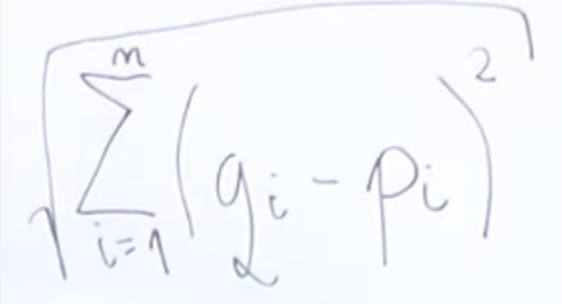
El primer renglón serían los Variable type (continuos o numéricos, discretos o categóricos, o strings) y variable kind (meta y class importantes)

▼ 5. Hierarchical Clustering

Uso del dataset de Iris

Se usará el método de Clustering Jerárquico

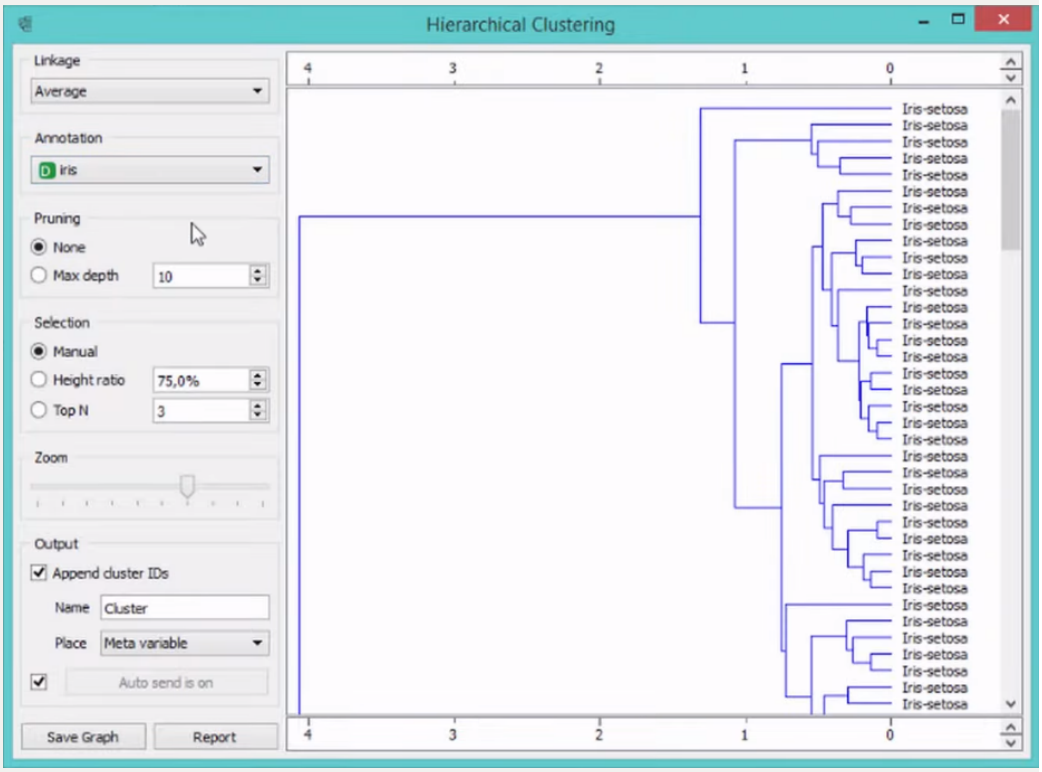
Explicación matemática del método



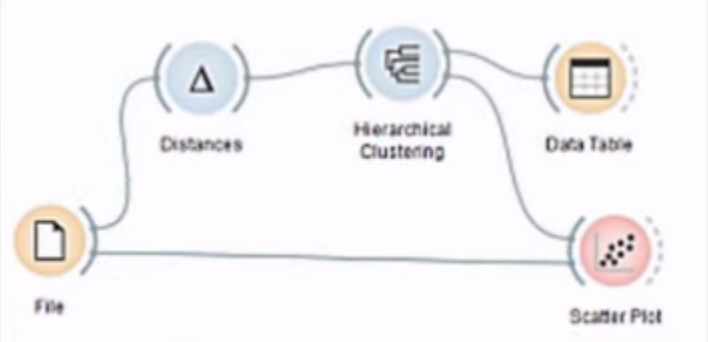
Básicamente es una fórmula de distancia euclidiana ya que “entre menor sean las distancias, mayor serán las similitudes”

Lo anterior sirve para clustering jerárquico

Se conecta entonces el widget File al widget Distances y éste al Hierarchical Clustering que mostrará un Dendrograma

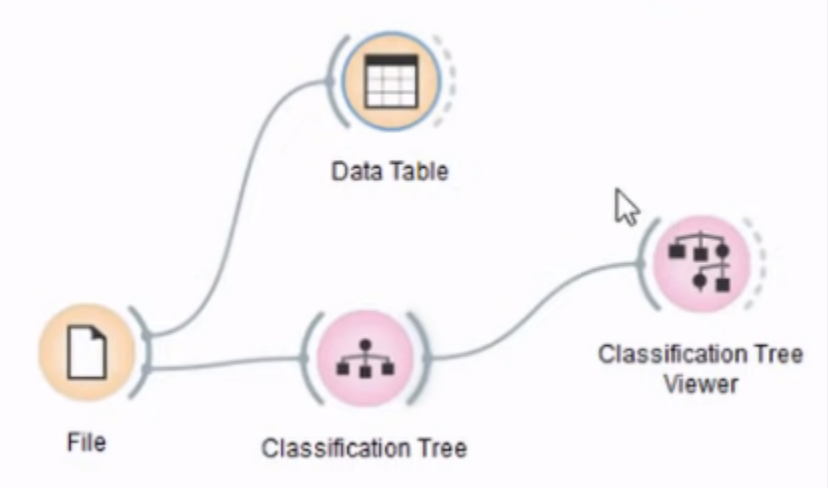


Puede enviarse esto a un Data Table para ver todo ordenado, pero también puede enviarse a un Scatter Plot conectado de la siguiente forma:

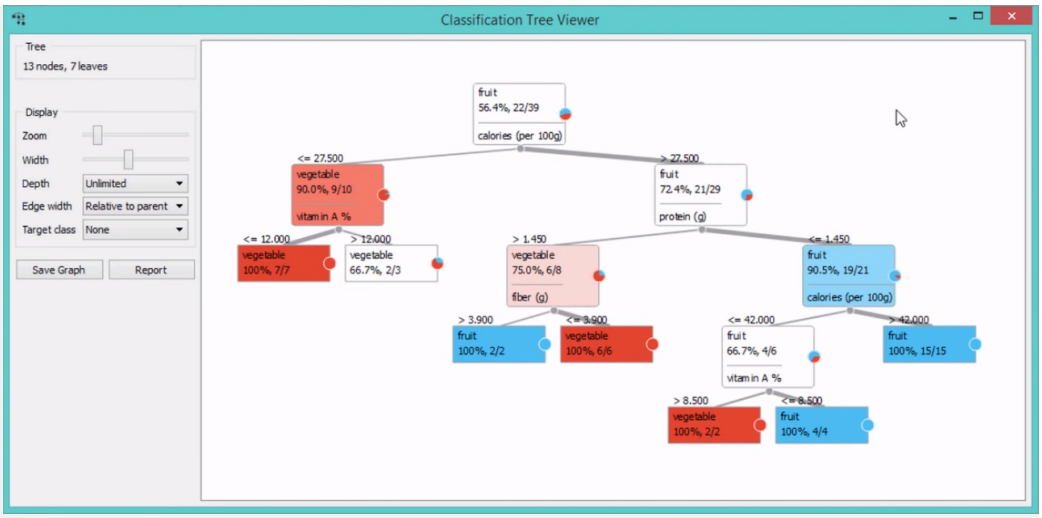


▼ 6. Making predictions

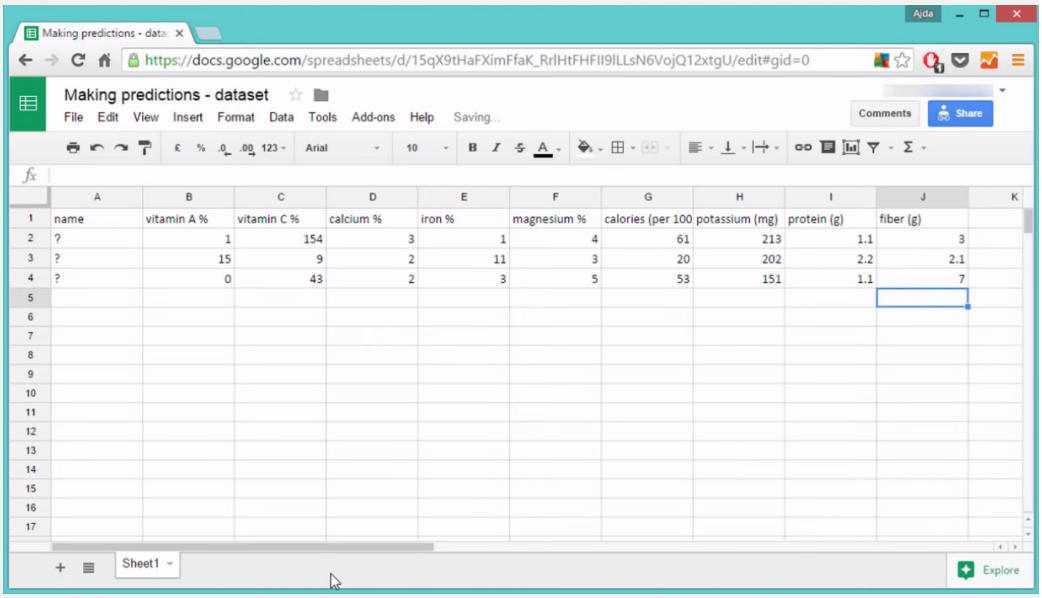
Se usará un dataset de frutas y vegetales para predecir si una instancia es fruta o verdura



Con el viewer puede verse qué atributo es más definitorio para las decisiones de nuestras predicciones:

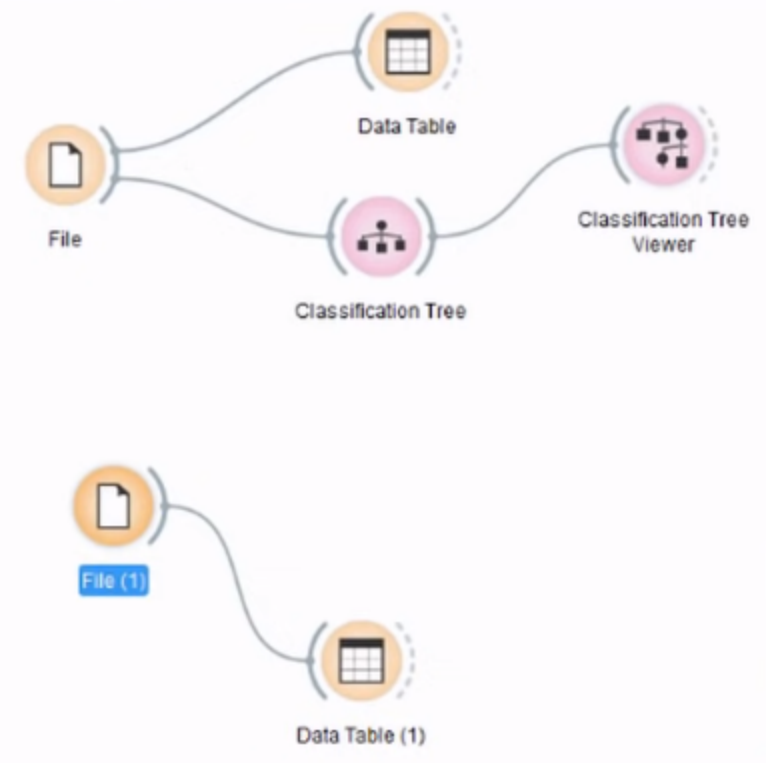


Se hace otra tabla donde tengan los mismos nombres de los atributos con un “?” para el name, dado que es la predicción que quiere hacerse

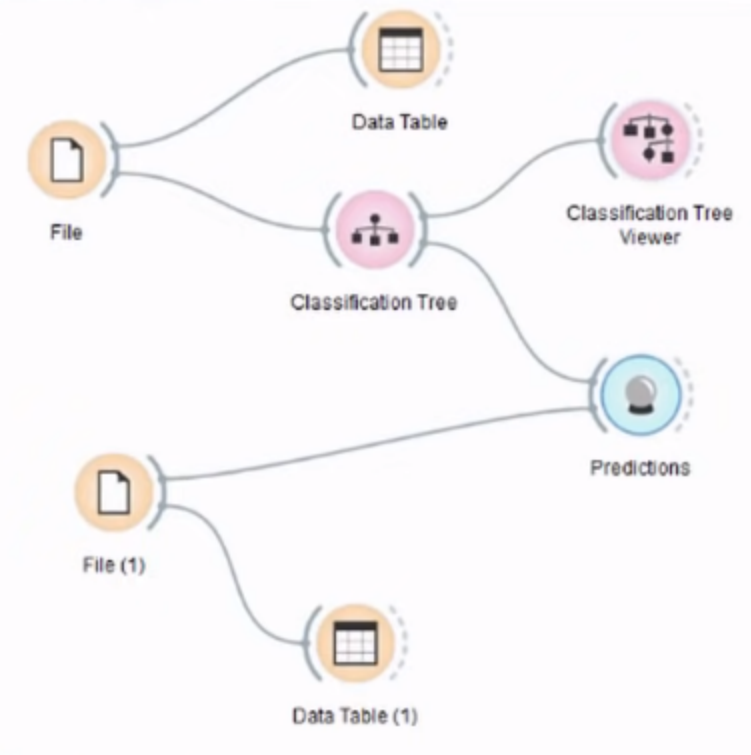


	A	B	C	D	E	F	G	H	I	J	K
1	name	vitamin A %	vitamin C %	calcium %	iron %	magnesium %	calories (per 100	potassium (mg)	protein (g)	fiber (g)	
2	?	1	154	3	1	4	61	213	1.1	3	
3	?	15	9	2	11	3	20	202	2.2	2.1	
4	?	0	43	2	3	5	53	151	1.1	7	
5											
6											
7											
8											
9											
10											
11											
12											
13											
14											
15											
16											
17											

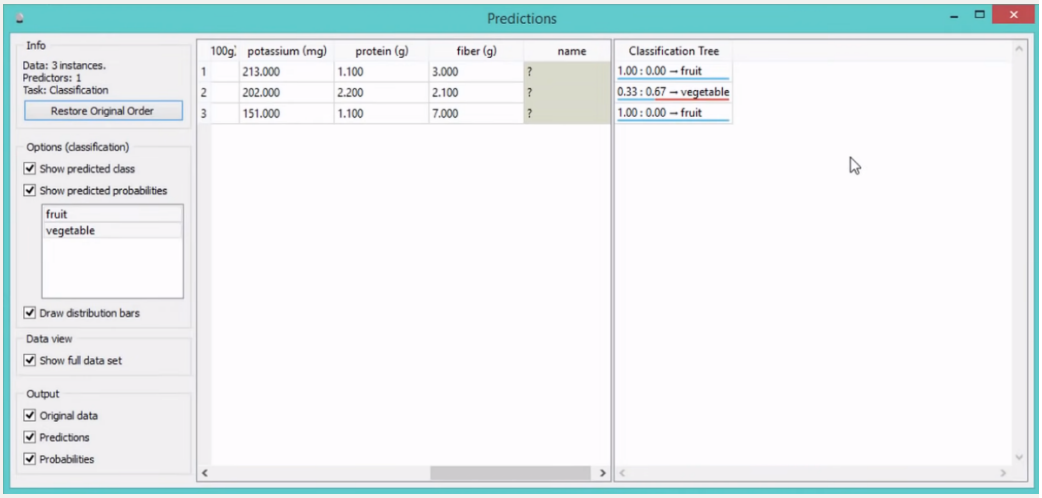
Se carga el archivo a Orange y se lee con un Data Table:



El File de abajo se debe conectar a Predictions y Classification Tree de la siguiente forma:

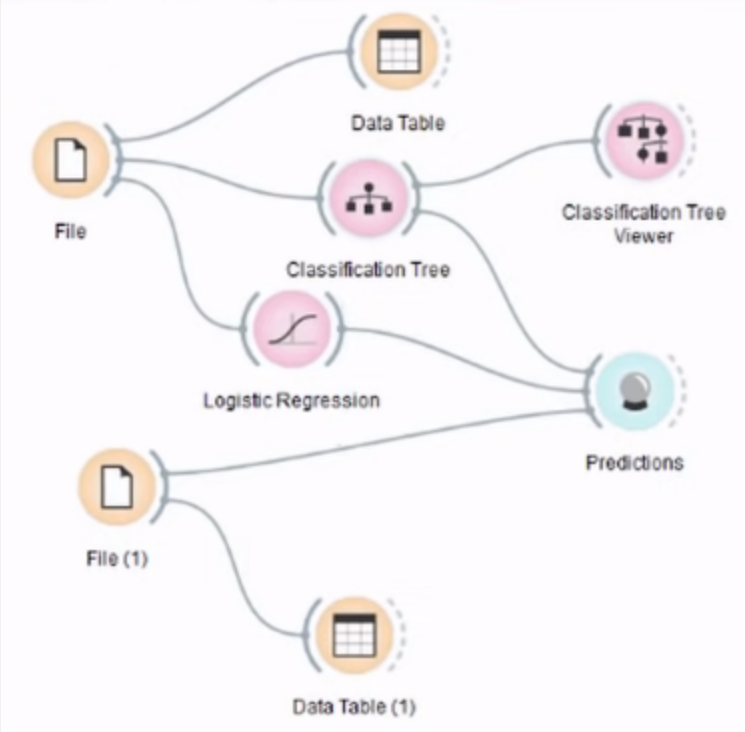


De esta forma puede ahora verse la predicción directamente en el widget:

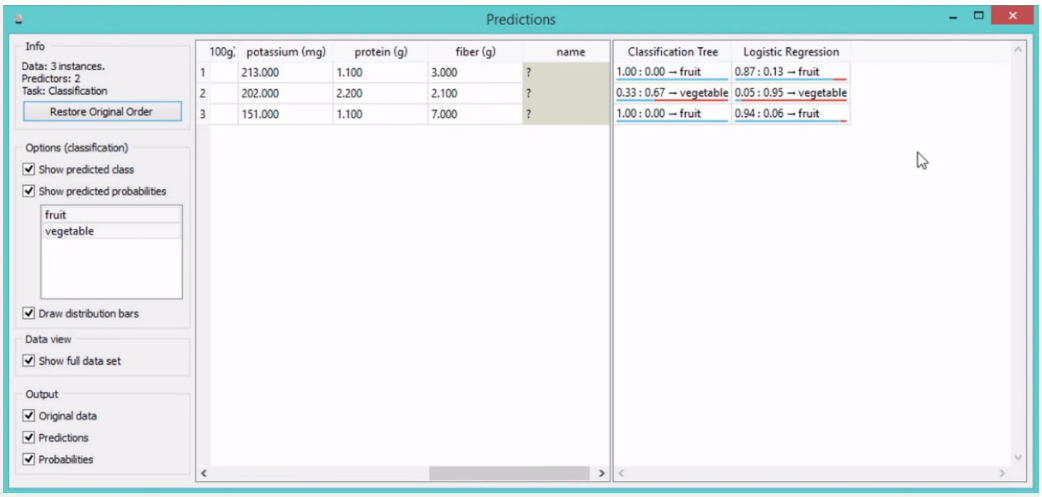


También pueden usarse otros clasificadores como regresión lógica

Se conecta de la siguiente forma:



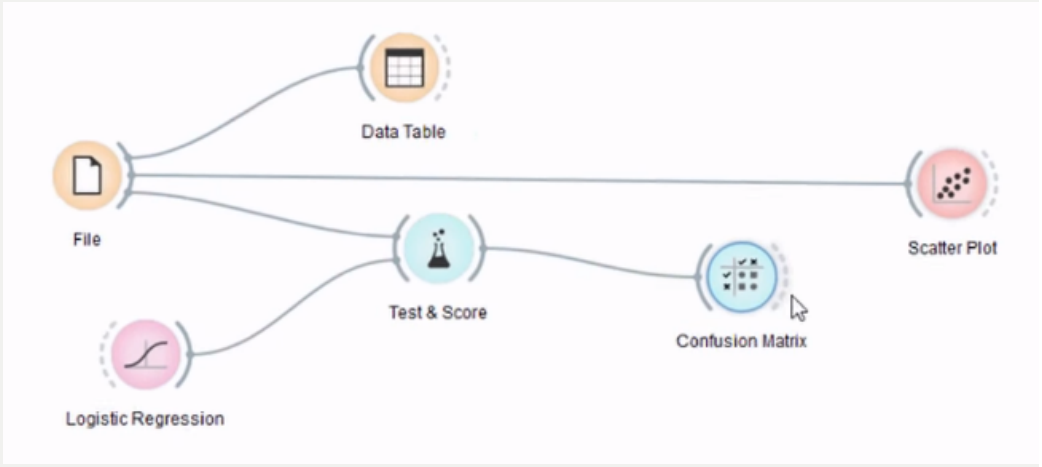
E incluso se verán ambas predicciones:



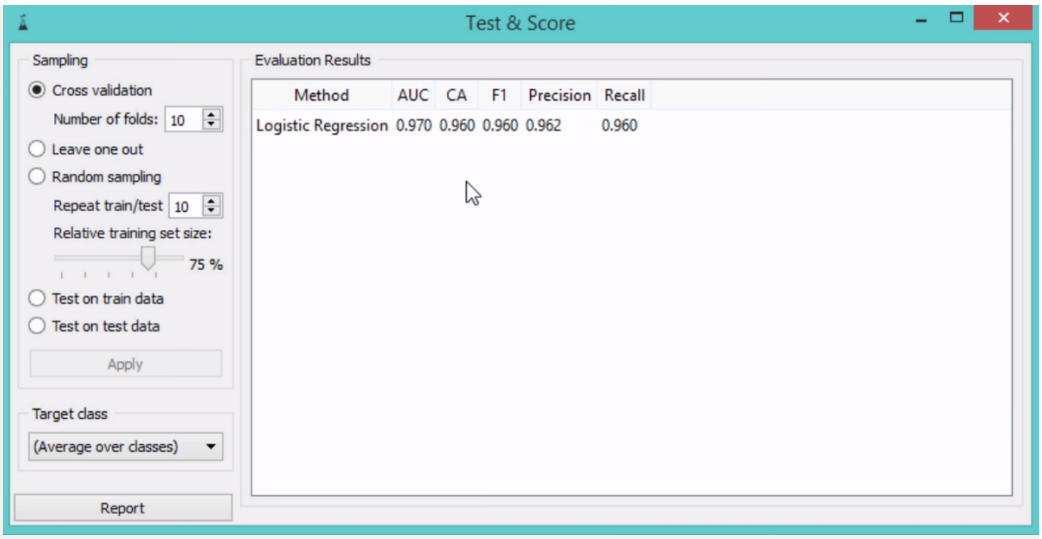
▼ 7. Model evaluation and scoring

¿Cuál método de clasificación funciona mejor?

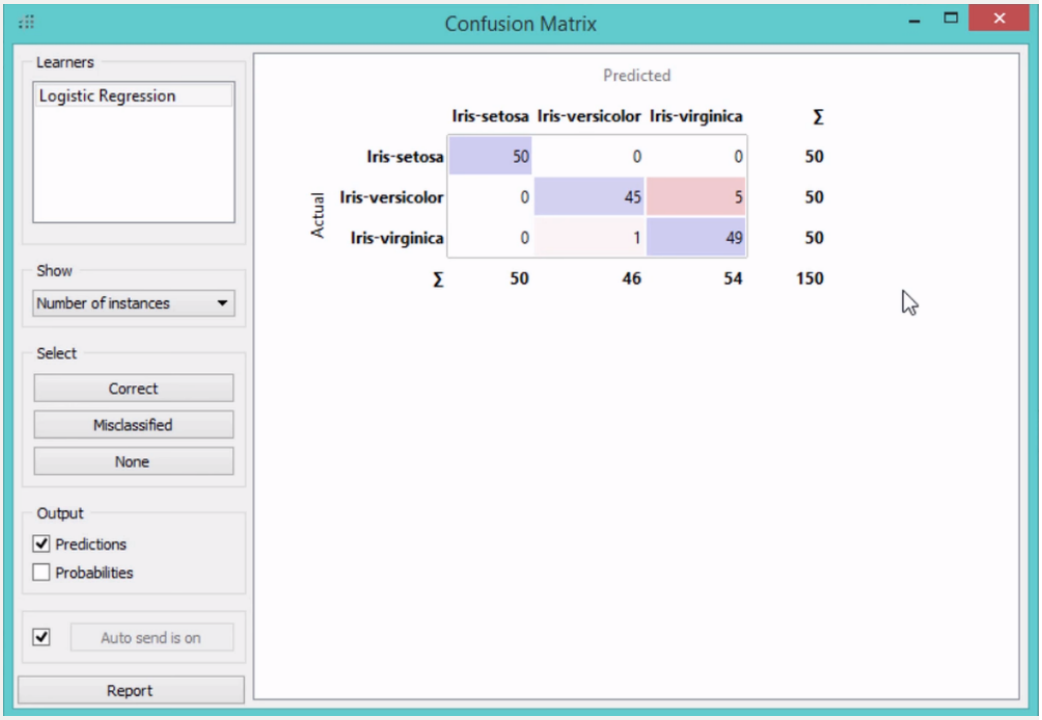
Se conectará la releción logística en otra rama para evitar overfitting:



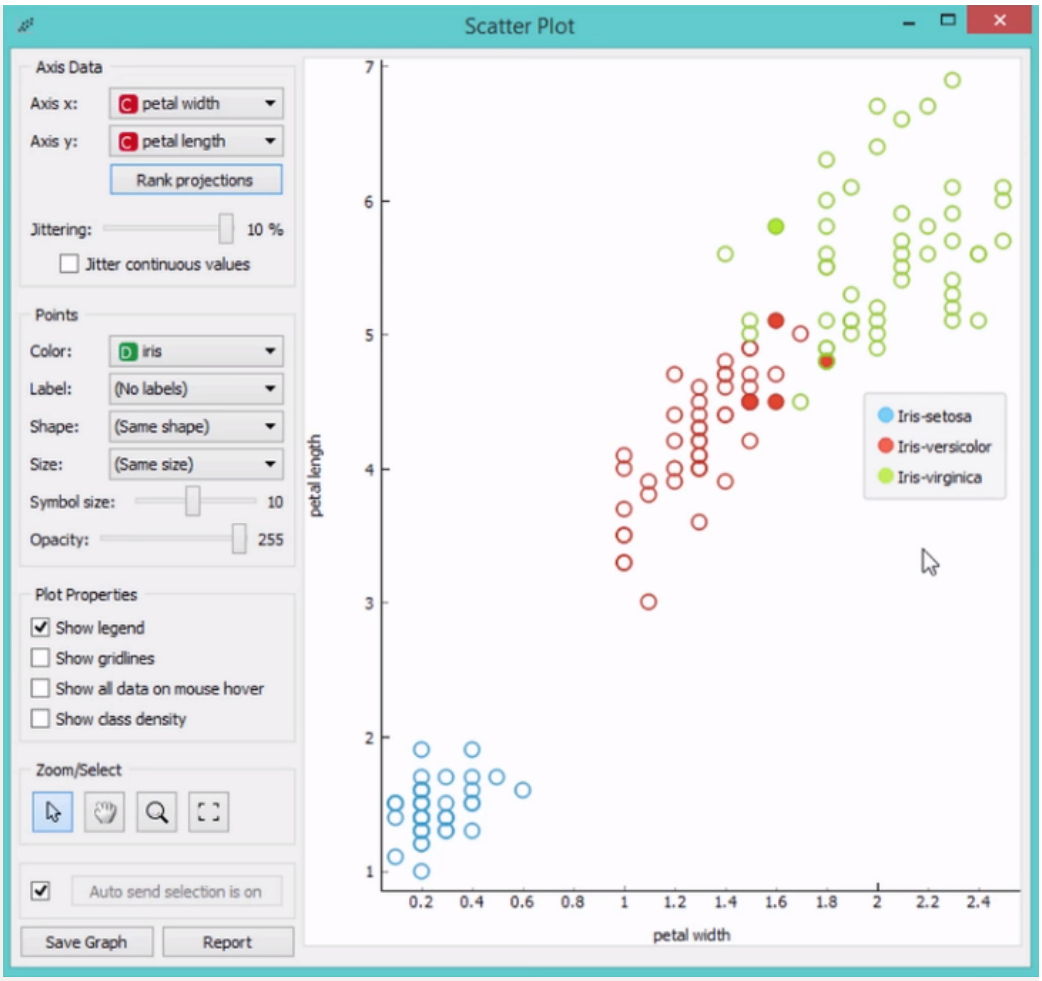
En Test & Score observamos lo siguiente:



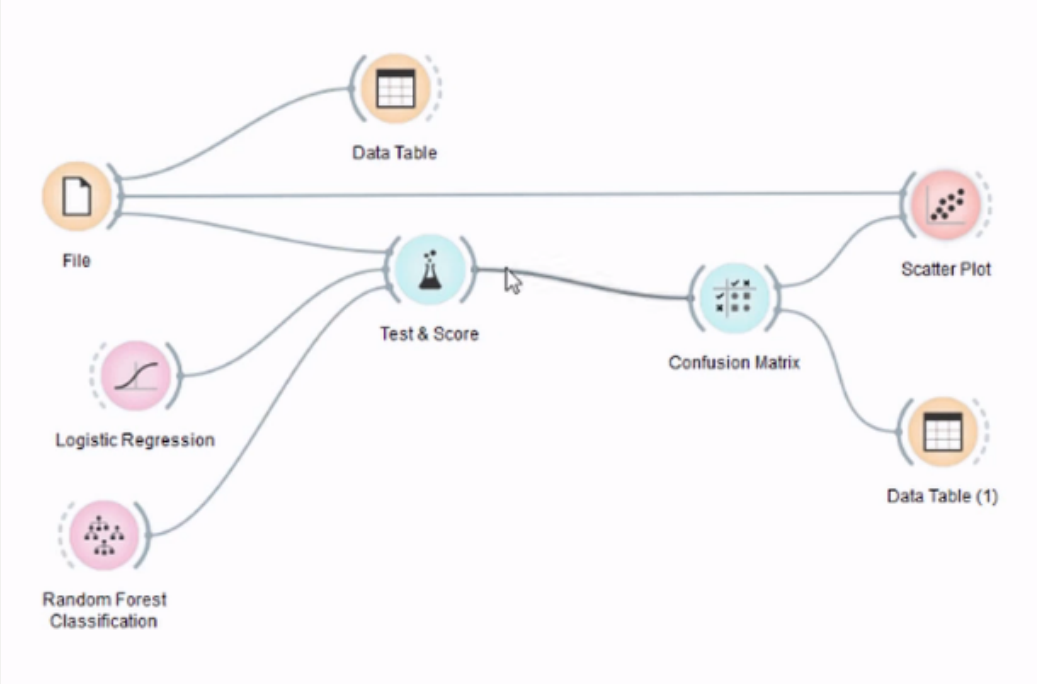
Se observa que CA (Classification Accuracy) sacó un 96% y puede observarse esto en la Matriz de Confusión:

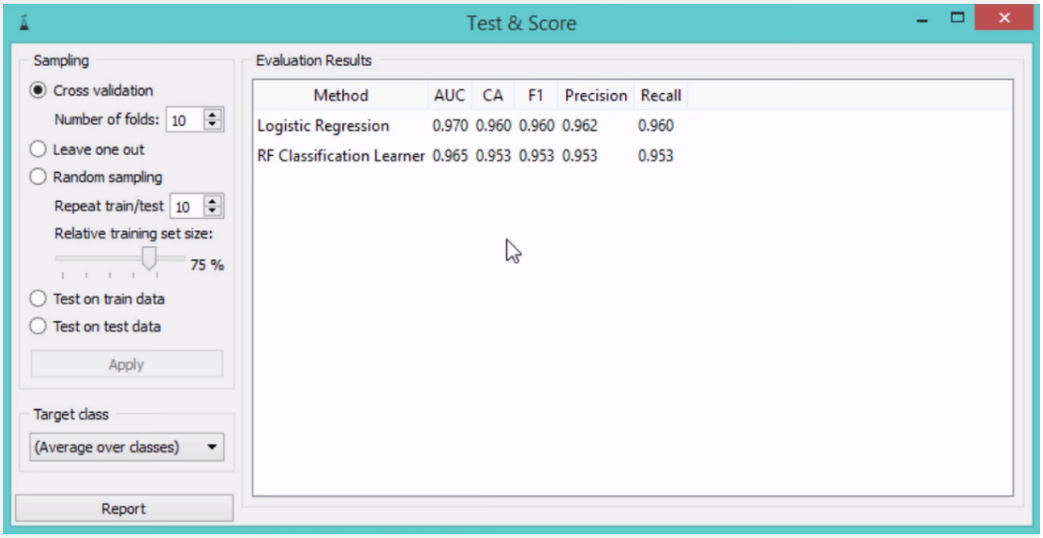


Si se conecta este widget de matriz de confusión con el widget que ya está de Scatter Plot, puede observarse que el error está en las flores muy cercanas:



Y como Logistic Regression no es la única que puede usarse, puede también usarse Random Forest:





AUC = Area Under receiver operating Characteristics

- ▼ **8. Add-ons**
- ▼ **9. Principal Component Analysis**
- ▼ **10. Feature Scoring and Ranking**
- ▼ **11. K-Means**
- ▼ **12. K-Means Explained**
- ▼ **13. Silhouette**

▼ **14. Image Analytics - Clustering**
▼ **15. Image Analytics - Classification**
▼ **16. Text Preprocessing**
▼ **17. Text Clustering**
▼ **18. Text Classification**
▼ **19. How to import text documents**
▼ **20. Multivariate Projection - Freeviz**

▼ **22-08-2022**

Revisión de la tarea

Revisión del KDD en la presentación

Como tarea buscar herramientas de ciencia de datos y mínimo buscar alguna para el proyecto (por equipos)

▼ **23-08-2022**

Nuevo tema: Limpieza de datos

Con base en la fuente:

https://elibro.net/es/ereader/uaa/71744?page=1

¿Qué es la limpieza de datos?

La anomalía es una propiedad de los valores de los datos a partir de la cual se obtiene una representación errónea del mini mundo que éstos reflejan. Esta puede ser originada por mediciones erróneas, entradas de datos sin validaciones, omisiones ocurridas mientras se coleccionan o se mantienen los datos o también puede ser el resultado de malas interpretaciones del análisis de los datos ó de cambios que se han producido en el mini mundo que no se han reflejado en la representación de los datos. Un tipo especial de anomalía es la redundancia, o sea, múltiples tuplas representan el mismo hecho, o partes de un hecho.

Las anomalías obstaculizan el uso efectivo y eficiente de los datos. Los datos que contienen anomalías son datos erróneos o sucios.

La anomalía es una propiedad de los valores de los datos a partir de la cual se obtiene una representación errónea del mini mundo que éstos reflejan. Esta puede ser originada por mediciones erróneas, entradas de datos sin validaciones, omisiones ocurridas mientras se coleccionan o se mantienen los datos o también puede ser el resultado de malas interpretaciones del análisis de los datos ó de cambios que se han producido en el mini mundo que no se han reflejado en la representación de los datos. Un tipo especial de anomalía es la redundancia, o sea, múltiples tuplas representan el mismo hecho, o partes de un hecho.

Por otra parte, en el área de KDD la **limpieza de datos** se define como el primer paso o preprocesamiento ([28],[29]). Varios sistemas de KDD y minería de datos resuelven las actividades de limpieza de datos con herramientas dependientes de dominios específicos.

▼ **24-08-2022**

Fuente del curso: Limpieza de datos

https://elibro.net/es/ereader/uaa/71744?page=1

Fuente del curso: Cómo usar Power BI

https://www.youtube.com/watch?v=pwJuFbyhZFE

▼ **25-08-2022**

No fui a clase

▼ **26-08-2022**

No hubo clase

▼ **29-08-2022**

Cosas para el proyecto:

1. Autocapacitarse para una herramienta libre como Weka, RapidMiner, R, Python, Orange, etc. [A fuerzas en Orange]
2. Detección de errores y propuesta de limpieza de datos
3. Usar una base de datos diferente a la de los demás equipos (Disney+ Movies en Kaggle)
4. Es posible que se pida el análisis de la base de datos de ICIs

Las cosas anteriores quedarían para dentro de dos semanas

▼ **30-08-2022**

Valores ausentes y su significación el los almacenes de datos

El proceso de limpieza de datos:

1. Deben ser detectados y corregidos los principales errores e inconsistencias, tanto para fuentes de datos simples o se integran datos de diversos orígenes
2. El enfoque debe estar basado en herramientas que limiten tanto la inspección manual de los datos como los esfuerzos de programación y debe permitir que puedan ser fácilmente cubiertos otros orígenes de datos
3. El proceso no debe resolverse aisladamente sino junto a las transformaciones de los datos relativas a los esquemas basándose éstas en los metadatos

En el examen se usará ya sea Orange, Power BI o Excel

Data warehouse

▼ **31-08-2022**

Práctica con Excel

▼ **01-09-2022**

Hablando del examen teórico - práctico

▼ **02-09-2022**

No hubo clase (se habló ayer sobre el examen. La info está en el grupo del Team Secuestro)

▼ Segundo Parcial

▼ **05-09-2022**

Viendo qué pez con el examen

▼ **06-09-2022**

No hubo clase

▼ **07-09-2022**

Calificaciones del primer parcial

▼ **08-09-2022**

No hubo clase

▼ **09-09-2022**

No hubo clase

▼ **12-09-2022**

No hubo clase

▼ **13-09-2022**

No hubo clase

▼ **14-09-2022**

No hubo clase

▼ **15-09-2022**

No hubo clase

▼ **19-09-2022**

No hubo clase

▼ **20-09-2022**

Vamos a programar en Python

Práctica de JupyterLab:

Bajar el archivo de Aula Virtual

▼ **21-09-2022**

Hacer examen refácil y reteórico de Python

▼ **22-09-2022**

Tenemos que subir la tarea completa

▼ **23-09-2022**

No hubo clase

▼ **26-09-2022**

Presentación estadística descriptiva

▼ **27-09-2022**

▼ **28-09-2022**

▼ **29-09-2022**

▼ **30-09-2022**

No hubo clase

▼ **03-10-2022**

No hubo clases por el congreso de ICI

▼ **04-10-2022**

No hubo clases por el congreso de ICI

▼ **05-10-2022**

No hubo clases por el congreso de ICI

▼ **06-10-2022**

No hubo clase

▼ **07-10-2022**

No hubo clase

▼ **10-10-2022**

No hubo clase

▼ **11-10-2022**

Requerimientos del examen

▼ **12-10-2022**

Preparación para las presentaciones

▼ **13-10-2022**

Preparación para las presentaciones

▼ **14-10-2022**

Preparación para las presentaciones

▼ **17-10-2022**

Preparación para las presentaciones

▼ **18-10-2022**

Preparación para las presentaciones

▼ **19-10-2022**

Preparación para las presentaciones

▼ **20-10-2022**

Preparación para las presentaciones

▼ **21-10-2022**

Preparación para las presentaciones

▼ **24-10-2022**

Presentamos nosotros

▼ **25-10-2022**

Preparación para las presentaciones

▼ **26-10-2022**

Preparación para las presentaciones

▼ **27-10-2022**

Preparación para las presentaciones

▼ **28-10-2022**

Preparación para las presentaciones

▼ Tercer Parcial

▼ **31-10-2022**

Alternativas para el tercer parcial:

1. Ver una sección del curso por integrante y hacer un proyecto con lo aprendido
2. Otra alternativa que no entendí si era más teórica o práctica

▼ **01-11-2022**

?

▼ **03-11-2022**

?

▼ **04-11-2022**

No hubo clases

▼ **07-11-2022**

El examen sería n integrantes, n secciones y dominio de esas secciones del curso

> Revisando tratando de entender con un dataset nuevo aplicando dichas técnicas

Contactar con el profe para liberarnos del tercer parcial (puede ser desde ahorita)

Exposición de Cristian

▼ **08-11-2022**

Exposciones

▼ **09-11-2022**

Exposciones

▼ **10-11-2022**

Exposciones

▼ **11-11-2022**

Exposciones

▼ **14-11-2022**

No entré a clase

▼ **15-11-2022**

No entré a clase

▼ **16-11-2022**

No entré a clase

▼ **17-11-2022**

No entré a clase

▼ **18-11-2022**

No entré a clase

▼ **22-11-2022**

No entré a clases

▼ **23-11-2022**

No entré a clases

▼ **24-11-2022**

No entré a clases

▼ **25-11-2022**

No entré a clases