MINERIA DE DATOS USANDO SISTEMAS INTELIGENTES

PRACTICA 1 – PREPROCESAMIENTO DE LOS DATOS

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

El archivo **Curso.xls** contiene información de los alumnos de un curso. Para cada alumno se ha registrado: la calificación obtenida en la actividad **Practica**, la proporción de instancias en las que participó a distancia (**Activ_Distancia**), la proporción de clases a las que asistió **(Activ_Presencial**), si **Trabaja** o no, si asistió a uno de los colegios de la Universidad (**Colegio_UNLP**) y la **Calificacion**que obtuvo.

1. Indique qué tipo de información brindan las siguientes representaciones gráficas:
    1. Diagrama de dispersión (scatter plot)
    2. Diagrama de caja (box plot)
    3. Histograma
    4. Diagrama de Barras

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

2. En la Figura 1 se pueden ver los diagramas de clase correspondientes al atributo **Practica** separando los alumnos que aprobaron el curso de los que no lo hicieron (atributo **Calificacion**). Utilícelos, en caso de ser posible, para indicar el valor de verdad de las siguientes afirmaciones. Si no es posible, justifique.
    a) La mayoría de los alumnos aprobaron el curso.
    b) Conocer el valor que un alumno ha obtenido en el atributo **Practica** no alcanza para determinar la **Calificacion** que obtuvo en el curso.
    c) Al menos el 50% de los alumnos que aprobaron el curso obtuvo una nota de **Practica** superior a 6 puntos.
    d) Al menos el 50% de los desaprobados obtuvo una nota de **Practica** inferior a 4 puntos.
    e) Todos los alumnos que obtuvieron 8 o más como nota de **Practica** aprobaron el curso.
    f) Es raro que un alumno que aprobó el curso obtenga una nota de **Practica** inferior a 1 punto.

3. Abra el archivo **Curso.xls** y visualice los metadatos.
    a) Indique cuáles son los atributos que presentan datos faltantes.
    b) ¿Puede indicar con sólo observar los metadatos cuántos alumnos poseen la información correspondiente a los 6 atributos completa? Es decir, ¿cuántos alumnos NO presentan datos faltantes?

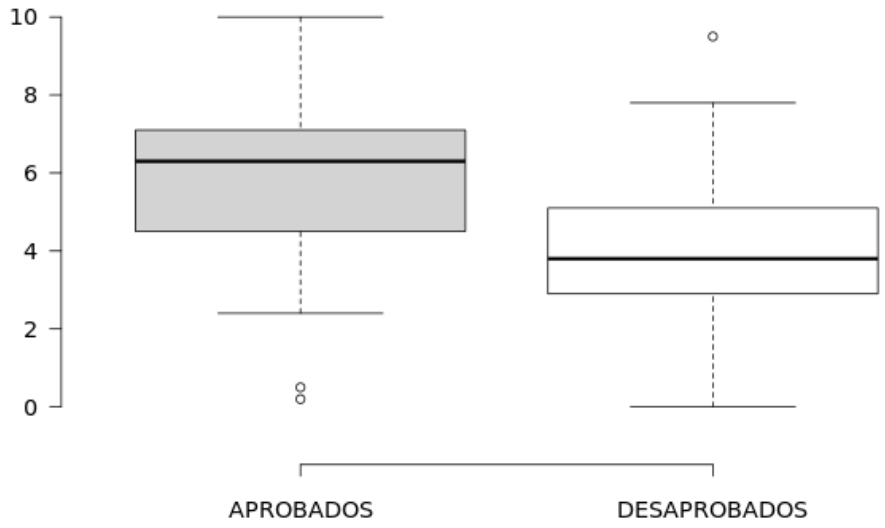

Figura 1. Diagramas de caja correspondientes al atributo PRACTICA separando los alumnos según el atributo CALIFICACION.

c) ¿Por qué es necesario completar los datos faltantes? Mencione al menos tres formas de realizar esta tarea. Indique en cada caso como influye en la distribución de valores del atributo.

d) Utilice el operador ***ReplaceMissingValues*** para completar los datos faltantes utilizando las opciones por defecto para todos los atributos salvo **Activ_Distancia** que deber completarse con el valor mínimo. Ejecute el proceso y verifique que ya no quedan datos faltantes. Indique cuales han sido los valores utilizados para completar cada atributo.

4. El atributo **Activ_Presencial** debe contener valores entre 0 y 1 indicando la proporción entre la cantidad de clases a las que asistió el alumno y el total de clases del curso. Ej: un valor 0.1 indica que el alumno asistió al 10% de las clases. Utilice un diagrama de caja para verificar que existen valores excesivamente fuera de rango en el atributo **Activ_Presencial**.Utilice el operador ***GenerateAttribute*** para generar un nuevo atributo llamado **ASISTENCIA** que posea los mismos valores que **Activ_Presencial** cuando éste posea un valor menor o igual a 1 y **Activ_Presencial/100** cuando supera 1. Rehaga el diagrama de caja para el atributo **ASISTENCIA** y verifique que no existen valores fuera de rango extremos.
5. El atributo **Trabaja** presenta errores en su codificación. Utilice el operador ***Map*** para convertir los valores "S" en "si" y "N" en "no".
6. Utilice un mismo operador ***DiscretizeByUserSpecification*** para discretizar los valores de los atributos **ASISTENCIA** y **Activ_Distancia** asignando la etiqueta BAJA o ALTA según si el valor del atributo es menor o igual a 0.5 o no respectivamente.

7. Utilice otro operador ***DiscretizeByBinning*** para discretizar los valores del atributo **Practica** dividiendo su rango en tres intervalos. Observe los metadatos e indique como quedaron formados los intervalos y cuántos alumnos hay en cada uno de ellos.
8. ¿Qué diferencia hay entre una discretización por intervalos (***DiscretizeByBinning)*** y una discretización por frecuencia ***(DiscretizeByFrecuency)***?

   Ejemplifique nuevamente su respuesta utilizando la información del archivo Curso.xls

9. 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. Luego, utilizando el diagrama de caja visto en clase, informe si hay horarios inusuales (fuera de rango) de retiro de bicicletas. Justifique su respuesta indicando los valores de los cuartiles y el criterio utilizado para decidir qué es un horario inusual.
   b. 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.
   c. A partir del atributo FECHA_HORA_RETIRO genere un segundo atributo con el número de mes en el cual la bicicleta fue retirada. Grafique manualmente el histograma correspondiente a este atributo utilizando 3 intervalos.

10. Analice la información del archivo **Sopas.xls** cuyo contenido se encuentra descripto en "Caso de Estudio 2: Sopas.pdf"
    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 las Sopas.
    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.