ACÁMICA

TEMA DEL DÍA

Árboles de decisión y KNN

Dos modelos predictivos de **clasificación** y **aprendizaje supervisado**. ¡Es una meeting con mucho contenido nuevo!

Agenda

Daily

Explicación: Árboles de decisión

Break

Explicación: KNN

Hands-on training

Cierre

Daily

Daily

Sincronizando...

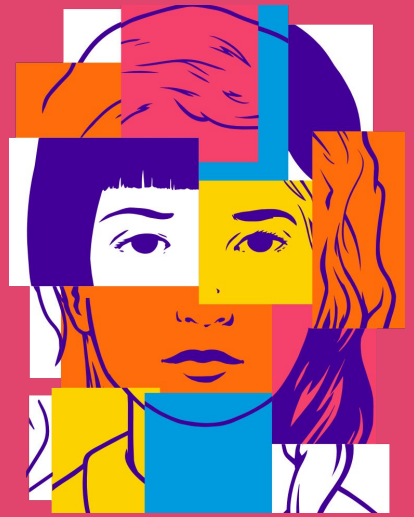
Toolbox

¿Cómo te ha ido?
¿Obstáculos?
¿Cómo seguimos?

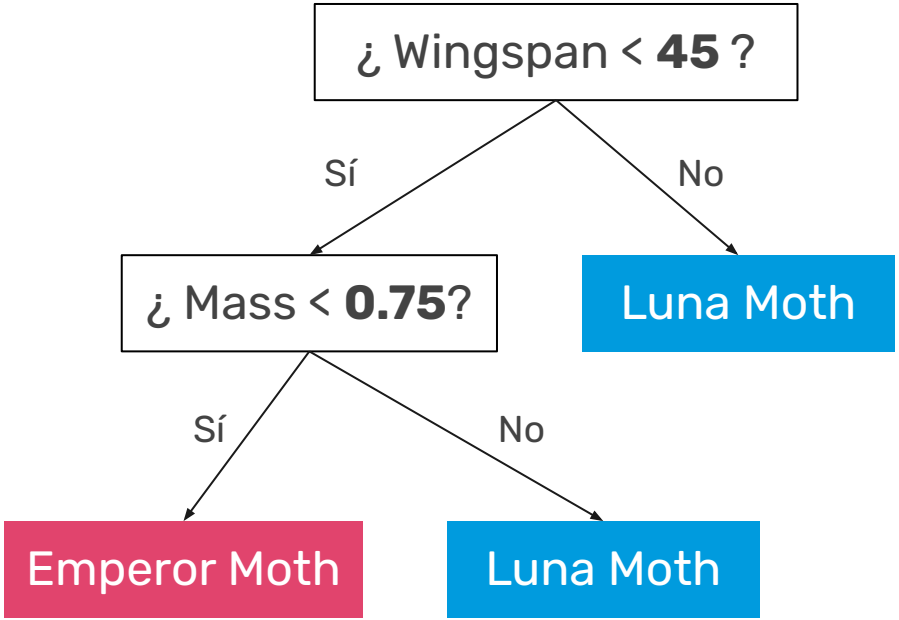
Challenge

¿Cómo te ha ido?
¿Obstáculos?
¿Cómo seguimos?

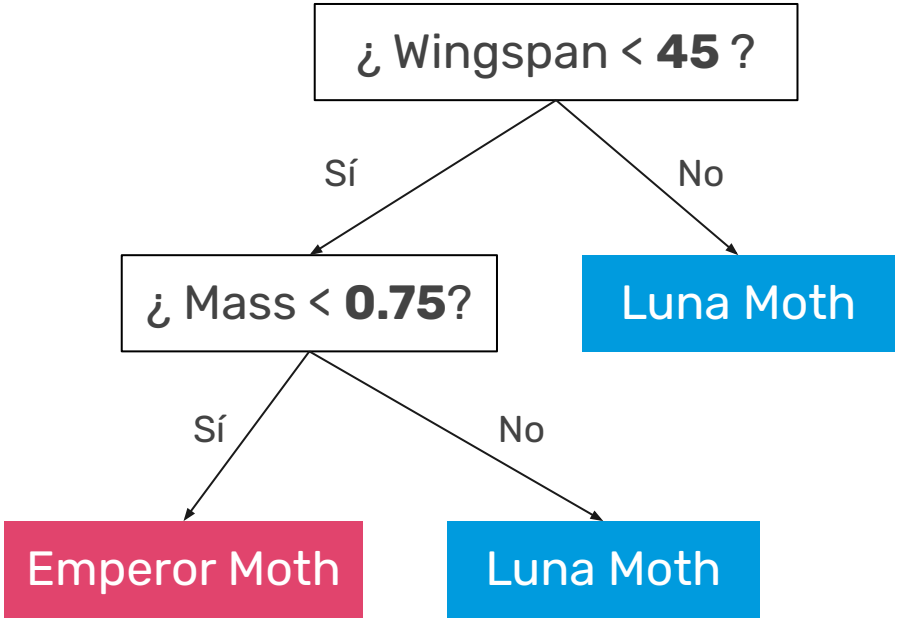
Árboles de Decisión



Un árbol de decisión “hace preguntas” y va clasificando de acuerdo a las respuestas.

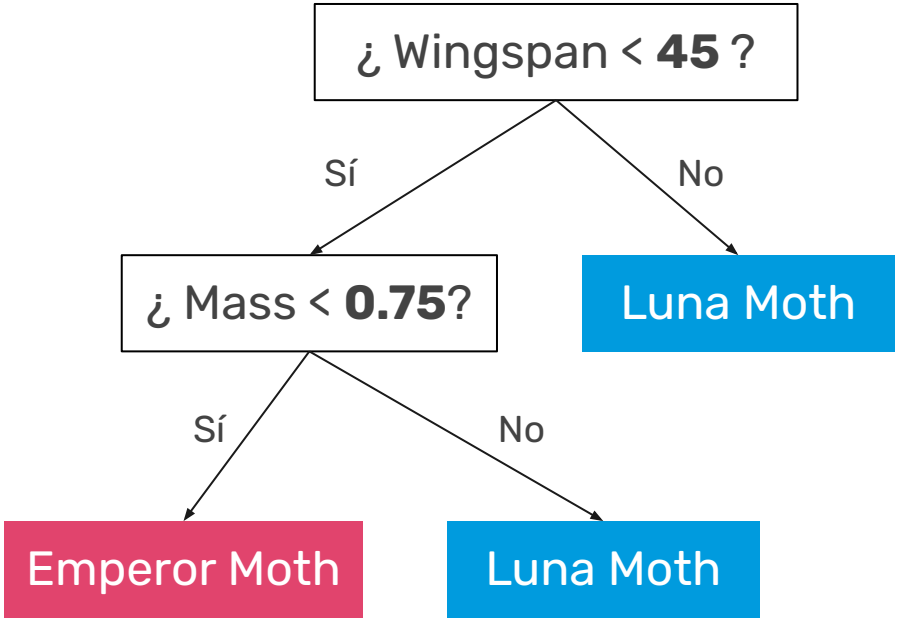


Un árbol de decisión “hace preguntas” y va clasificando de acuerdo a las respuestas.



¿Cómo decide qué preguntar?

Un árbol de decisión “hace preguntas” y va clasificando de acuerdo a las respuestas.

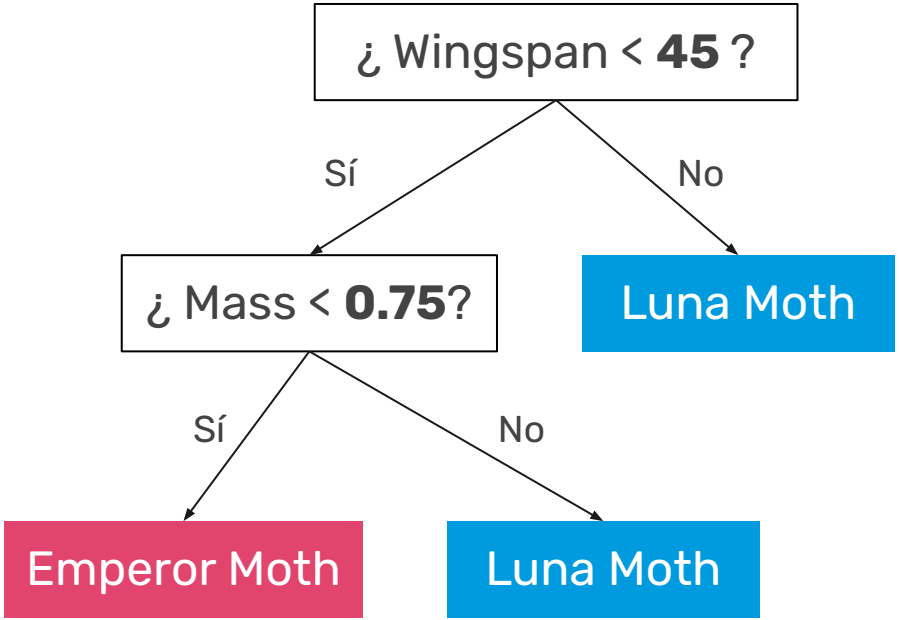


¿Cómo decide qué preguntar?

1. Impureza Gini
2. Entropía/Ganancia de información

Son cálculos que se hacen sobre los datos que ayudan a descubrir cuán bueno es un feature para separar las instancias por sus etiquetas.

Un árbol de decisión “hace preguntas” y va clasificando de acuerdo a las respuestas.



¿Cómo decide qué preguntar?

1. **Impureza Gini**
2. Entropía/Ganancia de información

Son cálculos que se hacen sobre los datos que ayudan a descubrir cuán bueno es un feature para separar las instancias por sus etiquetas.

Impureza Gini

Supongamos que tenemos este dataset para el ejemplo de las polillas del video (muy simplificado).

¿Cuál será un mejor atributo para "preguntar"?

Pero… ¿qué es un mejor atributo?

Intuitivamente, un mejor atributo será el que separe "mejor" las clases.

que las muestras obtenidas sean lo más "puras" posibles. Es decir, tengan instancias de una sola de las clases.

Masa	Envergadura	Tipo de polilla
Mayor a 0.75 gr	Mayor a 45 mm	Luna Moth
Menor a 0.75 gr	Mayor a 45 mm	Luna Moth
Mayor a 0.75 gr	Menor a 45 mm	Luna Moth
Menor a 0.75 gr	Mayor a 45 mm	Luna Moth
Menor a 0.75 gr	Menor a 45 mm	Emperor Moth
Menor a 0.75 gr	Mayor a 45 mm	Luna Moth
Menor a 0.75 gr	Menor a 45 mm	Emperor Moth
Mayor a 0.75 gr	Mayor a 45 mm	Luna Moth
Menor a 0.75 gr	Menor a 45 mm	Emperor Moth
Menor a 0.75 gr	Menor a 45 mm	Emperor Moth

Impureza Gini

A simple vista, es muy difícil determinar cuál atributo es mejor para separar clases, y eso que sólo tenemos diez instancias, dos atributos y solamente dos valores por atributo.

Para hacerlo eficientemente, necesitamos algún estadístico que cuantifique la pureza de las muestras.

Para eso existe la **Impureza Gini.**

Masa	Envergadura	Tipo de polilla
Mayor a 0.75 gr	Mayor a 45 mm	Luna Moth
Menor a 0.75 gr	Mayor a 45 mm	Luna Moth
Mayor a 0.75 gr	Menor a 45 mm	Luna Moth
Menor a 0.75 gr	Mayor a 45 mm	Luna Moth
Menor a 0.75 gr	Menor a 45 mm	Emperor Moth
Menor a 0.75 gr	Mayor a 45 mm	Luna Moth
Menor a 0.75 gr	Menor a 45 mm	Emperor Moth
Mayor a 0.75 gr	Mayor a 45 mm	Luna Moth
Menor a 0.75 gr	Menor a 45 mm	Emperor Moth
Menor a 0.75 gr	Menor a 45 mm	Emperor Moth

Impureza Gini: ¿Cómo funciona?

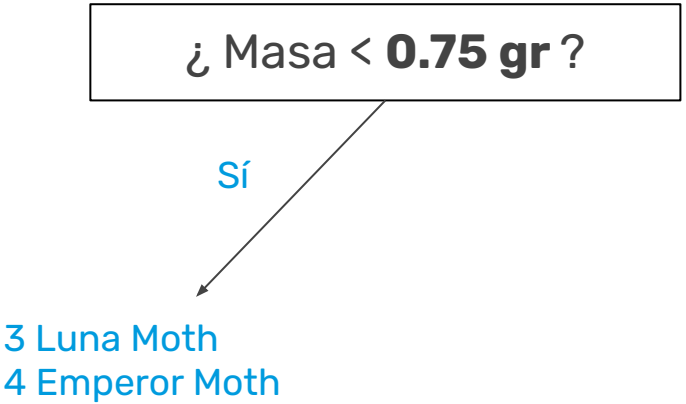
Probemos construyendo una *pregunta* por cada feature y veamos cual deja mejor separadas las instancias.

¿ Masa < **0.75 gr** ?

Masa	Envergadura	Tipo de polilla
Mayor a 0.75 gr	Mayor a 45 mm	Luna Moth
Menor a 0.75 gr	Mayor a 45 mm	Luna Moth
Mayor a 0.75 gr	Menor a 45 mm	Luna Moth
Menor a 0.75 gr	Mayor a 45 mm	Luna Moth
Menor a 0.75 gr	Menor a 45 mm	Emperor Moth
Menor a 0.75 gr	Mayor a 45 mm	Luna Moth
Menor a 0.75 gr	Menor a 45 mm	Emperor Moth
Mayor a 0.75 gr	Mayor a 45 mm	Luna Moth
Menor a 0.75 gr	Menor a 45 mm	Emperor Moth
Menor a 0.75 gr	Menor a 45 mm	Emperor Moth

Impureza Gini: ¿Cómo funciona?

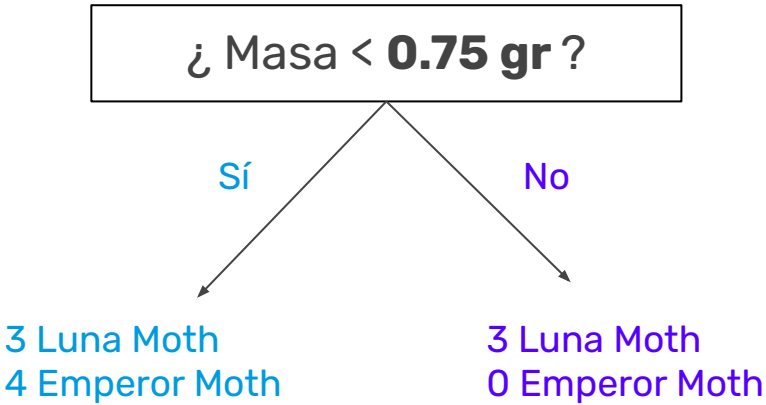
Probemos construyendo una *pregunta* por cada feature y veamos cual deja mejor separadas las instancias.



Masa	Envergadura	Tipo de polilla
Mayor a 0.75 gr	Mayor a 45 mm	Luna Moth
Menor a 0.75 gr	Mayor a 45 mm	Luna Moth
Mayor a 0.75 gr	Menor a 45 mm	Luna Moth
Menor a 0.75 gr	Mayor a 45 mm	Luna Moth
Menor a 0.75 gr	Menor a 45 mm	Emperor Moth
Menor a 0.75 gr	Mayor a 45 mm	Luna Moth
Menor a 0.75 gr	Menor a 45 mm	Emperor Moth
Mayor a 0.75 gr	Mayor a 45 mm	Luna Moth
Menor a 0.75 gr	Menor a 45 mm	Emperor Moth
Menor a 0.75 gr	Menor a 45 mm	Emperor Moth

Impureza Gini: ¿Cómo funciona?

Probemos construyendo una *pregunta* por cada feature y veamos cual deja mejor separadas las instancias.



Masa	Envergadura	Tipo de polilla
Mayor a 0.75 gr	Mayor a 45 mm	Luna Moth
Menor a 0.75 gr	Mayor a 45 mm	Luna Moth
Mayor a 0.75 gr	Menor a 45 mm	Luna Moth
Menor a 0.75 gr	Mayor a 45 mm	Luna Moth
Menor a 0.75 gr	Menor a 45 mm	Emperor Moth
Menor a 0.75 gr	Mayor a 45 mm	Luna Moth
Menor a 0.75 gr	Menor a 45 mm	Emperor Moth
Mayor a 0.75 gr	Mayor a 45 mm	Luna Moth
Menor a 0.75 gr	Menor a 45 mm	Emperor Moth
Menor a 0.75 gr	Menor a 45 mm	Emperor Moth

Impureza Gini: ¿Cómo funciona?

Probemos construyendo una *pregunta* por cada feature y veamos cual deja mejor separadas las instancias.

¿ Envergadura < **45 mm** ?

Masa	Envergadura	Tipo de polilla
Mayor a 0.75 gr	Mayor a 45 mm	Luna Moth
Menor a 0.75 gr	Mayor a 45 mm	Luna Moth
Mayor a 0.75 gr	Menor a 45 mm	Luna Moth
Menor a 0.75 gr	Mayor a 45 mm	Luna Moth
Menor a 0.75 gr	Menor a 45 mm	Emperor Moth
Menor a 0.75 gr	Mayor a 45 mm	Luna Moth
Menor a 0.75 gr	Menor a 45 mm	Emperor Moth
Mayor a 0.75 gr	Mayor a 45 mm	Luna Moth
Menor a 0.75 gr	Menor a 45 mm	Emperor Moth
Menor a 0.75 gr	Menor a 45 mm	Emperor Moth

Impureza Gini: ¿Cómo funciona?

Probemos construyendo una *pregunta* por cada feature y veamos cual deja mejor separadas las instancias.

¿ Envergadura < **45 mm** ?

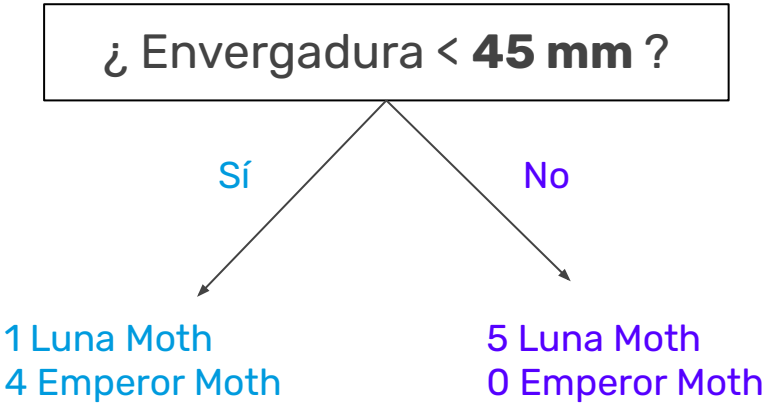
Sí

1 Luna Moth
4 Emperor Moth

Masa	Envergadura	Tipo de polilla
Mayor a 0.75 gr	Mayor a 45 mm	Luna Moth
Menor a 0.75 gr	Mayor a 45 mm	Luna Moth
Mayor a 0.75 gr	**Menor a 45 mm**	Luna Moth
Menor a 0.75 gr	Mayor a 45 mm	Luna Moth
Menor a 0.75 gr	**Menor a 45 mm**	Emperor Moth
Menor a 0.75 gr	Mayor a 45 mm	Luna Moth
Menor a 0.75 gr	**Menor a 45 mm**	Emperor Moth
Mayor a 0.75 gr	Mayor a 45 mm	Luna Moth
Menor a 0.75 gr	**Menor a 45 mm**	Emperor Moth
Menor a 0.75 gr	**Menor a 45 mm**	Emperor Moth

Impureza Gini: ¿Cómo funciona?

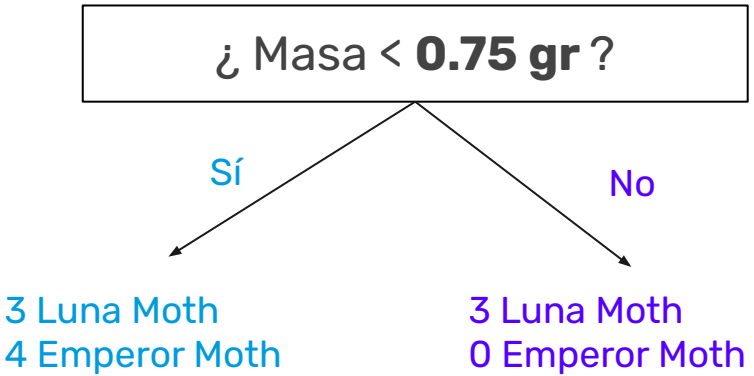
Probemos construyendo una *pregunta* por cada feature y veamos cual deja mejor separadas las instancias.

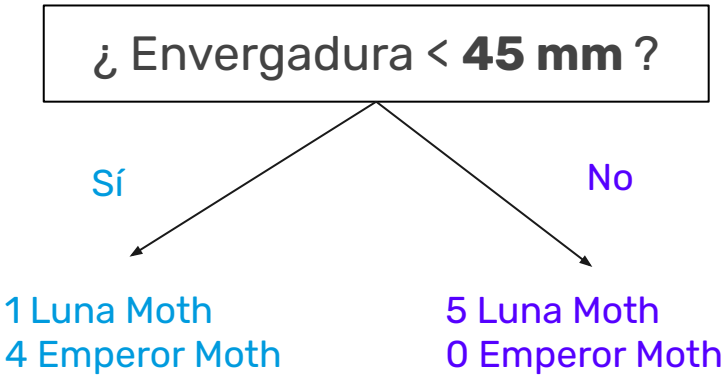


Masa	Envergadura	Tipo de polilla
Mayor a 0.75 gr	**Mayor a 45 mm**	Luna Moth
Menor a 0.75 gr	**Mayor a 45 mm**	Luna Moth
Mayor a 0.75 gr	Menor a 45 mm	Luna Moth
Menor a 0.75 gr	**Mayor a 45 mm**	Luna Moth
Menor a 0.75 gr	Menor a 45 mm	Emperor Moth
Menor a 0.75 gr	**Mayor a 45 mm**	Luna Moth
Menor a 0.75 gr	Menor a 45 mm	Emperor Moth
Mayor a 0.75 gr	**Mayor a 45 mm**	Luna Moth
Menor a 0.75 gr	Menor a 45 mm	Emperor Moth
Menor a 0.75 gr	Menor a 45 mm	Emperor Moth

Impureza Gini

¿Cuál de las dos preguntas *separó* mejor las clases?





Tratemos de cuantificarlo...

Impureza Gini

1. Calculamos la **Impureza Gini inicial** de la muestra.
2. Calculamos la **Impureza Gini** luego de hacer cada pregunta. Para ellos, hacemos un **promedio ponderado** de las impurezas resultantes en cada **hoja**, por pregunta.
3. Elegimos el atributo con **mayor reducción de impureza** (Ganancia Gini).
4. Si consideramos que las instancias ya están clasificadas suficientemente bien, FIN. Si no, seguimos construyendo el árbol de forma iterativa, tomando como muestra inicial la muestra de cada hoja y realizando los pasos 1 - 4.

Impureza Gini

1. Calculamos la **Impureza Gini inicial** de la muestra.

$$Gini_{inicial} = 1 - (\text{proporción de Luna Moth})^2 - (\text{proporción de Emperor Moth})^2$$

Impureza Gini

1. Calculamos la **Impureza Gini inicial** de la muestra.

$$Gini_{inicial} = 1 - (proporción\ de\ Luna\ Moth)^2 - (proporción\ de\ Emperor\ Moth)^2$$

Como son diez instancias, (6 Luna Moth y 4 Emperor Moth), entonces:

$$Gini_{inicial} = 1 - (6/10)^2 - (4/10)^2 = 0.48$$

Impureza Gini

1. Calculamos la **Impureza Gini inicial** de la muestra.

$$Gini_{inicial} = 1 - (\text{proporción de Luna Moth})^2 - (\text{proporción de Emperor Moth})^2$$

Como son diez instancias, (6 Luna Moth y 4 Emperor Moth), entonces:

$$Gini_{inicial} = 1 - (6/10)^2 - (4/10)^2 = 0.48$$

***Si la muestra tiene solamente miembros de una clase, entonces**

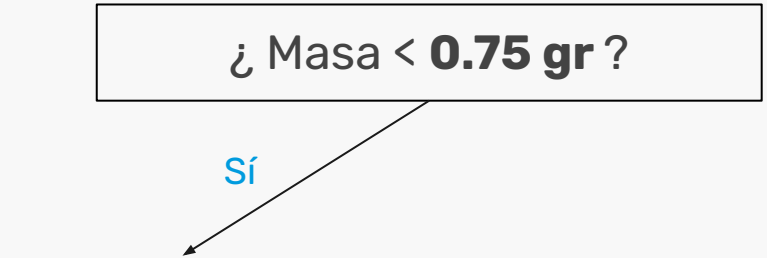
$$Gini = 1 - (\text{proporción única clase})^2 = 0$$

***y si tiene mitad y mitad:**

$$Gini = 1 - (1/2)^2 - (1/2)^2 = 0.5$$

Impureza Gini

2. Calculamos la **Impureza Gini** luego de hacer cada pregunta. Para ellos, hacemos un **promedio ponderado** de las impurezas resultantes en cada **hoja**, por pregunta.

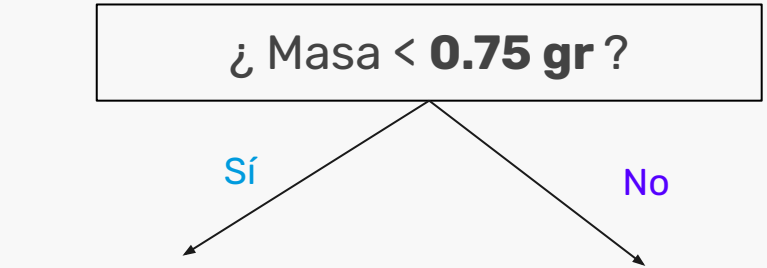


3 Luna Moth
4 Emperor Moth

Gini para esta **hoja:**
$= 1 - (3/7)^2 - (4/7)^2$
$= 0.4898$

Impureza Gini

2. Calculamos la **Impureza Gini** luego de hacer cada pregunta. Para ellos, hacemos un **promedio ponderado** de las impurezas resultantes en cada **hoja**, por pregunta.

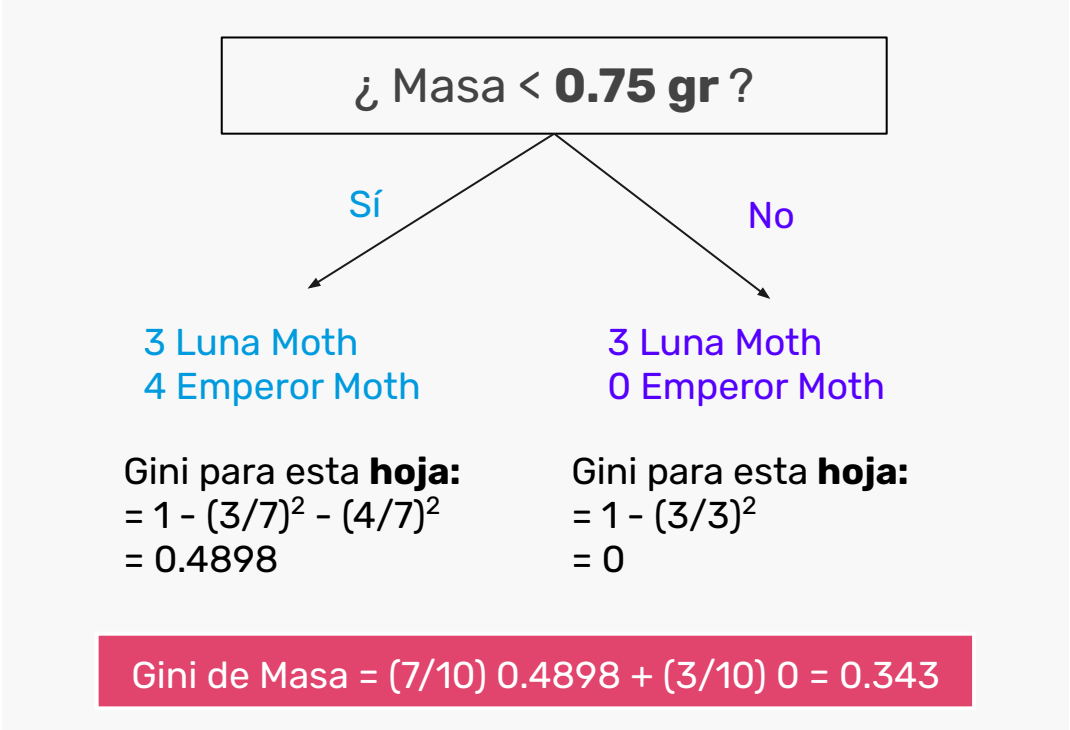


Gini para esta **hoja:**
$= 1 - (3/7)^2 - (4/7)^2$
$= 0.4898$

Gini para esta **hoja:**
$= 1 - (3/3)^2$
$= 0$

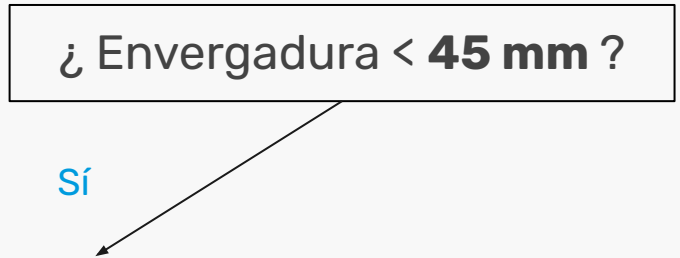
Impureza Gini

2. Calculamos la **Impureza Gini** luego de hacer cada pregunta. Para ellos, hacemos un **promedio ponderado** de las impurezas resultantes en cada **hoja**, por pregunta.



Impureza Gini

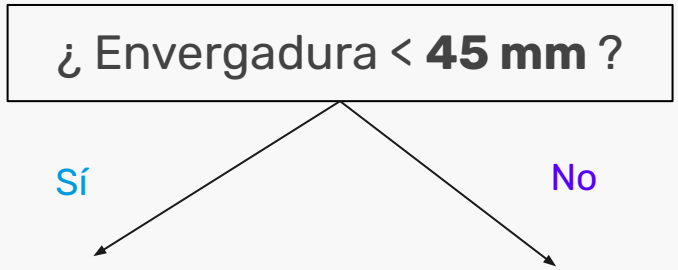
2. Calculamos la **Impureza Gini** luego de hacer cada pregunta. Para ellos, hacemos un **promedio ponderado** de las impurezas resultantes en cada **hoja**, por pregunta.



Gini para esta **hoja:**
$= 1 - (1/5)^2 - (4/5)^2$
$= 0.32$

Impureza Gini

2. Calculamos la **Impureza Gini** luego de hacer cada pregunta. Para ellos, hacemos un **promedio ponderado** de las impurezas resultantes en cada **hoja**, por pregunta.

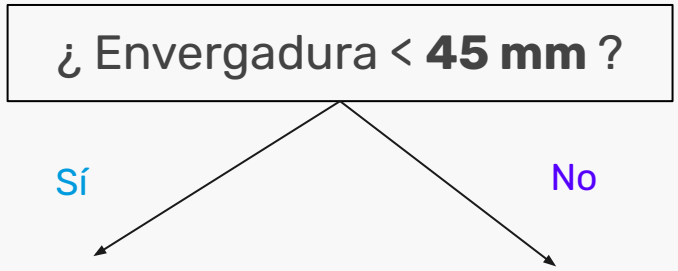


Gini para esta **hoja:**
$= 1 - (1/5)^2 - (4/5)^2$
$= 0.32$

Gini para esta **hoja:**
$= 1 - (5/5)^2$
$= 0$

Impureza Gini

2. Calculamos la **Impureza Gini** luego de hacer cada pregunta. Para ellos, hacemos un **promedio ponderado** de las impurezas resultantes en cada **hoja**, por pregunta.



Gini para esta **hoja:**
$= 1 - (1/5)^2 - (4/5)^2$
$= 0.32$

Gini para esta **hoja:**
$= 1 - (5/5)^2$
$= 0$

Gini de Env = (5/10) 0.32 + (5/10) 0 = 0.16

Impureza Gini

3. Elegimos el atributo con **mayor reducción de impureza** (Ganancia Gini).

- Masa: 0.48 - 0.343 = 0.137

- Envergadura: 0.48 - 0.16 = 0.32

Impureza Gini

3. Elegimos el atributo con **mayor reducción de impureza** (Ganancia Gini).

- Masa: 0.48 - 0.343 = 0.137

- Envergadura: 0.48 - 0.16 = 0.32

Impureza Gini

4. Si consideramos que las instancias ya están clasificadas suficientemente bien, **FIN.**

   Si no, seguimos construyendo el árbol de forma iterativa, tomando como muestra inicial la muestra de cada hoja y realizando los pasos 1 - 4.

En resumen...

1. Calculamos la **Impureza Gini inicial** de la muestra.
2. Calculamos la **Impureza Gini** luego de hacer cada pregunta. Para ellos, hacemos un **promedio ponderado** de las impurezas resultantes en cada **hoja**, por pregunta.
3. Elegimos el atributo con **mayor reducción de impureza** (Ganancia Gini).
4. Si consideramos que las instancias ya están clasificadas suficientemente bien, FIN. Si no, seguimos construyendo el árbol de forma iterativa, tomando como muestra inicial la muestra de cada hoja y realizando los pasos 1 - 4.

Árboles: Algunos comentarios

1. **Entropía/ganancia de información** es otro criterio que podemos utilizar para medir el grado de impureza de una muestra y elegir el atributo que más la reduce. Conceptualmente es MUY parecido.

Árboles: Algunos comentarios

1. **Entropía/ganancia de información** es otro criterio que podemos utilizar para medir el grado de impureza de una muestra y elegir el atributo que más la reduce. Conceptualmente es MUY parecido.
2. Existen otras métricas que se podrían utilizar, que tienen ventajas en algunas situaciones específicas (por ejemplo, **Gain Ratio**, que corrige la preferencia de ganancia de información por atributos con demasiados valores) .

Árboles: Algunos comentarios

1. **Entropía/ganancia de información** es otro criterio que podemos utilizar para medir el grado de impureza de una muestra y elegir el atributo que más la reduce. Conceptualmente es MUY parecido.
2. Existen otras métricas que se podrían utilizar, que tienen ventajas en algunas situaciones específicas (por ejemplo, **Gain Ratio**, que corrige la preferencia de ganancia de información por atributos con demasiados valores) .
3. Nosotros aquí mostramos un ejemplo de **Clasificación Binaria** (dos clases). Los árboles generalizan muy bien a problemas multiclase y de regresión.

Árboles: Algunos comentarios

1. **Entropía/ganancia de información** es otro criterio que podemos utilizar para medir el grado de impureza de una muestra y elegir el atributo que más la reduce. Conceptualmente es MUY parecido.
2. Existen otras métricas que se podrían utilizar, que tienen ventajas en algunas situaciones específicas (por ejemplo, **Gain Ratio**, que corrige la preferencia de ganancia de información por atributos con demasiados valores) .
3. Nosotros aquí mostramos un ejemplo de **Clasificación Binaria** (dos clases). Los árboles generalizan muy bien a problemas multiclase y de regresión.
4. Hay mucha jerga en árboles: hojas, raíz, nodo, poda (pruning), Gini, información, profundidad, etc. Es fácil marearse. Este artículo y la documentación de Scikit-Learn y, sobretodo, la práctica, les servirán para ir incorporándolos.

Árboles: Ventajas y desventajas

- Simple de entender, interpretar y visualizar. Esto es una gran ventaja, también, al momento de comunicar nuestro trabajo.
- Entrenamiento rápido.
- Modelo base para modelos más complejos (Random Forest, xgboost, etc.).
- ¡Muchas implementaciones y variantes!

- Poder de generalización relativamente bajo en muchas circunstancias.
- Desempeño inferior a modelos más modernos.
- ¡Muchas implementaciones y variantes!

En Scikit-Learn

El módulo que contiene la implementación de árboles de decisión en Scikit-Learn es *tree*.
Como siempre, la documentación es muy buena.

Sus principales clases son:

- DecisionTreeClassifier
- DecisionTreeRegressor (esta la usaremos más adelante cuando veamos regresión).

Recomendamos mirar sus atributos, métodos y ejemplos.

¡BREAK!

Ph. Credit: Drew Coffmann

K-vecinos más cercanos (KNN)

Machine Learning

↓

Aprendizaje Supervisado

↓

Clasificación

Modelos

- **Árbol de Decisión**
- **k-nearest neighbors** ⇐
- Support Vector Machines
- Random Forest
- Perceptrón
- etc…

KNN - K Nearest Neighbors

IDEA: Dada una nueva instancia de la cual no conocemos la etiqueta objetivo, vamos a asumir que su etiqueta será igual a la de las instancias “vecinas” en el training set.

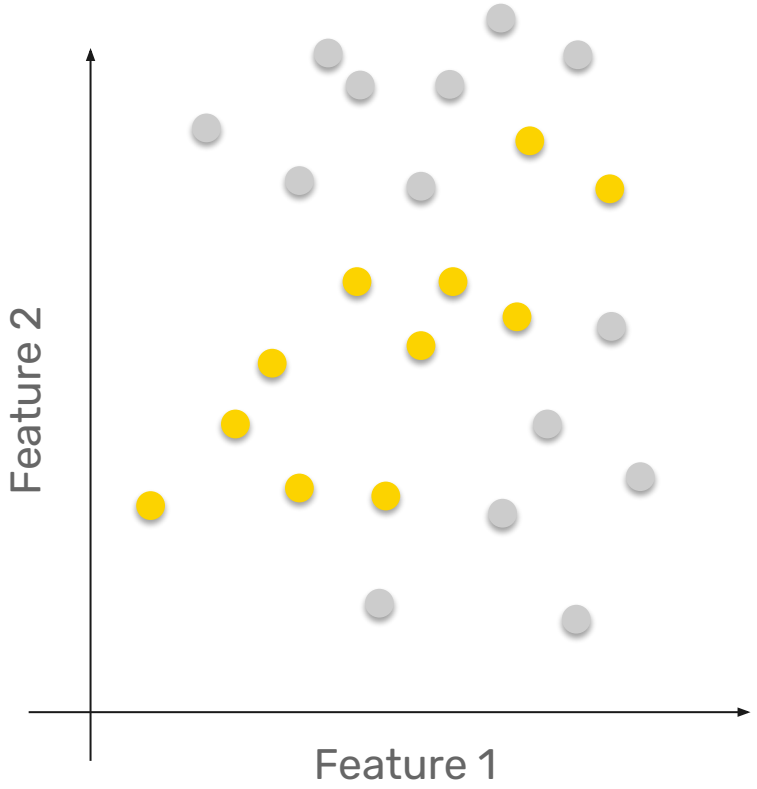
KNN - K Nearest Neighbors

IDEA: Dada una nueva instancia de la cual no conocemos la etiqueta objetivo, vamos a asumir que su etiqueta será igual a la de las instancias “vecinas” en el training set.

O dicho de otra forma...

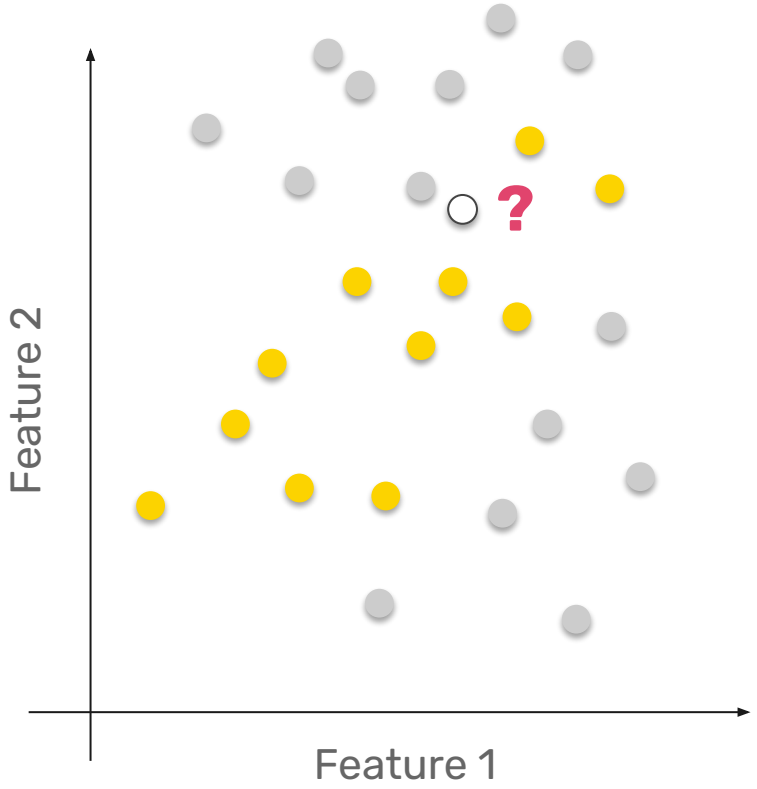
KNN - Ejemplo

Supongamos que tenemos un Dataset con 2 Features, en el cual cada instancia puede pertenecer a una de dos clases: “Gris” o “Amarillo”.



KNN - Ejemplo

Dada una nueva instancia, de la cual no conocemos su clase, vamos a recurrir a sus vecinos para clasificarla.

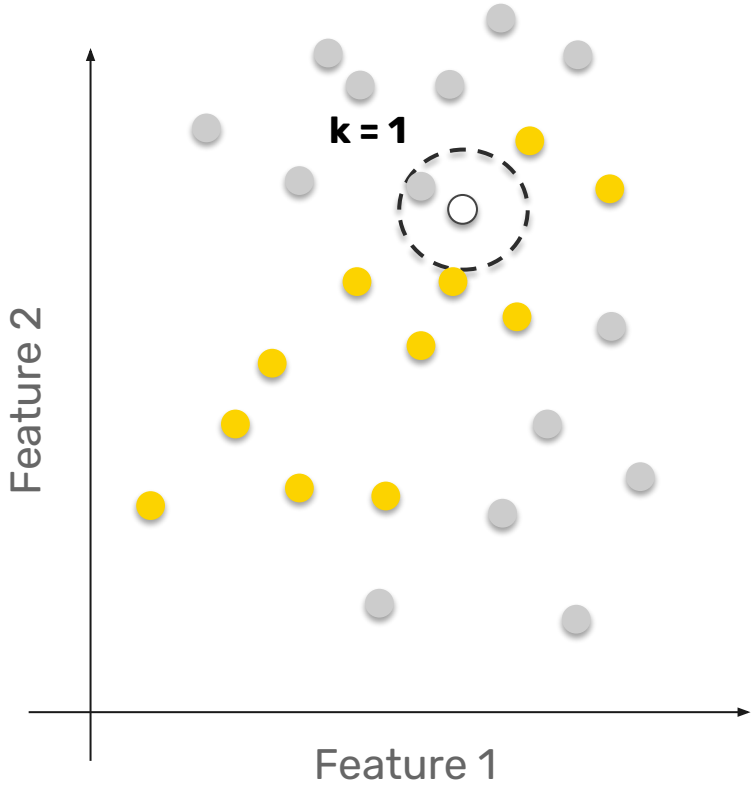


KNN - Ejemplo

¿Por K en nombre del algoritmo?

K es el número de vecinos que miramos para saber la clase de nuestra nueva instancia.

Si tomamos **K=1**, solo miraremos al vecino más cercano.



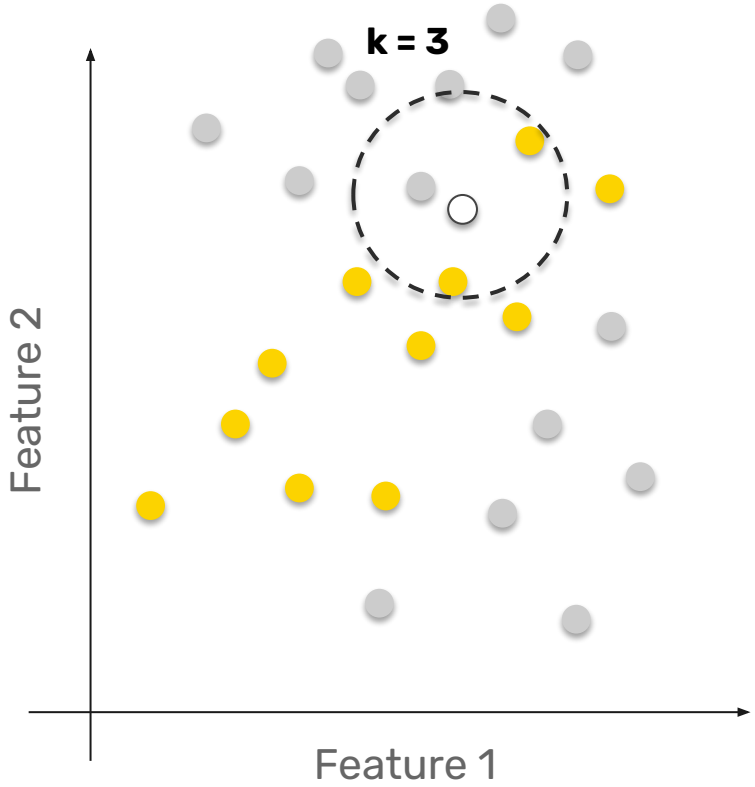
GRIS

KNN - Ejemplo

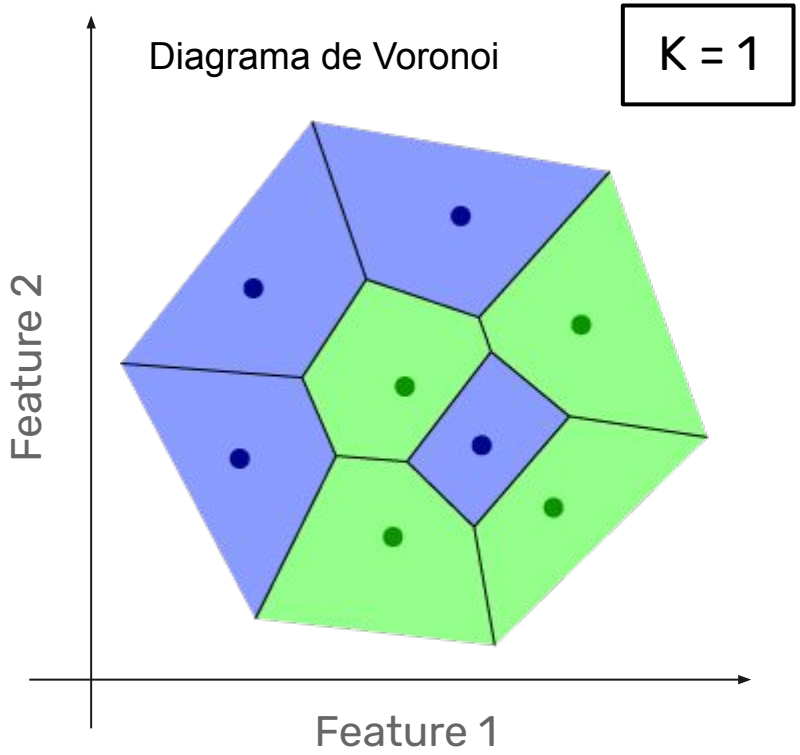
Si elegimos k > 1, se vota según el número de vecinos.

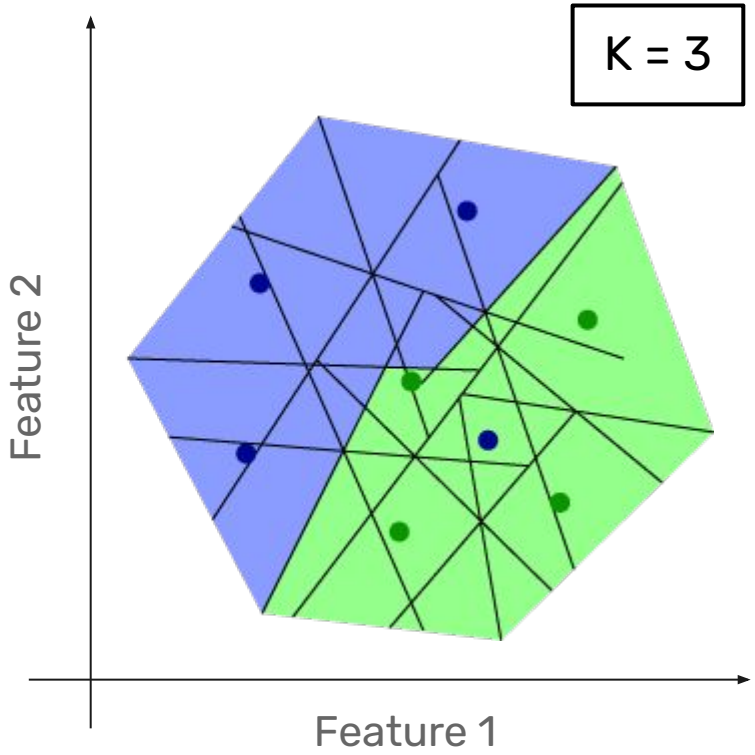
Por ejemplo, con k = 3 tenemos dos vecinos Amarillos y uno Gris.

AMARILLO



Fronteras de decisión según K





https://shapeofdata.wordpress.com/2013/05/07/k-nearest-neighbors/

Ventajas y Desventajas de KNN

- Simple de interpretar
- Entrenamiento rápido.

- Lento para clasificar (predecir)
- Ocupa mucho espacio en el disco (tiene que guardar todo el set de entrenamiento)

Hands-on training

Hands-on training

DS_Toolbox_12_DTyKNN

Para la próxima

1. Termina el notebook de hoy.
2. Lee la Toolbox 13.

ACÁMICA