Bases de Datos Masivas (11088)
Departamento de Ciencias Básicas

TRABAJO PRÁCTICO II: Preprocesamiento de datos

-Análisis, Limpieza, Transformación e Integración-

Introducción:

En este trabajo se abordan cuestiones relacionadas con las tareas de Preprocesamiento de datos previo a la etapa de descubrimiento del conocimiento. Entre las tareas que se abordan, se encuentra la integración, limpieza, selección y transformación de variables así como técnicas de reducción de dimensionalidad de un dataset, a efectos de reconocer aquellos atributos y escalas que mejor lo representan.

Se plantean ejercicios y datasets cuyas resoluciones serán realizadas mediante el lenguaje R.

Limpieza de datos:

1. *Datos faltantes.* Se cuenta con el dataset *encuesta_universitaria*.csv, el cual posee valores faltantes para la variable *tiempo_traslado*. Aplique los siguientes métodos a efectos de reemplazar esos valores:
    a. Sustituya los valores faltantes por la media encontrada para el atributo.
    b. Sustituya los valores faltantes de acuerdo al método de "hot deck imputation" descrita en el texto "Data mining and the impact of missing data.".
    c. Analice los resultados encontrados a partir de la aplicación de los métodos anteriores. Compare los mismos realizando gráficos sobre los valores resultantes en cada caso.
2. *Manejo de Ruido.* Para el dataset anterior, avance sobre las siguientes operaciones para los atributos numéricos (cuantitativos continuos):
    a. Verifique en primer lugar la distribución de los datos, utilice algún método gráfico para esto.

b. Realice un suavizado utilizando *binning* por *frecuencias iguales* y estime el valor del Bin por el cálculo de medias. Grafique las dos series resultantes y comente los resultados observados.
c. Utilizando suavizado por medias o medianas (elija!) calcular los bins con *anchos iguales* de 2 a 10 y compare los resultados gráficamente. ¿Qué ocurre conforme el bin aumenta?

3. *Detección de outliers.* Ahora, trabaje sobre el mismo atributo del dataset original con las siguientes consignas:
   a. Verifique la existencia de *outliers* en cada uno de los atributos. ¿Existen atributos que poseen valores anómalos?
   b. Seleccione uno de los *features* del dataset que a su entender posea *outiers* y aplique las técnicas de análisis y detección vistas en clase (IRQ, SD y LOF).
   c. Realice un análisis en torno a la diferencia de utilizar las diferentes técnicas, que implicancias tienen en la nueva distribución del dato (en caso que se opte por eliminar los valores anómalos) e indague sobre los valores categorizados como *outliers* por cada una de las técnicas. Concluya al respecto.

Reducción de dimensionalidad:

4. A partir del dataset *auto-mpg.data-original.txt*[1], se solicita trabajar sobre las siguientes consignas:
   a. Indague sobre la varianza[2] de cada uno de los atributos que conforman el dataset. ¿Existen atributos que podrían ser eliminados de acuerdo a la técnica de *Low Variance Factor*? Actúe en consecuencia.
   b. Evalúe la relación entre atributos a partir del coeficiente de correlación de Pearson y un análisis gráfico de heatmap[3] para estudiar la posibilidad de eliminar redundancia en el dataset. En caso de corresponder, aplique las técnicas de Reducing Highly Correlated Columns trabajadas en clase.

[1] Disponible en: https://archive.ics.uci.edu/ml/datasets/Auto+MPG
[2] Recuerde previamente normalizar el dataset, consulte la instrucción *scale*().
[3] Explore la instrucción *heatmap.2* de la librería *gplots.*

c. Por último, compare la importancia de cada uno de los atributos en función de la técnica de determinación de *Random Forest*[4] (suponiendo que intenta predecir la cantidad de cilindros de un auto). Analice la importancia de las variables de modo analítico y gráfico.

5. *Análisis de Componentes Principales.* Cargue en R el dataset *europa.dat* y conteste las siguientes consignas a través de las funcionalidades provistas por esa herramienta:
   a. Calcule la matriz de covarianzas. Que nos indica la misma sobre los atributos del dataset?
   b. Realice ahora el análisis de componentes principales. ¿Cuánto explica de la variación total del dataset la primera componente? ¿Y si se incorpora la segunda? ¿Y el primer auto-valor?
   c. Grafique el perfil de variación de las componentes en un gráfico de dispersión donde las X es la componente y la Y la varianza.
   d. Analice la matriz de loading. ¿Qué información provee? ¿Qué variables están más correlacionadas con la primera componente?
   e. Genere un gráfico de biplot y explique brevemente que información le provee el mismo.
   f. En función de los análisis realizados en los puntos anteriores. ¿Cuántas componentes principales elegiría para explicar el comportamiento del dataset? Justifique esa cantidad.

Transformación de datos:

6. *Discretización.* A partir del dataset *encuesta_universitaria.csv*, opere sobre el atributo *tiempo_traslado* de la siguiente manera:
   a. Transforme el atributo a discreto, definiendo 5 rangos de acuerdo al análisis de frecuencia de los valores encontrados para el atributo. ¿Qué tipo de variable se obtiene?
   b. Transforme el atributo a discreto, definiendo 5 rangos utilizando intervalos de clases.

[4] Se sugiere utilizar las instrucciones *randomForest, importance* y *varImpPlot* de la librería randomForest.

c. Analice los resultados encontrados. Compare los mismos realizando gráficos de frecuencia sobre los intervalos resultantes en cada caso. ¿Qué conclusiones se pueden obtener en términos del balanceo de las mismas de acuerdo a la técnica utilizada?

7. *Normalización.* Trabaje sobre las siguientes consignas:
    a. Entrada en calor: Para el siguiente dominio de edades E = {2, 3, 8, 17, 34, 89, 99} ¿Cuál sería el valor de *escalado decimal* para 65? ¿Y si utilizo *mínimo-máximo*?
    b. A partir del dataset *encuesta_universitaria.csv*, opere sobre el atributo *tiempo_traslado* de la siguiente manera:
        i. Normalice el atributo utilizando la técnica de mínimo-máximo. Generalizar la operación para cualquier atributo numérico.
        ii. Ahora, normalice el atributo mediante la técnica de z-score propuesta en el libro "Data Mining. Concepts & Techniques de Jiawei Han & otros".
        iii. Por último, utilice la técnica de escalado decimal para llevar adelante la tarea de normalización.
    c. Analice los resultados encontrados. Compare los mismos realizando gráficos sobre los atributos resultantes en cada caso.

Referencias sugeridas:

Principal component analysis. Hervé Adbi & otros. 2010.

Data mining and the impact of missing data. Marvin L. Brown & otros. 2003.

Data Mining. Concepts & Techniques. Jiawei Han and Micheline Kamber. 2006.