PRACTICA 1 – PREPROCESAMIENTO DE LOS DATOS

Ejercicio a entregar: 4. Deberá respetarse el formato de tabla indicado para la resolución.

Material de Lectura: Capítulo 4 del Libro Introducción a la Minería de Datos de Hernández Orallo

1. Automóviles

Analice la información del archivo **automobile.csv** cuyo contenido se encuentra descripto en "automobile_names.txt"

a. Indique qué tipo de gráfica puede construir con los atributos. Ejemplifique cada caso.
b. Utilizando distintas representaciones gráficas, describa la distribución de los atributos, e indique si observa relaciones entre los mismos.
c. La Minería de Datos permite extraer dos tipos de conocimiento: descriptivo y predictivo. Ejemplifíquelos para el caso de los autos.
d. Calcule el coeficiente de correlación lineal entre los atributos numéricos. Relacione los valores obtenidos con los diagramas de dispersión de cada par de atributos.

2. Bicicletas

En el siguiente link encontrará información referida al uso de bicicletas que el Gobierno de la Ciudad de Buenos Aires pone a disposición de la población en forma gratuita como medio de transporte:

https://recursos-data.buenosaires.gob.ar/ckan2/bicicletas-publicas/recorrido-bicis-2016.csv

Estas bicicletas están ubicadas en distintos puntos de la ciudad y se encuentran disponibles las 24 horas del día durante todo el año. En el archivo encontrará información referida a las estaciones de origen y destino, la hora de partida y la duración de los viajes realizados por las bicicletas durante el año 2016.

a. A partir del atributo FECHA_HORA_RETIRO genere un atributo nuevo que contenga únicamente el horario en el cual la bicicleta fue retirada. Para ello puede obtener el substring que contiene la hora con las funciones de texto de **GenerateAttributes**, y luego convertir ese substring a un número entero con el mismo operador. También puede utilizar la función *date_parse* de **GenerateAttributes** para convertir el string a un tipo date, y luego utilizar *date_get* para obtener la hora.

b. Utilizando un histograma donde cada hora represente una barra, informe si hay horarios inusuales de retiro de bicicletas. Justifique su respuesta utilizando la frecuencia relativa de cada hora para decidir qué es un horario inusual (datos) y en qué horas tradicionalmente circula la gente por una ciudad (su conocimiento sobre el dominio)
c. Indique el valor de verdad de la siguiente proposición: “Se obtendrán los mismos resultados si se discretiza por rango el atributo generado en a) utilizando 4 intervalos que si se lo discretiza por frecuencia utilizando 4 intervalos”. Justifique su respuesta.

3. Estudiantes

El archivo ***estudiantes.csv*** contiene información de alumnos de dos colegios secundarios de Portugal relevada para analizar la adicción al alcohol en el nivel secundario y la forma en que éste afecta a los estudiantes tanto en sus relaciones como en sus estudios.

Atributo	**Descripción**
escuela	"GP" - Gabriel Pereira o "MS" - Mousinho da Silveira
Sexo	F o M
Edad	valor numérico entre 15 y 22
ambito	"Rural" o "Urbano"
grupo_familiar	Tamaño del grupo familiar (<=3 o >3 integrantes)
estado_padres	Si viven "juntos" o "separados"
educ_madre	Nivel educativo de la madre
educ_padre	Nivel educativo del padre
trabajo_madre	Tipo de trabajo de la madre
trabajo_padre	Tipo de trabajo del padre
razon_escuela	Motivos por los cuales eligió la escuela
Tutor	Indica quien es el tutor legal del estudiante
duracion_viaje	Tiempo que tarda en llegar a la escuela
tiempo_estudio	Cantidad de hs. Que estudia por semana
asignaturas_perdidas	Cant.de materias desaprobadas. Vale N si 0<=N<3 sino 3
Becas	Si tiene beca de estudios
apoyo_familiar	Si su familia lo ayuda económicamente
clases_particulares	Si toma clases particulares pagas
act_extracurriculares	Si realiza actividades extracurriculares

Atributo	**Descripción**
fue_guarderia	Si fue a guardería
quiere_educ_superior	Si piensa ir a la Universidad
internet_casa	Si posee internet en su casa
en_pareja	Si vive en pareja
rel_familia	Tipo de relación que tiene con sus familiares
tiempo_libre	Cantidad de tiempo libre que dispone
sale_con_amigos	Si sale con amigos
alcohol_diario	Cantidad de alcohol que consume diariamente
alcohol_semanal	Cantidad de alcohol que consume semanalmente
Salud	Estado de salud
ausencias	Cantidad de ausencias a la escuela
nota_1er_parcial	Calificación del 1er. Parcial
nota_2do_parcial	Calificación del 2do. Parcial
nota_final	Nota Final

A. Indique qué tipo de información brindan las siguientes representaciones gráficas:
    a. Diagrama de dispersión (scatter plot)
    b. Diagrama de caja (box plot)
    c. Histograma
    d. Diagrama de Barras

Realice al menos una de cada una de las representaciones anteriores utilizando la información del archivo ***estudiantes.csv*** y explique cómo interpretarlas.

B. Discretice el atributo EDAD en tres intervalos de dos formas distintas:
    a. Utilizando una discretización por rango (operador ***DiscretizeByBinning***)
    b. Utilizando discretización por frecuencia (operador ***DiscretizeByFrecuency***)
    c. Analice y compare los resultados obtenidos. Explique cómo se determina en cada caso los intervalos a utilizar.

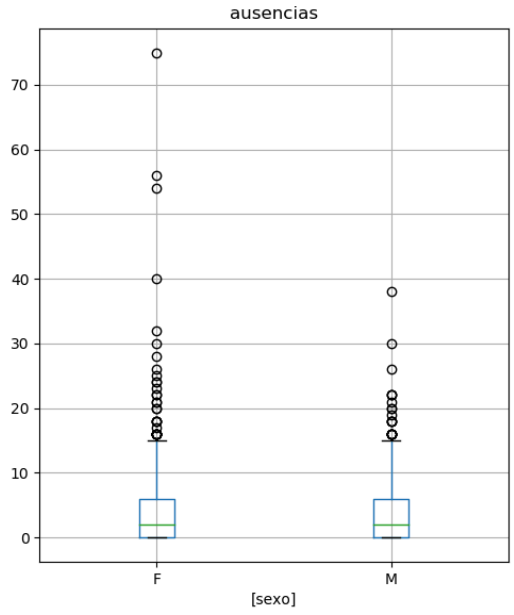
C. Se considera que un estudiante es mayor de edad si tiene al menos 18 años. Genere un nuevo atributo que tome el valor “SI” cuando la edad del estudiante sea mayor o igual a 18 años y “NO” en caso contrario. Luego grafique el resultado obtenido mediante un diagrama de barras.

D. Los atributos *EDUC_MADRE* y *EDUC_PADRE* indican la educación recibida por el padre y por la madre de cada estudiante, respectivamente. Los valores posibles son “ninguna”, “primaria (hasta 4to)”, “primaria (hasta 9no)”, “secundaria” y “universitaria”. Utilice el operador ***Map*** para mapear estos valores de la siguiente

forma: “ninguna” = 0, “primaria (hasta 4to)” = 1, “primaria (hasta 9no)” = 2, “secundaria” = 4 y “universitaria” = 6. Recuerde aplicar el operador ***Parse Number*** para convertir los valores de estos atributos en numéricos. Luego de la numerización, compare el nivel educativo de las madres con el de los padres.

E. En la figura más abajo pueden verse los Diagramas de Caja de Tukey correspondientes al atributo **Ausencias** del archivo **estudiantes.csv** separando los ejemplos por el atributo *sexo.* Calcule la mediana, los cuartiles Q1 y Q3, el rango intercuartil, y los intervalos en donde hay valores atípicos leves y extremos. Marcar estos valores en el gráfico.
Luego indique el valor de verdad de las siguientes afirmaciones. En caso de no poder hacerlo, justifique.
   a. Es atípico que un estudiante tenga más de 20 ausencias.
   b. Los cuartiles del atributo *AUSENCIAS* son los mismos para ambos sexos por lo que puede afirmarse que la cantidad de mujeres y varones con más de 6 ausencias coinciden.
   c. Al menos el 25% de las mujeres tiene asistencia perfecta.
   d. Es atípico encontrar un varón que no haya faltado nunca.
   e. La cantidad de mujeres con valores atípicos leves en el atributo AUSENCIAS es mayor que la de varones.

	F	*M*
Máximo		
3er.cuartil		
2do. cuartil		
1er. cuartil		
Mínimo		
Rango Intercuartil		
Atípicos leves		
Atípicos extremos		



d.

4. Estrellas

El archivo **estrellas2023.csv** contiene información sobre estrellas de una zona del espacio previamente inexplorada. Utilizando este archivo, realice las siguientes operaciones. Incluya en su respuesta los cálculos realizados.

A. Discretice por frecuencia el atributo **Edad** en dos intervalos llamados **Baja** y **Alta**. Indique los rangos de los dos intervalos resultantes, así como la cantidad de ejemplos que hay en cada intervalo.

	Baja	Alta
Intervalos		
Valores		

B. Discretice por rango el atributo **Edad** en dos intervalos llamados **Baja** y **Alta**. Indique los rangos de los dos intervalos resultantes, así como la cantidad de ejemplos que hay en cada intervalo.

	Baja	Alta
Intervalos		
Valores		

C. Calcule la correlación lineal entre los atributos **Edad** y **Temperatura**. Indique la intensidad de la correlación (no hay correlación/débil/fuerte) y el tipo (positiva/negativa)

Valor	
Intensidad	
Tipo	

D. Dibuje un Diagrama de Caja de Tukey de la variable **Edad** e inclúyalo en la respuesta. Indique también los valores del cuadro:

Mediana	
Q1	
Q3	
RI	
Bigote superior:	
Bigote inferior:	
Intervalos de valores atípicos leves	
Valores atípicos leves	
Intervalos de valores atípicos extremos	
Valores atípicos extremos	

5. Ejercicio integrador: Ofertas de trabajo

El archivo **trabajos.xlsx** contiene ejemplos de ofertas de trabajo de una plataforma online. Para poder utilizarlo en un proceso de minería de datos se requiere un **extenso** proceso de limpieza, como con cualquier conjunto de datos recolectados de sistemas existentes. En particular, hay muchos atributos con información codificada como texto, algo muy común en cualquier conjunto de datos.

A. Mire con detenimiento el conjunto de datos, en particular, los valores de los atributos “Job Title” y “Job Description”.
B. Elimine atributos que no aportan información: : index, Competitors, Company Name
C. Genere a partir de los existentes los siguientes atributos:
    a. Sueldo mínimo: codificado como número
    b. Sueldo máximo: codificado como número

c. Sueldo promedio: codificado como número
d. Ubicación: Del trabajo, codificada como el código del estado de dos letras (por ej, NY o MA), sin la ciudad. Quitar el resto de los atributos de ubicación (Location y Headquarters).
e. Antigüedad de la compañía, codificada como número, en base a la Fecha de fundación. Quitar la fecha de fundación.
f. Senior: Si el trabajo es para una persona con mucha experiencia (atributo binario)
   Nota: Para identificar si un trabajo es "Senior", consultar si los atributos "Job Title" o "Job Description" contienen las palabras: "Senior", "Sr", "Sr.", ¿Qué otras palabras se te ocurren para mejorar este atributo?
g. Junior: si el trabajo es para una persona con poca experiencia (atributo binario).
   Nota: Para identificar si un trabajo es "Junior", consultar si los atributos "Job Title" o "Job Description" contienen las palabras: "Junior" o "Jr". ¿Qué otras palabras se te ocurren para mejorar este atributo?
h. "Public" y "Private", atributos binarios en base al atributo "Type of ownership". Eliminar ese atributo.
i. "Estimated size" y "Estimated revenue", en base a "Size" y "Revenue", pero los nuevos atributos deben ser numéricos, y deben corresponder al máximo de los valores que aparecen de forma textual en el atributo. Por ejemplo, si el valor textual es *100 to 500 million (USD)*, el valor numérico debería ser *500.000.000*. En el caso de empresas que no tienen este valor, reemplazar por el promedio.
j. Los títulos de los trabajos contienen mucha variación y en algunos casos palabras que no contribuyen a la descripción o son muy específicas. Cree 10 atributos binarios con las palabras o frases clave más comunes, como "Data", "Engineer" o "Scientist", y elimine el atributo "Job Title".
k. Las descripciones de los trabajos están en lenguaje natural y es difícil utilizarlas en un proceso de minería. Identificar 10 habilidades muy ocurrentes que se mencionan en las descripciones y crear un atributo binario por cada una.
l. Convierta los valores del atributo "Industry" que tengan menos de 20 ocurrencias a "Other".

D. Elimine los atributos "Job Title" y "Job Description"
E. Elimine ejemplos con valores faltantes.
F. Elimine ejemplos repetidos. ¿Por qué es conveniente hacer este paso ahora y no antes?
G. Experimente y elija 3 gráficos complementarios distintos que le parezca que puedan describir el conjunto de datos de forma visual y resumida.