Tarea 08

Nombres y código:

Ruiz Ortiz Cesar Adrián 208020817

Padilla Martin Juan Pablo 217294261

Álvarez Gutiérrez David Alejandro 217294016

Sección: D01

Materia: MINERIA DE DATOS

Profesor: ISRAEL ROMAN GODINEZ

CARRERA: INGENIERIA EN INFORMATICA

CICLO ESCOLAR: 2020 A

Descripción del problema

Instrucciones: Elabore un documento de texto (*.docx) donde se muestre el desarrollo de las actividades de limpieza de datos. Para ello, utilice las técnicas mostradas durante la clase y apóyese con las diapositivas que están en la plataforma. En los tres primeros casos, pueden usar herramientas como Excel u Orange DataMining.

1. Determine cuál es la mejor estrategia de substitución de valores faltantes (media, mediana, algoritmo de aprendizaje) para los atributos del conjunto de datos y realícela (explique cada una de las decisiones de substitución).

Pronóstico: La mejor estrategia, es usar la moda, ya que, para los atributos categóricos, es la mejor estrategia.

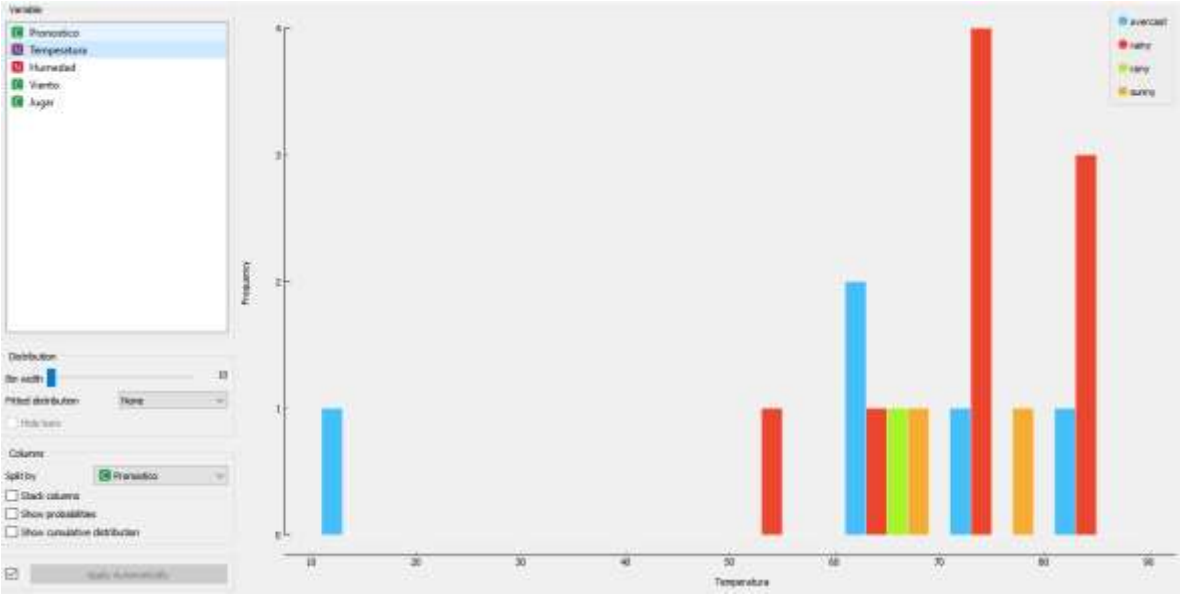
Temperatura: Como se tiene una distribución sesgada, la mejor estrategia es utilizar la mediana.

Humedad: Como se tiene una distribución sesgada, la mejor estrategia es utilizar la mediana.

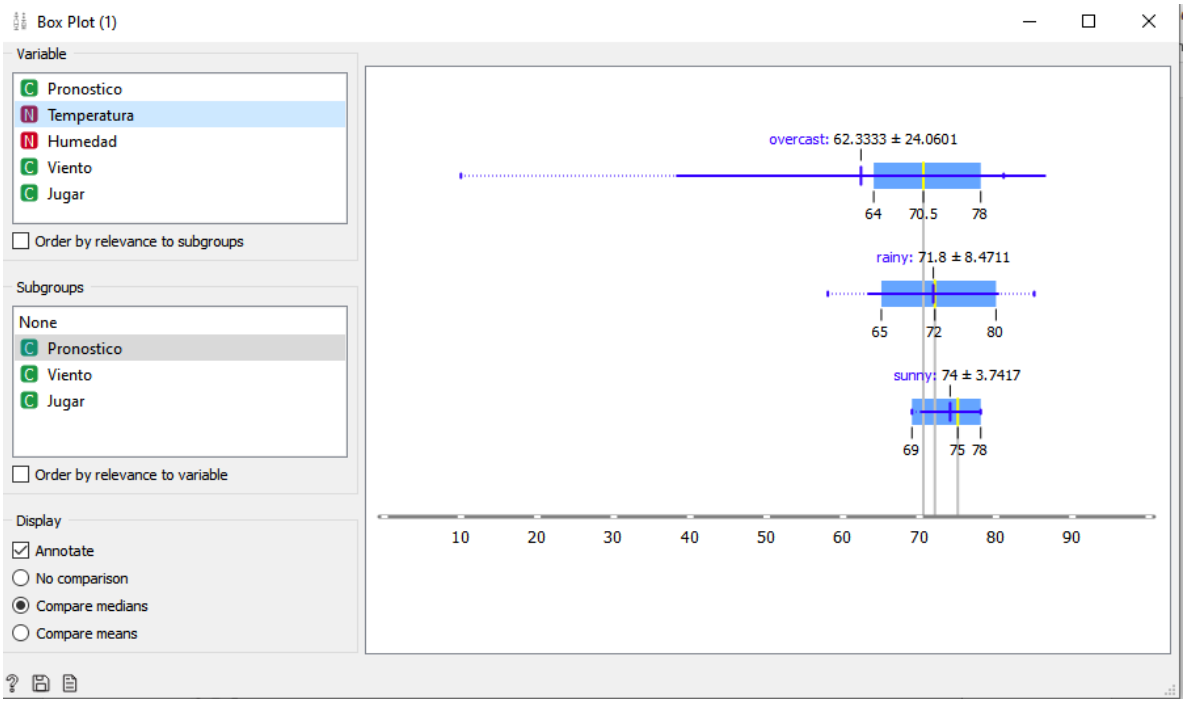
Viento: La mejor estrategia, es usar la moda, ya que, para los atributos categóricos, es la mejor estrategia.

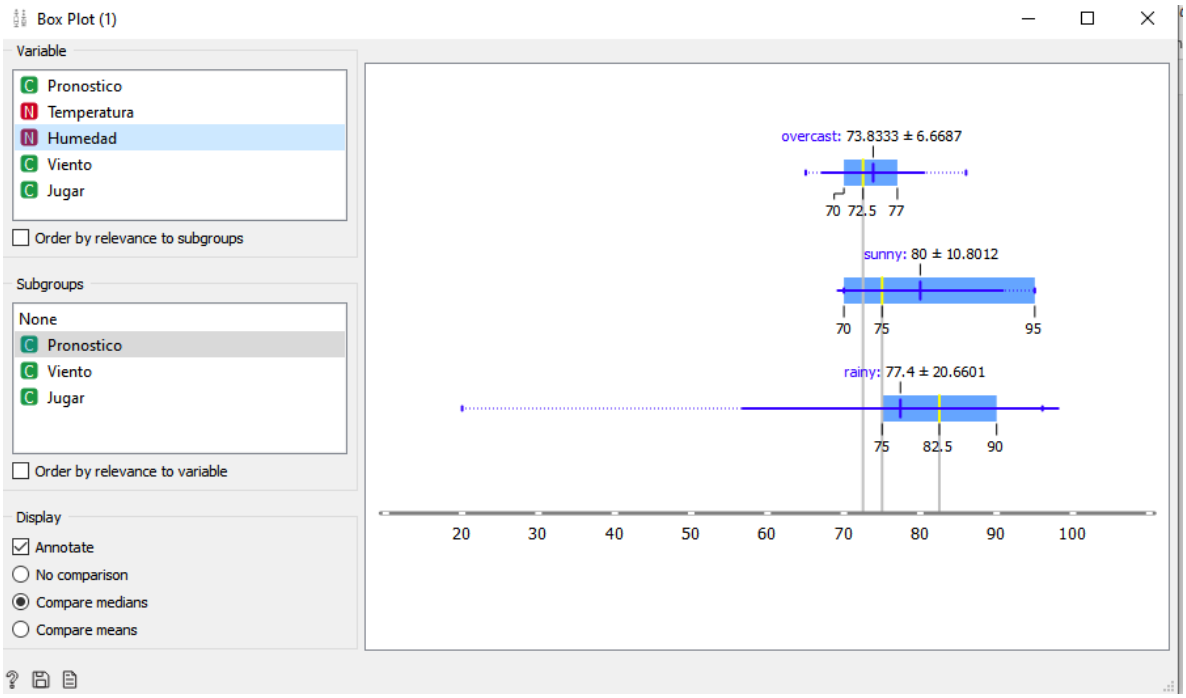
Jugar: La mejor estrategia, es usar la moda, ya que, para los atributos categóricos, es la mejor estrategia.

2. Identifique los outliers que existen en el documento, indicando su número de instancia, y determine cuál es la mejor estrategia de sustitución o eliminación del valor. Para la determinación del outlier utilice la herramienta de box plot.

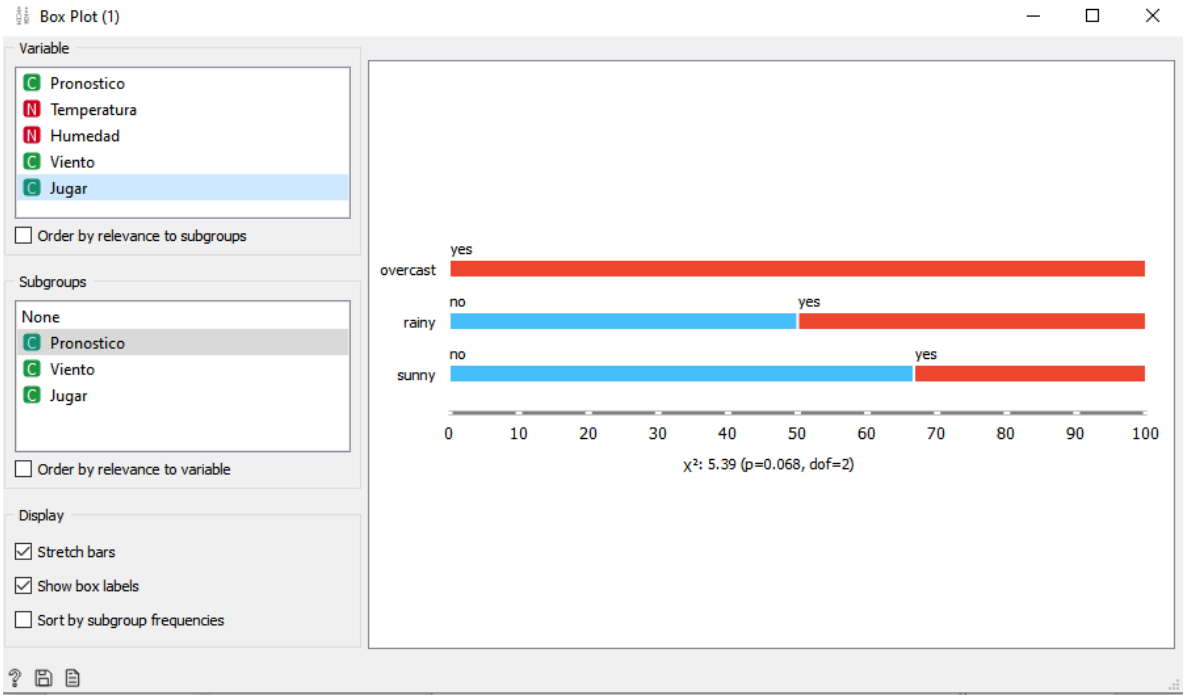


Podemos ver que, con los valores faltantes, mal escritos y outliers, queda de esta forma, por ejemplo, el atributo temperatura.




Box Plot (1)
Variable
Pronostico
Temperatura
Humedad
Viento
Jugar
Order by relevance to subgroups
Subgroups
None
Pronostico
Viento
Jugar
Order by relevance to variable
Display
Annotate
No comparison
Compare medians
Compare means
overcast: 73.8333 ± 6.6687
70 72.5 77
sunny: 80 ± 10.8012
70 75 95
rainy: 77.4 ± 20.6601
75 82.5 90
20 30 40 50 60 70 80 90 100


Box Plot (1)
Variable
Pronostico
Temperatura
Humedad
Viento
Jugar
Order by relevance to subgroups
Subgroups
None
Pronostico
Viento
Jugar
Order by relevance to variable
Display
Stretch bars
Show box labels
Sort by subgroup frequencies
FALSE
TRUE
overcast
FALSE
TRUE
rainy
TRUE
sunny
0 10 20 30 40 50 60 70 80 90 100
X²: 2.59 (p=0.274, dof=2)



3. Identifique si en la tabla 2 existe algún falso predictor. En caso de que exista, indique cuál es.

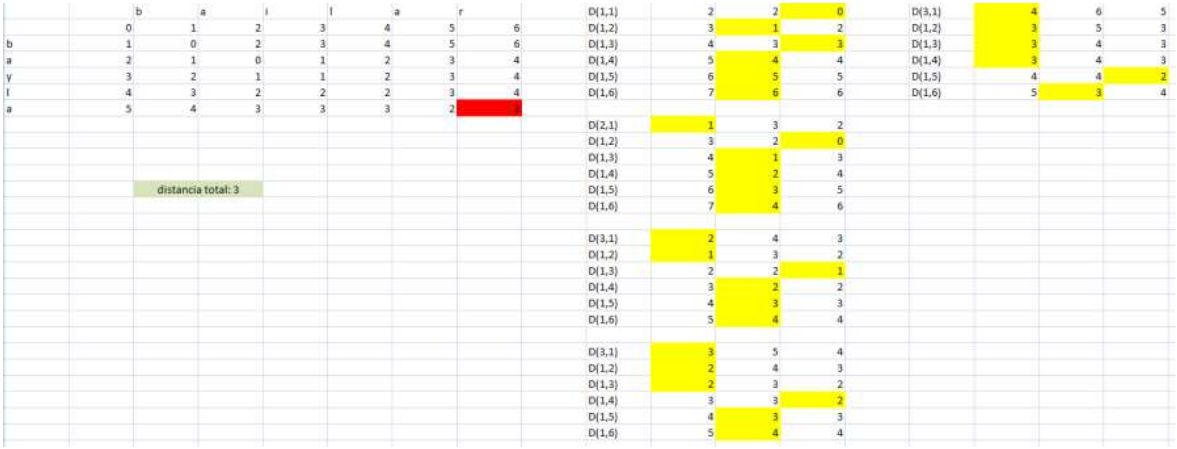
En este caso no existe correlación entre la clase y los atributos categóricos (jugar y viento), por lo que no podrían ser falsos predictores. Esto lo calculamos usando pares (pares, clase), para identificar la correlación. También podemos usar la correlación chi-cuadrada, ya que son atributos categóricos

4. Calcule la distancia de Levenshtain para el siguiente par de palabras. Para ello

desarrolle el proceso del cálculo de la matriz utilizando papel, lápiz y calculadora.

D(correr, corriendo)

D(bailar, bayla)



D(ACGTTTGCA, ATGCA)