**Curso: tópicos de inteligencia artificial**

**Preprocesamiento de datos**

¿Qué incluye la Preparación de Datos? El Preprocesamiento de Datos engloba a todas aquellas técnicas de análisis de datos que permite mejorar la calidad de un conjunto de datos de modo que las técnicas de extracción de conocimiento/minería de datos puedan obtener mayor y mejor información (mejor porcentaje de clasificación, reglas con más completitud, etc.)

Los datos reales pueden ser impuros, pueden conducir a la extracción de patrones/reglas poco útiles. Esto se puede deber a:

* Datos Incompletos: falta de valores de atributos
* Datos con Ruido
* Datos inconsistentes (incluyendo discrepancias)

La preparación de datos puede generar un conjunto de datos más pequeño que el original, lo cual puede mejorar la eficiencia del proceso de Minería de Datos.

El preprocesamiento puede incluye:

* Selección relevante de datos: eliminando registros duplicados, eliminando anomalías
* Reducción de Datos: Selección de características, muestreo o selección de instancias, discretización.

La preparación de datos genera “datos de calidad”, los cuales pueden conducir a patrones/reglas de calidad. Por ejemplo, se puede:

* Recuperar información incompleta.
* Eliminar outliers
* Resolver conflictos

Importar datos desde un archivo csv.

#importar dataset  
dataset=pd.read\_csv("Data.csv")

Dividir los conjuntos de información en diferentes vectores arreglos de datos. En Python las matrices se especifican como [fila,columna].

#conjunto de características  
X=dataset.iloc[:,0:3].values  
#vector de clases  
Y=dataset.iloc[:,3].values

Reemplazar los datos faltantes o nulos por un valor numérico, por lo general se suele reemplazar por la media de todos los datos de la columna (variable) donde faltan los valores, considerando que los datos siguen una distribución normal.

from sklearn.impute import SimpleImputer

from sklearn.compose import ColumnTransformer

#reemplazar datos faltantes con la media de los demás datos  
imputer=SimpleImputer(missing\_values=np.nan, strategy='mean', fill\_value=None, verbose=0, copy=True, add\_indicator=False)  
imputer.fit(X[:, 1:3])  
X[:, 1:3]=imputer.transform(X[:, 1:3])

Una variable categórica es aquella que permite clasificar una serie de datos por medio de valores fijos asociados a una cualidad o categoría concreta. La variable categórica, a diferencia de las variables cardinales o continuas (que permiten cálculos numéricos), clasifica a los individuos o casos. Normalmente toman valores representados por números enteros, como el uno o el cero, pero estos son solo eso, representaciones. Estas variables se pueden representar como variables dummy o indicadoras. Estas variables toman dos valores usualmente, cero y uno. Los dos valores significan que la observación pertenece a una de dos categorías. Las variables dummy o indicadoras sirven para identificar categorías o clase a las que pertenecen las observaciones.

Codificando datos categóricos en numéricos.

from sklearn.preprocessing import LabelEncoder, OneHotEncoder

#Codificando datos categoricos  
labelencoder\_X=LabelEncoder()  
X[:,0]=labelencoder\_X.fit\_transform(X[:,0])

Convertir a variables dummy

#Traducir categoricos a dummy  
ct=ColumnTransformer([("Country", OneHotEncoder(), [0])], remainder="passthrough")  
X=ct.fit\_transform(X)

La estandarización de los datos mejora los resultados cuando los valores entre variables tienen dimensiones muy diferentes. Entre -1 y 1 para que ninguna característica influya más que otra. La estandarización ajusta los datos para que su media sea 0 y la normalización ajusta los valores entre un mínimo y un máximo definido*. La variable Y no se debe escalar para clasificación*.

Graphical user interface, application, table, Word

Description automatically generated

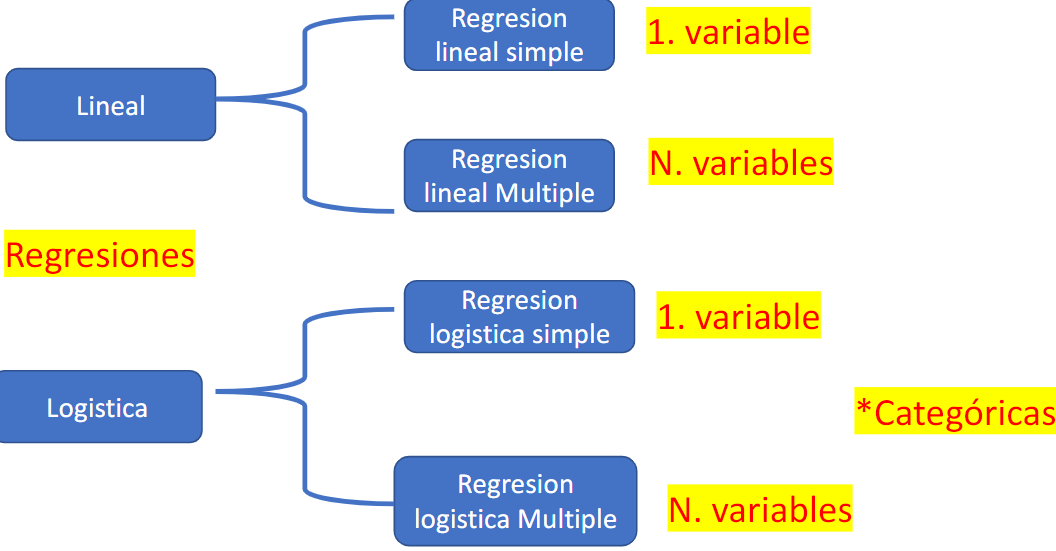
#escalar datos  
sc\_X=StandardScaler()  
X=sc\_X.fit\_transform(X\_train)

Dividir set de entrenamiento en entrenamiento y pruebas.

#separar conjunto de datos  
X\_train, X\_test, Y\_train, Y\_test=train\_test\_split(X,Y,test\_size=0.2, random\_state=0)

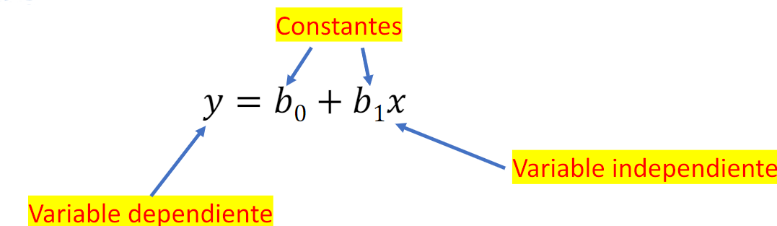
**Regresiones**

El análisis de regresión es una herramienta de frecuente uso en estadística. La cual permite investigar las relaciones entre diferentes variables cuantitativas. Esto, mediante la formulación de ecuaciones matemáticas. Visto de otro modo, dicho análisis es un proceso o modelo que analiza el vínculo entre una variable dependiente y una o varias variables independientes. Así, a partir de dicho estudio, se halla una relación matemática. Una de las principales aplicaciones del análisis de regresión es la proyección con diferentes escenarios. Esto, teniendo en cuenta el grado de influencia (en estadística se conoce a esto como correlación) sobre la variable dependiente.



**Regresión lineal simple**

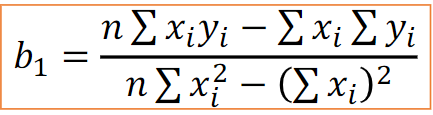
Pretende relacionar una variable dependiente con una o más variable independientes. Supóngase que se tiene un conjunto de n pares de observaciones (xi,yi), se busca encontrar una recta que describa de la mejor manera cada uno de esos pares observados.



Chart, scatter chart

Description automatically generated

El criterio de **mínimos cuadrados** nos proporciona un valor de y uno de y, tal que



A picture containing text, orange, clock

Description automatically generated

n = número de muestras

= recta propuesta

=punto en el espacio

