¿Què se requiere para tener calidad de datos?

Utilidad similar a Calidad

Datos de calidad = datos ùtiles

Variables continuas: Cualquier valor en un rango del espectro de número. Ejemplo coordenadas

Variables discretas: Solo un valor especifico, ejemplo genero

--------------------------------------------------------------------------------------------------------

Visualización de Datos

Identificar por tres caminos distintos como se pueden quitar los outlyers y valores atípicos. Para después comparar entre el la cantidad de registros de la población inicial vs la cantidad de registros después de la depuración.

Esto solo es para variables continuas.

Método 1. Por percentiles

Los percentiles muestran que tan concentrada está la población por categorías. Entonces lo que se tiene que hacer es:

Observar que tan concentrada está la población en las categorías y acotar éstas de acuerdo a un criterio de sensibilidad (felling)

Mètodo 2. Por la media

Bajo el supuesto de que todas las variables se comportan de una manera normal (Normal(Mu, sigma2)) se puede acotar mediante la desviacion tipica (raiz cuadrada de la varianza) por (generalmente) 3. Entonces lo que se tiene que hacer es:

Encontrar la desviación estándar

Encontrar la media

Establecer los límites superior e inferior

lim\_sup = media + n\*desv\_esta

lim\_inf = media - n\*desv\_esta

Y con estos, acotar nuestra población a estos parámetros y anotar cual es la nueva cantidad de población con la que contamos. Y graficar.

Método 3 por rango intercualtil (Grafico de Caja)

Existe un método donde solo se toma el primer cuartil (Q1 25%) y el tercer cuartil (Q3 75%) de la población y se obtiene esta diferencia para obtener el rango intercuantil (IQR = Q3 - Q1). Después tenemos que obtener el valor mínimo y el máximo que puede obtener ésta variable.

Valor\_min = Q1 - (1.5 \* IQR)

Valor\_max = Q3 + (1.5 \* IQR)

Y con estos, acotar nuestra población a estos parámetros y anotar cual es la nueva cantidad de población con la que contamos. Y graficar.

--------------------------------------------------------------------------------------------------------

Retirar Outlyers

Identificar por tres caminos distintos como se pueden quitar los outlyers y valores atipicos. Para despupes comparar entre el la cantidad de registros de la poblaciòn inicial vs la cantidad de registros despuès de la depuraciòn.

Esto solo es para variables continuas.

Mètodo 1. Por percentiles

Los percentiles muestran que tan concentrada està la poblaciòn por categorias. Entonces lo que se tiene que hacer es:

Observar que tan concentrada està la poblaciòn en las categorias y acotar èstas deacuerdo a un criterio de sensibilidad (felling)

Mètodo 2. Por la media

Bajo el supuesto de que todas las variables se comportan de una manera normal (Normal(Mu, sigma2)) se puede acotar mediante la desviacion tipica (raiz cuadrada de la varianza) por (generalmente) 3. Entonces lo que se tiene que hacer es:

Encontrar la desviacion estandar

Encontrar la media

Establecer los limites superior e inferior

lim\_sup = media + n\*desv\_esta

lim\_inf = media - n\*desv\_esta

Y con estos, acotar nuestra poblaciòn a estos paràmetros y anotar cual es la nueva cantidad de poblaciòn con la que contamos. Y gràficar.

Mètodo 3 por rango intercualtil (Grafico de Caja)

Existe un metodo donde solo se toma el primer cuantil (Q1 25%) y el tercer cuantil (Q3 75%) de la poblaciòn y se obtiene esta diferencia para obtener el rango intercuantil (IQR = Q3 - Q1). Despuès tenemos que obtener el valor minimo y el màximo que puede obtener èsta variable.

Valor\_min = Q1 - (1.5 \* IQR)

Valor\_max = Q3 + (1.5 \* IQR)

Y con estos, acotar nuestra poblaciòn a estos paràmetros y anotar cual es la nueva cantidad de poblaciòn con la que contamos. Y gràficar.

--------------------------------------------------------------------------------------------------------

Nulos, missing y vacìos

Normalizar una variable

matriz spars. Es una matriz con valores en su mayorìa, vacìos

¿Què es un valor asuente?

Es aquel que no almacena ningun valor de datos, es decir ausencia de datos

La falta de informaciòn tambièn es informaciòn.

La inputaciòn no es buena necesariamente

Mètodos de imputaciòn

Continuas

Imputaciòn de la media

Imputaciòn de la media

Imputaciòn de la moda

Transformaciòn

Predicciòn

Discretas

Imputaciòn de al mopda

Prediccòn

Transformaciòn

Transformaciòn entropica \*\*(from creditt scoring)

while of evidence

you can use with discrete and continuos

Pickle es un objeto de memorìa ram y que se puede cargar sin voñlverlos a contruir

--------------------------------------------------------------------------------------------------------

¿Por què reducir dimensiones?

\* Reduce el tiempo y el espacio de almacenamiento requerido

Maldiciòn de la dimensionalidad. Mucho no siempre es mejor.

Mètodos de reducciòn

Analisis de multicolinealdad

Analisis de componentes principales

Analisis de varizanza explicada

Importancia de variables

Poder predictivo de caracteristicas (transformaciòn entropica)

Transformaciòn entropica. solo es para variables binarias

--------------------------------------------------------------------------------------------------------

--------------------------------------------------------------------------------------------------------

--------------------------------------------------------------------------------------------------------

--------------------------------------------------------------------------------------------------------

--------------------------------------------------------------------------------------------------------

--------------------------------------------------------------------------------------------------------

--------------------------------------------------------------------------------------------------------

--------------------------------------------------------------------------------------------------------

--------------------------------------------------------------------------------------------------------

--------------------------------------------------------------------------------------------------------