Curso

5

MACHINE LEARNING & BIG DATA

Supervised Learning: Árboles de Decisión

Algoritmos Supervisados

JOSÉ NELSON ZEPEDA DOÑO

Cluster de Estudio: Advanced Analytics

Este material es el resumen de muchos autores que por medio de sus libros y documentos nos ofrecen fuentes riquísimas de conocimiento sobre los temas de Big Data y Machine Learning.
Algunas citas, figuras y tablas pueden ser encontradas de forma textual tal como lo indica el autor en su material original.

Nelson Zepeda
MIP • V 1.0
San Salvador El Salvador
Phone 503 79074137 • @nelsonzepeda733

Tabla de Contenido

Capítulo 1

Aprendizaje Supervisado

Uno de los usos principales del aprendizaje supervisado consiste en hacer predicciones a futuro basadas en comportamientos o características que se han visto en los datos.

MAchine Learning, son un conjunto de métodos/algoritmos diseñados para encontrar patrones y tendencias en los datos.

ML, se encuentra en la intersección entre las matemáticas y la estadística con la ingeniería de software y las ciencias de la computación.

De acuerdo al tipo de problema en análisis, podemos clasificar las técnicas en 2 grandes familias:

1. Aprendizaje Supervisado: En este proceso de aprendizaje la variable de salida está bien definida (variable objetivo), es decir estas técnicas nos son útiles cuando nos interesa hacer predicciones sobre una variable objetivo.

2. Aprendizaje No Supervisado: Este proceso de aprendizaje no implica tener una variable objetivo bien identificada, su objetivo no es hacer predicciones.

El aprendizaje supervisado permite buscar patrones en datos históricos relacionando todos los campos con un campo especial, llamado campo objetivo (Target).

A la vez, el aprendizaje supervisado tiene 2 grandes ramas:

- Clasificación

- Regresión

Un sistema de clasificación predice una categoría, mientras que una regresión predice un número.

Un ejemplo de clasificación es la clasificación de emails, los correos se "categorizan" como "spam" o como "legítimos" siendo esta la variable objetivo. Otro ejemplo clásico de clasificación en el mundo del machine learning es la predicción de bajas en, por ejemplo, una compañía de servicios. El objetivo en este caso es detectar los patrones de comportamiento de los clientes que sirven para predecir si se van a ir a la competencia. En este caso los clientes se clasifican como "baja" o "no baja" siendo esta la variable objetivo. La regresión, en cambio, predice un número, como por ejemplo cuál va a ser el precio de un artículo, o el número de reservas que se harán en mayo en un hotel.

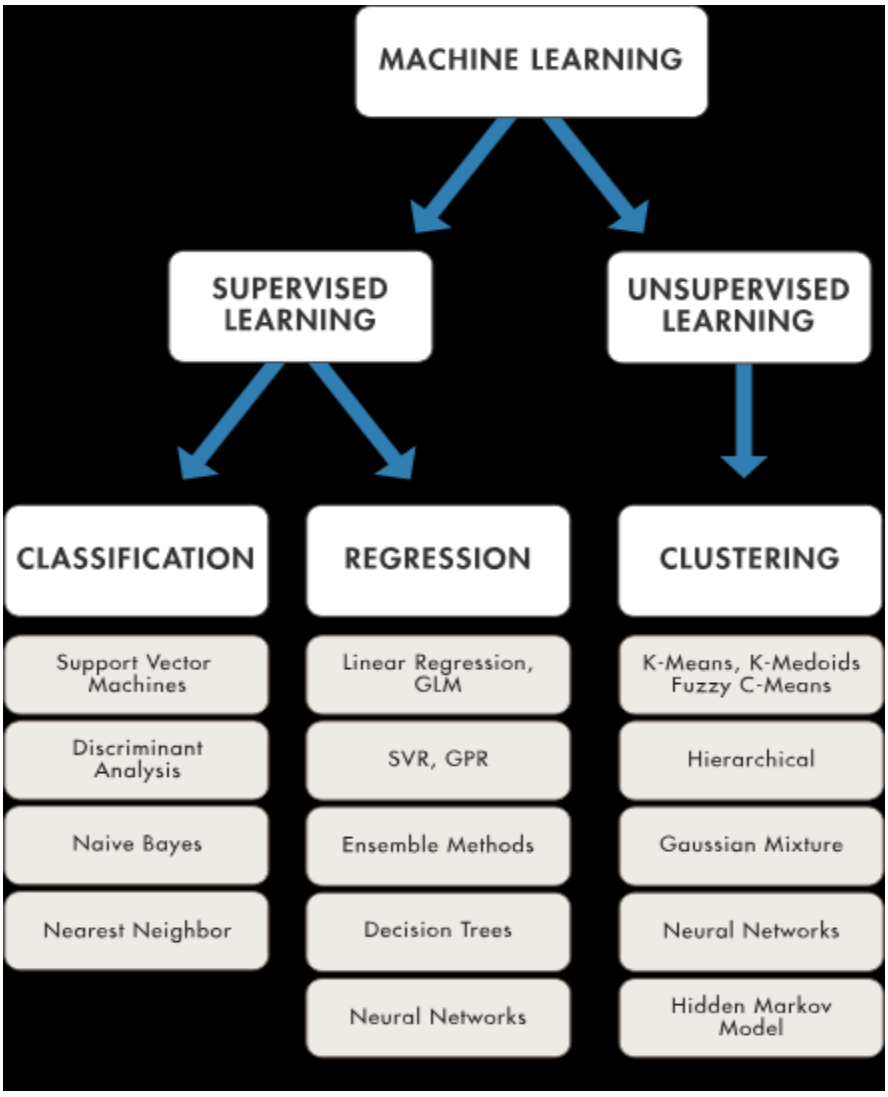

Figura 1-1 Familias de Algoritmos.

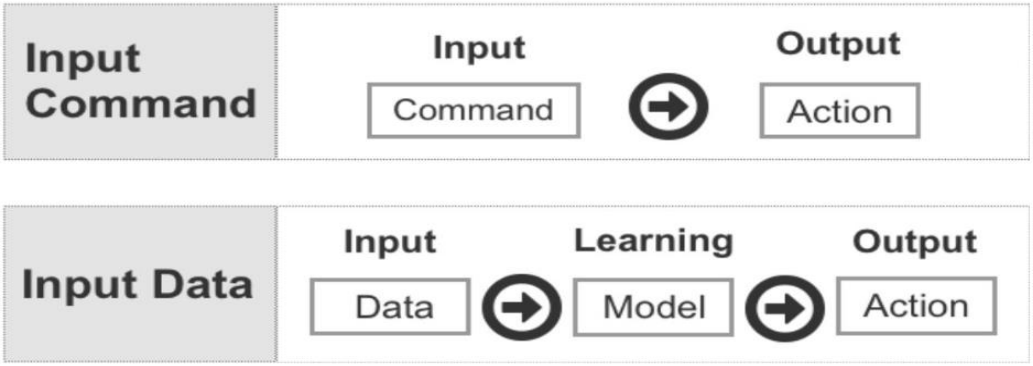

Figura 1-2 Aprendizaje vs No aprendizaje

Machine Learning es una rama de la Inteligencia Artificial, la siguiente figura muestra como una rama está contenida en otra:

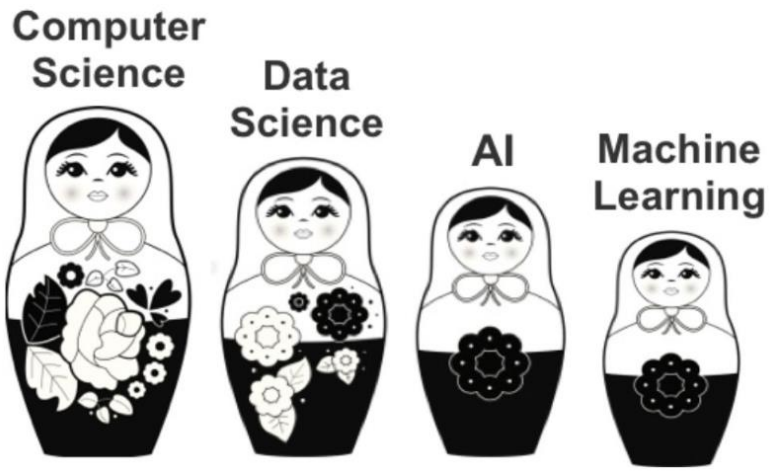

Figura 1-3 Machine Learning and Artificial Intelligence

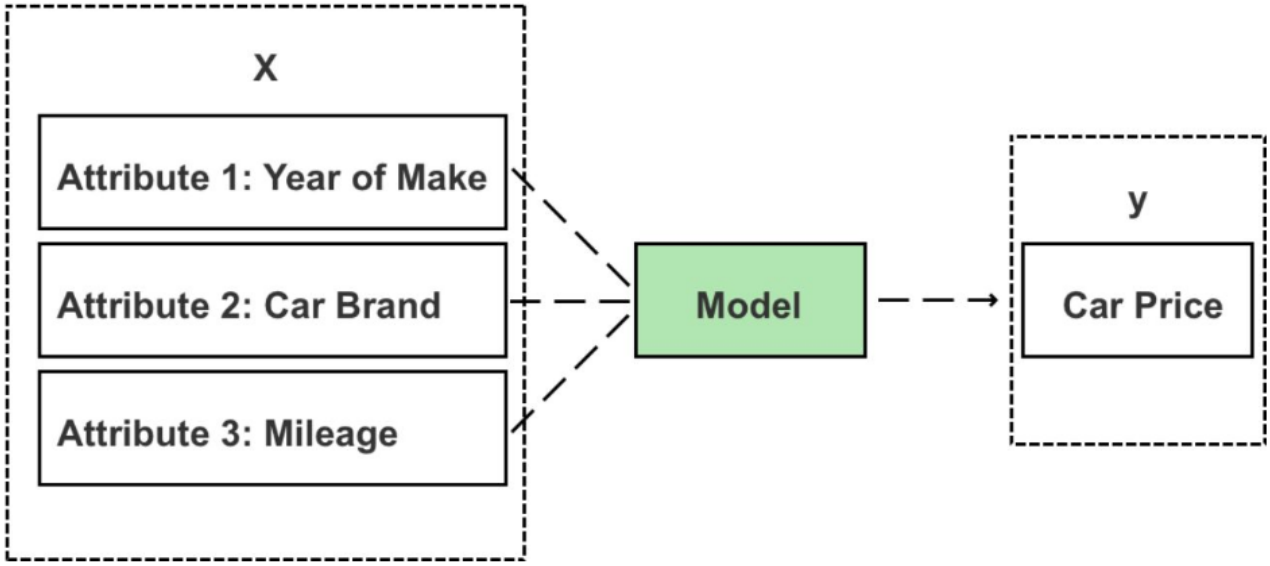

Figura 1-4 Modelo de predicción del precio de un vehículo.

Técnicas de Evaluación de Resultados

Siempre se debe evaluar el modelo para determinar si realizará un buen trabajo de predicción para nuevos y futuros datos. Dado que las futuras instancias tienen valores de destino desconocidos, se debe comprobar la métrica de precisión y otros indicadores del modelo de Machine Learning en relación con los datos de los que ya

se conoce la respuesta de destino y utilizar esta comprobación como un aproximado de la capacidad predictiva para futuros datos[1].

Las estrategias para particionar los datos en una porción para entrenamiento y una porción para testing son:

1. Holdout: Los registros se dividen en dos subconjuntos, el de entrenamiento y el de testeo. Usualmente la división es 2/3, 1/3

2. Sampling Method: Aplicación del método Holdout sobre muestras aleatorias.

3. K-Fold Cross Validation: Se dividen los datos en K subconjuntos y se construyen K modelos, cada modelo es testeado con los datos que no pertenecen a ese subconjunto.

4. Leave one-out method: Cada observación actúa como parte del conjunto de datos de prueba, mientras que el grupo restante funciona como grupo de entrenamiento

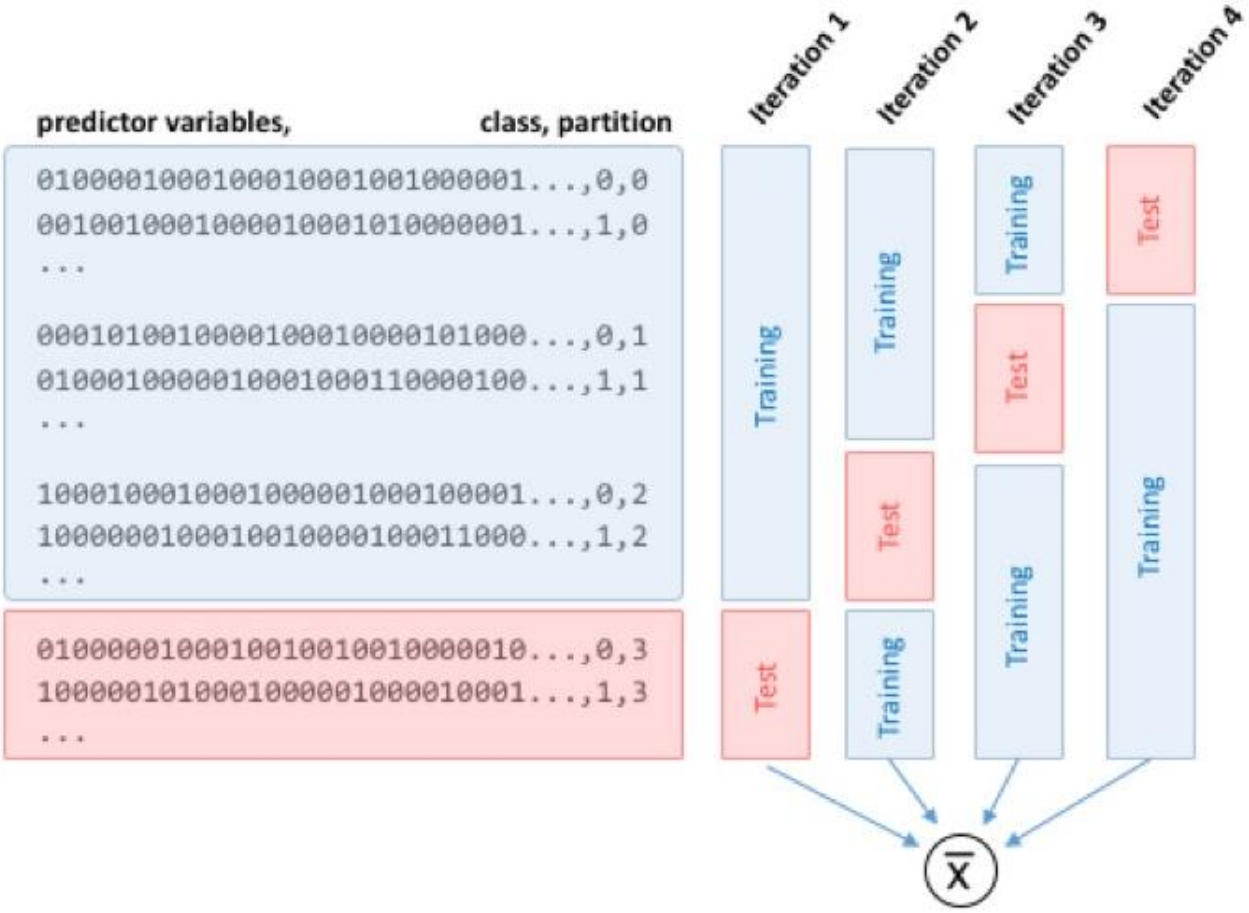

Figura 1-5 Validación Cruzada

[1] https://docs.aws.amazon.com/es_es/machine-learning/latest/dg/evaluating_models.html

Capítulo 2

Árboles de Decisión

William T. Greenwood, define a la toma decisión como una resolución de problemas, la cual necesita, para llegar a la mejor alternativa, un diagnóstico; realizado por una búsqueda adecuada, evaluación de las alternativas, y la elección de la mejor decisión.

Los árboles de decisión son un método usado en distintas disciplinas como modelo de predicción. Estos son similares a diagramas de flujo, en los que llegamos a puntos en los que se toman decisiones de acuerdo a una regla.

Otros conceptos

- Un árbol de decisión es un gráfico que mediante los métodos de construcción de ramas ilustra el curso de acción para una salida determinada.
- Un árbol de decisión es un flujo que ayuda a tomar decisiones por medio de la evaluación de diferentes condicionales.
- Un árbol de decisión ayuda a analizar escenarios y consecuencias de tomar una decisión y que regularmente sería muy difícil de visualizar.

Existen 2 tipos de árboles de decisión, clasificación y regresión.

En un "árbol de clasificación", Cada ramificación contiene un conjunto de atributos o reglas de clasificación asociadas a una etiqueta de clase específica, que se halla al final de la ramificación.

Por otro lado, si la variable predicha es un número real, como un precio. los árboles de decisión con resultados posibles, infinitos y continuos se llaman "árboles de regresión".

Elementos de un árbol de decisión.

Los árboles de decisión son un método usado en distintas disciplinas como modelo de predicción. Estos son similares a diagramas de flujo, en los que llegamos a puntos en los que se toman decisiones de acuerdo a una regla.

- Un nodo de probabilidad, muestra las probabilidades de ciertos resultados.
- Un nodo de decisión, muestra una decisión que se tomará.

- Un nodo terminal muestra el resultado definitivo de una ruta de decisión.
- Nodo Raíz: Nodo que contiene la pregunta original que inicia el árbol.
- Nodo Hijo: Es el resultado de dividir un nodo en 2 o más sub-conjuntos de datos.
- Poda (Prunning): proceso de remover un sub conjunto de datos del árbol.
- Ramas: Es una conexión entre nodos

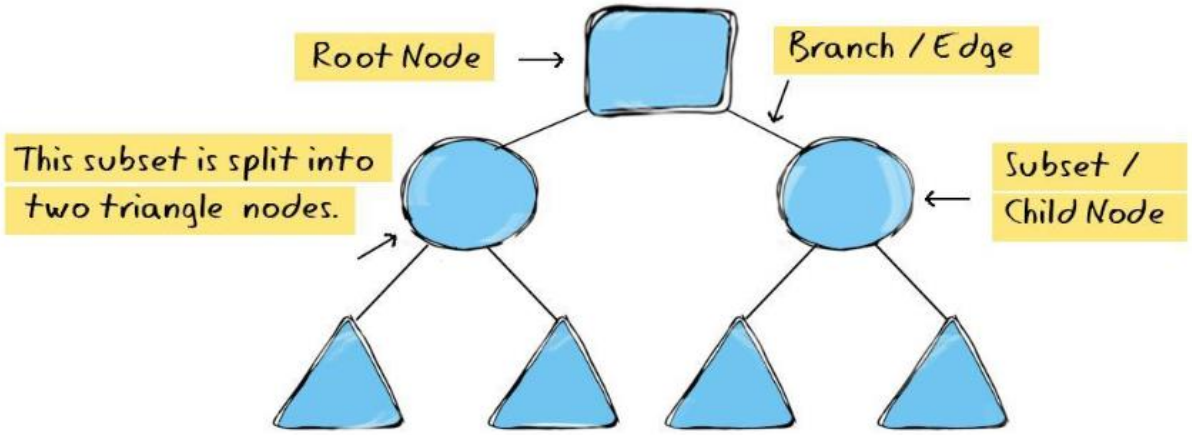

Figura 2-1 Elementos de un Árbol de Decisión

Construcción de un árbol de decisión.

Un modelo basado en árboles de decisión implica diferentes pasos que se describen en la siguiente figura:

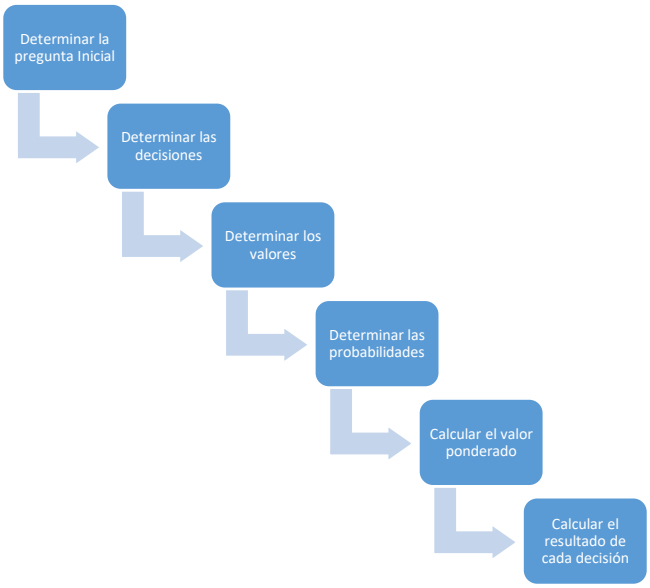

Figura 2-2 Proceso de Generación de un Árbol de Decisión

Ventajas y desventajas de un árbol de decisión.

Ventajas:

- El costo del uso del árbol para predecir los datos disminuye con cada punto de datos adicional.
- Funciona para los datos numéricos o categóricos.
- Puede modelar problemas con múltiples resultados.
- Usa un modelo de caja blanca (lo que hace que los resultados sean fáciles de explicar).
- La fiabilidad de un árbol se puede cuantificar y poner a prueba.
- Tiende a ser preciso independientemente de si viola las suposiciones de los datos de origen.
- Pueden ayudar a visualizar un problema

Desventajas:

- Cuando se presentan datos categóricos con múltiples niveles, la información obtenida se inclina a favor de los atributos con mayoría de niveles.
- Los cálculos pueden volverse complejos al lidiar con la falta de certezas y numerosos resultados relacionados.
- Las conjunciones entre nodos se limitan a AND, mientras que los gráficos de decisión admiten nódulos relacionados mediante OR.

Los diversos algoritmos para construir un árbol de decisión son:

- CART
- CHAID
- ID3
- C4.5
- C5.0

Índices de Pureza en los Nodos.

Un árbol de decisión se construye en base a divisiones BINARIAS tanto para atributos numéricos como no numéricos, el criterio para seleccionar la mejor división se basa en el índice de Gini y en la entropía los cuales miden la pureza de los datos.

En la práctica, un mayor Gini implica menor pureza y se puede definir como la probabilidad de no sacar dos registros de la misma clase del nodo.

Las formulas correspondientes son:

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

$$Entropía(t) = -\sum_j p(j \mid t) \log_2 p(j \mid t)$$

Ejemplo de cálculo de Gini

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

C1	**0**
C2	**6**

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1

Gini = 1 – P(C1)² – P(C2)² = 1 – 0 – 1 = 0

C1	**1**
C2	**5**

P(C1) = 1/6 P(C2) = 5/6

Gini = 1 – (1/6)² – (5/6)² = 0.278

C1	**2**
C2	**4**

P(C1) = 2/6 P(C2) = 4/6

Gini = 1 – (2/6)² – (4/6)² = 0.444

Desempeño del Algoritmo.

Matriz de confusión es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. Uno de los beneficios de las matrices de confusión es que facilitan ver si el sistema está confundiendo dos clases.

Si en los datos de entrada el número de muestras de clases diferentes cambia mucho la tasa de error del clasificador no es representativa de lo bien que realiza la tarea el clasificador. Si por ejemplo hay 990 muestras de la clase 1 y sólo 10 de la clase 2, el clasificador puede tener fácilmente un sesgo hacia la clase 1. Si el clasificador clasifica todas las muestras como clase 1 su precisión será del 99%. Esto no significa que sea un buen clasificador, pues tuvo un 100% de error en la clasificación de las muestras de la clase 2[2]

		Valor Predicho		
		Gato	Perro	Conejo
Valor Real	Gato	5	3	0
	Perro	2	3	1
	Conejo	0	2	11

Figura 2-3 Ejemplo de Matriz de Confusión

ROC (Receiver Operating Curve), Indica la relación entre los falsos positivos y los verdaderos positivos permitiéndonos establecer el nivel de calidad o robustez que nuestro modelo tiene para predecir, algunos atores hacen referencia a esta curva como AUC. Se basa de los siguientes ratios:

	Positive (Actual)	Negative (Actual)
Positive (Predicted)	True Positive (TP)	False Positive (FP)
Negative (Predicted)	False Negative (FN)	True Negative (TN)

Figura 2-4 Matriz de Confusión binaria

$$TruePositveRate = \frac{TP}{TP + FN}$$

[22] https://es.wikipedia.org/wiki/Matriz_de_confusi%C3%B3n

$$FalsePositiveRate = \frac{FP}{FP + TN}$$

F-Measure: Tambien conocido como Score F1, llamadas Precision y Recall. La precisión es el porcentaje de instancias clasificadas correctamente.

$$\mathrm{Pr}ecision = \frac{TP}{TP + FP}$$

$$\mathrm{Re}call = \frac{TP}{TP + FN}$$

$$F - Measure = \frac{2.\mathrm{Pr}ecision.\mathrm{Re}call}{\mathrm{Pr}ecision + \mathrm{Re}call}$$

Bibliografía

- Decision Trees
  By Chris Smith, 2017

- R Data Analysis Cookbook
  by Kuntal Ganguly,2017