Práctica 1

Análisis y Preprocesamiento de Datos

Objetivos

El objetivo de esta práctica es introducir al análisis y preprocesamiento de los datos.

Temas

- Atributos. Identificación, clasificación y análisis.
- Representaciones Gráficas. Diagrama de Barra, de Caja, de Dispersión e Histograma.
- Normalización y estandarización de Datos.

Lectura

Cap. 4 del Libro Introducción a la Minería de Datos de Hernández Orallo.

El archivo ***autos.csv*** contiene características de automóviles junto con su calificación de riesgo de seguro asignada. Dicha calificación corresponde al grado en que el coche es más arriesgado de lo que indica su precio. A los coches se les asigna inicialmente un símbolo de factor de riesgo asociado a su precio. Luego, si es más arriesgado (o menos), este símbolo se ajusta subiéndolo (o bajándolo) en la escala. Los actuarios llaman a este proceso "simbolización". Un valor de +3 indica que el automóvil es arriesgado, -3 que probablemente sea bastante seguro. https://archive.ics.uci.edu/ml/datasets/Automobile

Atributo	Descripción
symboling	Factor de riesgo: -3 (bastante seguro), -2, -1, 0, 1, 2, 3 (poco seguro).
normalized-losses	pérdidas normalizadas: numérico de 65 a 256.
make	marca: alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo
fuel-type	tipo de combustible: diesel, gasolina.
aspiration	aspiración: std, turbo.
num-of-doors	número de puertas: cuatro, dos.
body-style	tipo de carrocería: techo duro, wagon, sedán, hatchback, descapotable.
drive-wheels	tracción: 4x4, fwd, rwd.
engine-location	ubicación del motor: delantero, trasero.
wheel-base	distancia entre ejes: numérico de 86,6 a 120,9.
length	longitud: numérico de 141,1 a 208,1.
width	ancho: numérico de 60,3 a 72,3.
height	Altura: numérico de 47,8 a 59,8.
curb-weight	peso en vacío: numérico de 1488 a 4066.
engine-type	tipo de motor: dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
num-of-cylinders	número de cilindros: ocho, cinco, cuatro, seis, tres, doce, dos.
engine-size	tamaño del motor: numérico de 61 a 326.
fuel-system	sistema de combustible: 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
bore	diámetro: numérico de 2,54 a 3,94.
stroke	carrera: numérico de 2,07 a 4,17.
compression-ratio	relación de compresión: numérico de 7 a 23.
horsepower	potencia: numérico de 48 a 288.
peak-rpm	velocidad máxima: numérico de 4.150 a 6.600 rpm.
city-mpg	Rendimiento en ciudad (en millas por galón): numérico de 13 a 49.
highway-mpg	Rendimiento en ruta (en millas por galón): numérico de 16 a 54.
price	Precio en USD: numérico de 5118 a 45400.

Ejercicio 1

Complete el siguiente cuadro con la cantidad de atributos de cada tipo que contiene el archivo ***autos.csv.***

Tipo de atributo		Cantidad
Cuantitativo o numérico	*Discreto*	
	Continuo	
Cualitativo o categórico	*Nominal*	
	Ordinal	

Ejercicio 2

Enuncie dos tareas predictivas que pueda realizar a partir de los datos del archivo ***autos.csv***.

Ejercicio 3

Indique qué tipo de información brindan las siguientes representaciones gráficas:

a) Diagrama de Barras
b) Histograma
c) Diagrama de caja
d) Diagrama de dispersión

Realice al menos una de cada una de las representaciones anteriores utilizando los datos del archivo ***autos.csv*** y explique cómo interpretarlas.

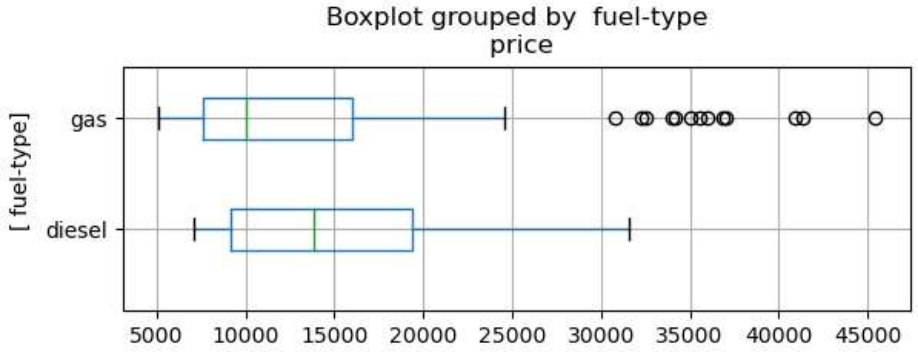
Ejercicio 4

Complete el siguiente cuadro y dibuje el diagrama de caja del atributo **"horsepower"**

Medida	Valor
Mínimo	
Máximo	
Q1	
Q2 o mediana	
Q3	
RIC	
Bigote superior	
Bigote inferior	
Intervalos de valores atípicos leves	
Valores atípicos leves	
Intervalos de valores atípicos extremos	
Valores atípicos extremos	

Ejercicio 5

Los valores del atributo **"price"** fueron agrupados según el tipo de combustible (atributo "fuel-type"). La figura muestra los diagramas de caja correspondientes.



Complete el siguiente cuadro y luego indique los valores de verdad de las afirmaciones utilizando únicamente los valores de las medidas calculados:

Medida	diesel	gas
Mínimo		
Máximo		
Q1		
Q2		
Q3		
RIC		
Bigote Inferior		
Bigote Superior		

a) Al menos el 25% de los autos que utilizan combustible diésel cuestan más de 19000 USD.
b) Es atípico encontrar un auto naftero (fuel-type=gas) que cueste menos de 7600 USD.
c) Utilizando únicamente la información del cuadro es posible afirmar que la mayoría de los vehículos utiliza nafta (fuel-type=gas).
d) Todos los valores de precio atípicos de los vehículos nafteros (fuel-type=gas) son leves.

Ejercicio 6

Calcule la correlación lineal entre los atributos **"curb-weight"** y **"highway-mpg"**. Indique la intensidad de la correlación (no hay correlación/débil/fuerte) y el tipo (positiva/negativa). Explique el significado del valor de correlación obtenido.

	curb-weight	highway-mpg
Valor		
Intensidad		
Tipo		
Significado		

Ejercicio 7

En base a los valores de la siguiente matriz de correlación indique el valor de verdad de las siguientes afirmaciones.

	wheel-base	length	width	bore	stroke	horsepower	peak-rpm	city-mpg	highway-mpg	price	symboling
wheel-base	1	0,87	0,80	0,49	0,16	0,35	-0,36	-0,47	-0,54	0,58	-0,53
length	0,87	1	0,84	0,61	0,13	0,56	-0,29	-0,67	-0,70	0,69	-0,36
width	0,80	0,84	1	0,56	0,18	0,64	-0,22	-0,64	-0,68	0,75	-0,23
bore	0,49	0,61	0,56	1	0	0,58	-0,26	-0,59	-0,59	0,54	-0,13
stroke	0,16	0,13	0,18	0	1	0,09	-0,07	-0,04	-0,04	0,08	-0,01
horsepower	0,35	0,56	0,64	0,58	0,09	1	0,13	-0,80	-0,77	0,81	0,07
peak-rpm	-0,36	-0,29	-0,22	-0,26	-0,07	0,13	1	-0,11	-0,05	-0,10	0,27
city-mpg	-0,47	-0,67	-0,64	-0,59	-0,04	-0,80	-0,11	1	0,97	-0,69	-0,04
highway-mpg	-0,54	-0,70	-0,68	-0,59	-0,04	-0,77	-0,05	0,97	1	-0,70	0,03
price	0,58	0,69	0,75	0,54	0,08	0,81	-0,10	-0,69	-0,70	1	-0,08
symboling	-0,53	-0,36	-0,23	-0,13	-0,01	0,07	0,27	-0,04	0,03	-0,08	1

a) Es de esperar que los autos más baratos (atributo “price”) tengan un menor rendimiento de combustible en ruta (atributo “highway-mpg”).
b) A mayor potencia de motor (atributo “horsepower”), mayor precio (atributo “price”).
c) Es posible que un auto con poca potencia (bajo valor del atributo “horsepower”) tenga un mayor rendimiento de combustible en ruta (atributo “highway-mpg”).
d) Los atributos “bore” y “stroke” son independientes.
e) Los elementos de la diagonal principal de la matriz de correlación siempre valen 1.
f) Los valores de la matriz son incorrectos porque el coeficiente de correlación lineal siempre es un valor positivo.
g) El valor del atributo “symboling” correspondiente al factor de riesgo del auto será menor en los autos de mayor precio.

Ejercicio 8

Discretice por rango el atributo **“engine-size”** en dos intervalos llamados **Chico** y **Grande**. Indique los los intervalos utilizados para discretizar, así como la cantidad de ejemplos que hay en cada intervalo.

	Chico	Grande
Intervalos		
Cantidad de Valores		

Ejercicio 9

Discretice por frecuencia el atributo **"engine-size"** en dos intervalos llamados **Chico** y **Grande**. Indique los intervalos utilizados para discretizar, así como la cantidad de ejemplos que hay en cada intervalo.

	Chico	Grande
Intervalos		
Cantidad de Valores		

Explique porque los ejemplos no quedaron divididos en dos intervalos con la misma cantidad de valores.

Ejercicio 10

Estudio de la calidad del Whisky. Una Asociación de Defensa de los Consumidores realizó un estudio de 35 marcas de whisky en venta en el mercado francés a fin de seleccionar aquellas marcas que presentan la relación "calidad/precio" más favorable para el consumidor final. Se puede verificar además el valor real de la "categoría" asignada a cada whisky la cual juega un rol importante en las campañas publicitarias de las diferentes marcas.

La calidad del whisky varía fundamentalmente en función de la proporción de malta que contiene y el tiempo de añejamiento. Es por eso que estos elementos son argumentos de publicidad de las marcas de categoría superior. Sin embargo, la modalidad del proceso de fabricación puede modificar la calidad de whisky. Por ello se tuvo en cuenta la opinión de catadores expertos.

Se observaron las 35 marcas de whisky que están en venta en almacenes que presentan puntos de venta en todo el país. El precio de una botella de whisky es el precio de venta que figura en la etiqueta del artículo en el punto de venta. La proporción de malta, el tiempo de añejamiento y la categoría del whisky son características que figuran obligatoriamente en la etiqueta de cada botella, en conformidad con la ley de comercialización de alcoholes. La nota de calidad global fue atribuida por un jurado que analizó cada whisky sin conocer la marca, ni el precio de venta, nilas características legales de los mismos.

La información recolectada es la que aprece en el archivo **Whisky.csv** donde

- Id_WHISKY: identificador de la botella de whisky. Es un entero entre 1 y 35.
- PRECIO: es el precio de una botella de whisky en francos.
- MALTA: es la graduación de malta en porcentaje
- CATEGORIA: categoría comercial del whisky
- AÑEJAMIENTO: Tiempo de añejamiento en meses
- CALIDAD: Calificación de cada whisky por un jurado de expertos catadores.

a) ¿Cuáles son las variables para analizar y cómo clasificaría a c/u?
b) Indique al menos dos formas de graficar la información de cada variable.
c) Plantee al menos un problema predictivo y otro descriptivo a resolver sobre este conjunto de ejemplos.
d) Complete los valores faltantes del atributo PRECIO utilizando la media. ¿Cuántos valores faltantes completó? ¿Qué valor utilizó para completar los valores de PRECIO faltantes?

e) Luego de haber completado los valores faltantes del atributo PRECIO utilizando la media, calcule los cuartiles del atributo PRECIO y utilícelos para indicar el valor de verdad de las siguientes afirmaciones:
   i. El 50% de las botellas de whisky cuesta a lo sumo 83 francos.
   ii. El atributo PRECIO posee dos valores atípicos leves.
   iii. Sería atípico encontrar una botella de whisky que cueste menos de 46,5 francos.
   iv. Se observa una mayor dispersión de precios entre los primeros dos cuartiles que entre el segundo y el tercero.

f) Realice el diagrama de barras correspondiente al atributo PRODUCTO e indique cual es el valor de la moda.

g) Complete la siguiente matriz de correlación e indique el valor de verdad de las siguientes afirmaciones:

	Precio	Malta	Añejamiento	Calidad
Precio				
Malta				
Añejamiento				
Calidad				

   i. No se observan valores de correlación negativos.
   ii. Sólo un par de atributos presenta un valor de correlación lineal fuerte.
   iii. Sólo un par de atributos presenta un valor de correlación lineal leve.
   iv. La correlación lineal entre los atributos “Precio” y “Añejamiento” es leve.

Ejercicio 11

Realice un análisis sobre los valores de los atributos del dataset AUTOS.CSV. Para cada atributo que no pueda ser procesado directamente, indique que problema tiene (valores nulos o vacíos, valores categóricos, valores atípicos o outliers, etc.) y como solucionarlo.

Ejercicio 12

Dada la siguiente tabla con mediciones de 2 características correspondientes a mediciones de altura y peso de personas:

Altura	1.65	1.81	1.70	1.62	1.74	1.70	1.80	1.73	1.68
Peso	75	86	82	78	77	87	90	83	80

a) Aplique las siguientes normalizaciones y grafíquelas con un diagrama de caja:

MinMax: $\frac{x_i - min(x)}{max(x) - min\,(x)}$ Standard: $\frac{x_i - media(x)}{stddev(x)}$ Robust: $\frac{x_i - Q1(x)}{Q3(x) - Q1(x)}$

b) Agregue la siguiente medición (2.20, 120) y repita el punto a)

c) Compare los diagramas de caja entre las normalizaciones de los puntos a) y b) y comente las diferencias.