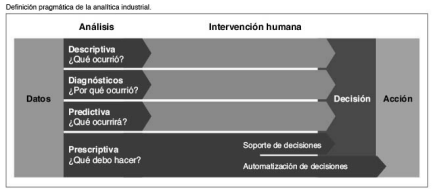
G8: Modelamiento predictivo para Ciencias de datos Cheat Sheet

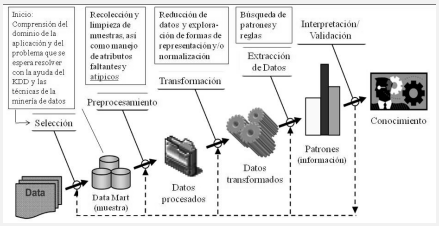
by esepulveda via cheatography.com/154288/cs/33130/

1.1 ANÁLISIS PREDICTIVO



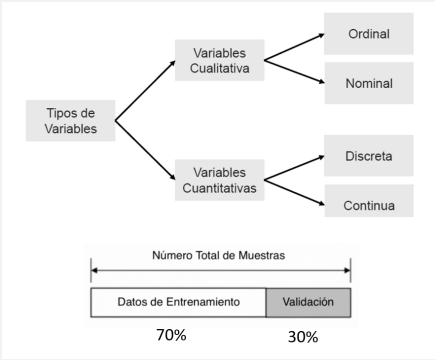
El análisis predictivo consiste en la tecnología que aprende de la experiencia para predecir el futuro comportamiento de individuos para tomar mejores decisiones. I NFORMS Institute for Operations Research and the Management Scie nces

1.4 PROCESO KDD



Proceso de Extracción del Conocimiento conocido como Knowledge Discovery in Databases KDD

1.5 KDD - Selección de los Datos

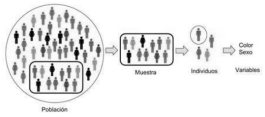


Exploración de los datos : Base Original -> Tablón de Datos (Minable)

Muestra de entrenamiento : Un subconjunto para entrenar un Modelo

Muestra de validación : Un subconjunto para probar el modelo entrenado

2.2 CONCEPTOS ESTADISTICA DESCRIPTIVA

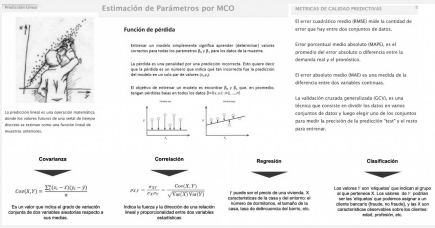


La población, es el conjunto total de objetos o personas de interés en un estudio. *una característica relevante es todos sus elementos deben cumplir con un conjunto predefinido de características*.

La Muestra es el subconjunto de la población, la cual se utiliza para estudiar las características de la población en general. Estas deben ser: Aleatorias y representativas.

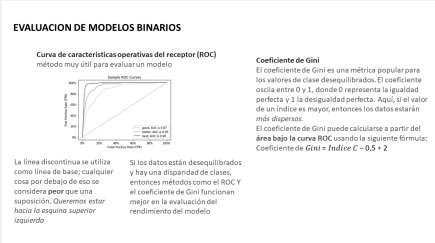
Variable Aleatoria cualquier característica que tome dos o más valores en una población.

3 PREDICIÓN LINEAL

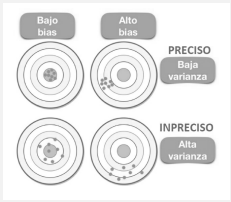


MCO Minimo cuadrado Ordinario

5.2 EVALUACION DE MODELOS BINARIOS



1.2 BIAS O SESGO Y VARIANZA



¿Qué es el Bias?

El bias o sesgo puede ser pensado como un modelo que no ha tenido en cuenta toda la información disponible en el set de datos, lo que dificulta predicciones precisas.

¿Qué es la varianza?

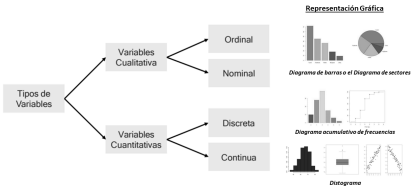
La Varianza es una medida de dispersión que se utiliza para representar la variabilidad de un conjunto de datos respecto de la media aritmética de los mismo.

1.6 KDD -Preprocesamiento

1. Análisis Descriptivo Univariado	Evaluar mediad de tendencia central (Mediana, Moda), de dispersión. Los datos null se replazan por 0
2. Análisis Descriptivo Multivariado	Análisis de Correlaciones, Gráficos de Dispersión, Etc.
3. Análisis Descriptivo Temporal (de Proporciones)	Debe contar con estabilidad temporal de proporciones.
4. Análisis Descriptivo Temporal (de Predicción)	Busqca predecir el comportamiento de una variable en particular.

MODELOS KDD - SEMMA - CRISP

2.3 TIPOS DE VARIABLES ALEATORIAS



Variable Cualitativa

Ordinal presenta modalidades no numéricas, en las que existe un orden por Ej: notas en un examen.

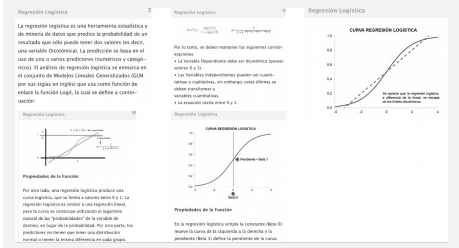
Nominal presenta modalidades no numéricas que no admiten un criterio de orden. Ej: El estado civil, con las siguientes modalidades: sol, cas, sep, divor y viudo.

Variable Cuantitativa

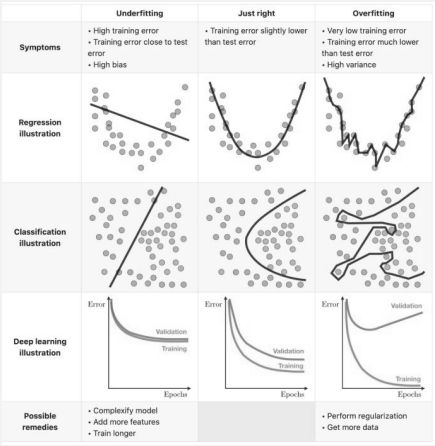
Discreta puede asumir un número contable de valores: Ej: # hijos en una familita.

Continua puede asumir un número incontable de valores. Ej: peso de una persona.

4 REGRESIÓN LOGISTICA



1.3 SOBREAJUSTE E INFRAJUSTE



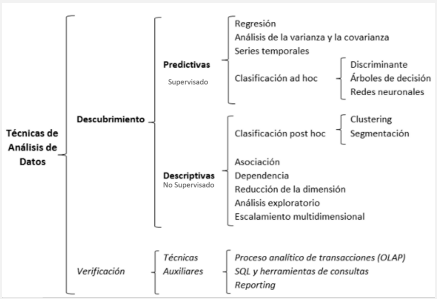
Overfitting

Se produce cuando un modelo modela demasiado bien los datos de entrenamiento., por lo que no es capaz de generalizar, y cuando le lleguen nuevos datos obtendrá pésimos resultados.

Underfitting

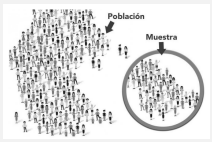
Se produce cuando nuestro modelo no es capaz de identificar patrones. Por lo que tendrá siempre
pésimos resultados.

1.7 KDD - Minería de Datos



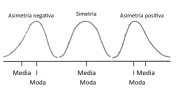
Tipos de aprendizaje de datos

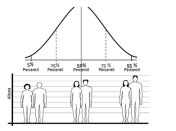
2.1 ESTADÍSTICA DESCRIPTIVA



La estadística descriptiva es un conjunto de técnicas numéricas y gráficas para describir y analizar un grupo de datos, sin extraer conclusiones (inferencias) sobre la población/universo a la que pertenecen.
En términos generales la estadística descriptiva busca: Describir o Caracterizar un Grupo de Datos.

2.4 CONCEPTOS ESTADÍSTICA DESCRIPTIVA





La moda se define, como el valor que más se repite dentro de la muestra de observaciones. Aplica tanto para variables cuantitativas como cualitativas
La Mediana es el valor hasta el cual se encuentran el 50 % de los casos, bajo el supuesto que los datos de la muestra están ordenados de menor a mayor.

Los percentiles es una medida estadística la cual divide una serie de datos ordenados de menor a mayor en cien partes iguales. Se trata de un indicador que busca mostrar la proporción de la serie de datos que queda por debajo de su valor.
Los percentiles más populares son: ***La mediana, la cual es el percentil 50***, los cuartiles que se definen como el percentil 25, 50 y 75 y los deciles que corresponden a los percentiles 10, 20, 30, 40, 50, 60, 70, 80 y 90.

5.1 ÁRBOLES DE DECISIÓN

Los **árboles de decisión** son modelos predictivos formados por reglas binarias(sí/no) con las que se consigue repartir las observaciones en función de sus atributos y predecir así el valor de la variable respuesta

Los métodos basados en árboles se han convertido en uno de los referentes dentro del ámbito predictivo debido a los buenos resultados que generan en problemas muy diversos

Ventajas	Desventajas
Los árboles son fáciles de interpretar	La capacidad predictiva de los modelos basados en un único árbol es bastante inferior a la conseguida con otros modelos
Los árboles pueden incluir tanto predictores numéricos como categóricos	Son sensibles a datos de entrenamiento dados lanzazados
No se ven muy influenciados por outliers	No son capaces de extrapolar fuera del rango de los predictores observados en los datos de entrenamiento
Son muy útiles en la exploración de datos, ya que permiten de forma rápida y eficiente las variables (predictores) más importantes	
Pueden aplicarse a problemas de regresión o clasificación	

Calibración de Muestras: es recomendable desarrollar modelos donde la asignación de probabilidad este
Balanceada con el fin de aumentar la dispersión y evitar la concentración de variables, existen 3 maneras

-Selección de muestras mismo tamaño: Implica seleccionar el N de una de las dos muestras dicotómicas y seleccionar a través de algún mecanismo de muestreo la misma cantidad de casos para ambas muestras.

-Ponderación de casos: Implica determinar los pesos de ambas muestras dicotómicas y ajustar la regresión logística con la ponderación asignada á fin de obtener la misma participación de las muestras dicotómicas

-modificar el intercepto (el "a" de la ecuación de regresión y = a + bx): Implica restarle al intercepto de la regresión logística ajustada el valor del WOE global a fin de extender la función al equilibrio