1. Supercised Learning

El Supercised Learning se puede clasificar en dos grandes grupos:

1. Classification:
   a) Regresión logística.
   b) K-Nearest Neighbors (KNN).
   c) Decision Trees.
   d) Random Forest.
2. Regression.
   a) Linear Regression.

Existen métodos y algoritmos que se pueden usar en ambos problemas:

Supervised Learning	**Classification**	**Regression**
Logistic Regression	✓	
k-Nearest Neighbors (KNN)	✓	✓
Decision Trees	✓	✓
Random Forest	✓	✓
Linear Regression		✓

Cuadro 1:

Métricas Clasificación:

- Precisión.
- Recall (exhaustividad).
- Valor-F (F_1)

Ver video 03b min 40. Matriz de confucion.

1.1. Classification

En el contexto del aprendizaje automático supervisado, la variable predictora (también conocida como variable independiente o característica) es una variable que se utiliza para predecir el valor de la variable objetivo (también conocida como variable dependiente o target). Las variables predictoras son los datos de entrada al modelo y la variable objetivo es el resultado que se desea predecir.

1.1.1. Regresión logística.

La regresión logística es un tipo de algoritmo de aprendizaje supervisado que se utiliza principalmente para resolver tareas de "clasificación" binarias, aunque contiene la palabra regresión". Se lo utiliza para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las **variables predictoras** o independientes. Un ejemplo de clasificación podría ser la detección de spam: un programa de aprendizaje automático puede aprender a marcar el spam después de recibir ejemplos de correos electrónicos spam marcados (**variable objetivo** "target" sería una variable binaria que indica si un correo electrónico es spam o no) por los usuarios y ejemplos de correos electrónicos regulares no spam (también llamados "ham"). Notebook ejemplo [2].

1.1.2. K-Nearest Neighbors (KNN).

k vecinos más cercanos es un método de clasificación no paramétrico. Video de youtube [3].
Se utiliza principalmente para la clasificación de datos no lineales y para resolver problemas de clasificación en los que los datos son muy complejos y desestructurados.

Resumen:

- Es sensible a conjuntos de datos no balanceados.
- Es sensible aoutliers.

1.1.3. Decision Trees

Existen varios algoritmos ID3, C4.5 y CART. Pagina buena [1]:

• **ID3 - Iterative Dichotomiser 3:** Genera un árbol de decision a partir de un conjunto de ejemplos.
Este algoritmo usa las metricas de *entropía* y *ganancia de la información*. Cuando se usa la librería sklearm *sklearn.tree.DecisionTreeClassifier* en el parámetro *criterion* colocar *entropy*.

- Un nodo principal llamado raíz en la parte superior.
- Nodos terminales. como su nombre lo indica, son nodos donde termina el flujo y que ya no son raiz de ningún otro nodo. Estos nodos terminales deben contener una respuesta, o sea, la clasificación a que pertenece el objeto que ha conducido hasta él.
- Los demás nodos representan preguntas con respecto al valor de uno de los atributos.

- Las líneas nodos representan preguntas con respecto al valor de uno de sus atributos.
- Las líneas representa las posibles respuestas que los atributos pueden tomar.

Algoritmo básico

1. Calcular la entropía para todas las clases.
2. Calcular la entropía para cada valor posible de cada atributo.
3. Seleccionar el mejor atributo basado en la reducción de la entropía. usando el calculo de la ganancia de la información.
4. Iterar, para cada sub-nodo, Excluyendo el nodo raíz, que ya fue usado.

- **C4.5:** Utiliza la metrica *gini*.

1. Mitiga el sobreajuste por que emplea inherentemente el proceso de poda de un solo paso.
2. Funciona para datos discretos y continuos.
3. Es útil para datos incompletos.

Metricas:

- **Entropía:**
  La medida del desorden o la medida de la pureza. Básicamente, es la medida de la impureza o aleatoriedad de los datos.

  Para calcular la entropia de n clases se utiliza la fórmula:

$$H(S) = \sum_{i=1}^{n} -p_i \cdot \log_2(p_i) \tag{1}$$

  Dónde:

  - S: es una lista de valores posibles.
  - p_i: es la probabilidad de los valores.
  - i: cada uno de los valores.

  Importante:

  - Para una muestra homogénea la entropia es igual a cero 0. Si no existe aleatoriedad, es decir, una moneda cargada.

- La máxima entropía viene dada por $\log_2(n)$, n son los posibles valores de salida. Si $n = 2$ (true o false) entonces, la máxima entropía es 1. O sea es la máxima incertidumbre, ejemplo moneda equilibrada.

- **Ganancia de la Información:** La ganancia de la información se aplica a cuantificar qué característica, de un conjunto de datos dados, proporciona la máxima información sobre la clasificación.

1.1.4. Random Forest - Bosques aleatorios.

Usa una técnica, o meta-algoritmo llamado Bootstrap aggregating. Se crean m tablas reducidas en atributos y para cada una de ellas entrenamos un árbol.

1.2. Regresión

1.2.1. Linear Regression

La regresión lineal es un método estadístico que se utiliza para estudiar la relación entre una variable dependiente (también conocida como variable objetivo) y una o más variables independientes (también conocidas como variables predictoras). El objetivo de la regresión lineal es encontrar la línea que mejor se ajuste a los datos y pueda utilizarse para hacer predicciones.

En el caso de una sola variable independiente, la línea de regresión se puede representar mediante la ecuación $y = a + bx$, donde y es la variable dependiente, x es la variable independiente, a es la intersección en y y b es la pendiente de la línea. Los valores de a y b se determinan utilizando los datos disponibles para minimizar la suma de los errores cuadrados entre los valores observados y los valores predichos por la línea de regresión. Ver ejemplo

1.2.2. Regresión Lineal Múltiple

Consiste en predecir una respuesta numérica y en base a múltiples variables predictoras $x_1, x_2, ..., x_n$, suponiendo una relación lineal.

$$y = a_1 \cdot x_1 + a_2 \cdot x_2 + ... + a_n \cdot x_n + b \tag{2}$$

• **Métricas Regresión:** Las métricas de evaluación son medidas utilizadas para evaluar el rendimiento de un modelo de aprendizaje automático. En el caso de los problemas de regresión en el aprendizaje supervisado, estas métricas nos ayudan a determinar qué tan bien nuestro modelo está haciendo predicciones cuantitativas, como valores continuos.

Metricas para la regresión:

- Raíz del error cuadrático medio **(RMSE)**.
- Error absoluto medio **(MAE)**.

- Error cuadratico medio **(MSE)**.
- Suma Residual de los cuadrados **RSS**.

Para cada una de las métricas:

- m : número de instancias
- h : Funcion hipotesis, es el modelo entrenado. En este caso regresion lineal.
- x : Todos los valores de entrada, todas las columnas.

Error cuadrático medio (MSE):

$$MSE(X,h) = \frac{1}{m} \cdot RSS = \frac{1}{m} \cdot \sum_{i=1}^{m} (h(x^{(i)}) - y^{(i)})^2 \tag{3}$$

Error absoluto medio (MAE):

$$MAE(X,h) = \frac{1}{m} \cdot \sum_{i=1}^{m} |h(x^{(i)} - y^{(i)}| \tag{4}$$

Raíz del error cuadrático medio (RMSE):

$$RMSE(X,h) = \sqrt{\frac{1}{m} \cdot \sum_{i=1}^{m} (h(x^{(i)} - y^{(i)}))^2} \tag{5}$$