Universidad Central de Venezuela

Facultad de Ciencias

Escuela de Computación

Materia: Minería de Datos

Alumno: Jefferson Santiago

Tarea #1: Preprocesamiento de datos.

En el siguiente texto se presenta una explicación de las decisiones tomadas para realizar la vista minable.

En principio el proceso se centro en remover columnas para disminuir la dimensionalidad del problema.

1- **Estandarización de columnas / transformación de los tipos de datos iniciales / imputación o transformación de instancias:** Se estandarizaron o transformaron tipos de datos de columnas como:
   - **Periodo académico a renovar:** Se estandarizó la columna utilizando expresiones regulares, entre otras cosas. De manera que los registros solo tengan valores del tipo (primer semestre, segundo semestre del año en números romanos)-(año), ejemplo: II-2015, I-2014.
   - **Fecha de nacimiento:** Se estandarizó esta columna con el fin de que todos los registros queden estrictamente con un formato de fecha estándar, tal como: dd/mm/aaaa.
   - **Edad:** Para estandarizar esta columna lo que se hizo fue sólo retirar la palabra “AÑOS” de algunos registros. Con el fin de que cada registro sólo tenga un valor numérico.
   - **Estado civil:** Los registros de esta columna se numerizaron. Para la numerización 1 a 1, se renombraron los valores de la siguiente manera:
     - El valor 0 para **Soltero (a).**
     - El valor 1 para **Casado (a).**
     - El valor 2 para **Unido (a).**
     - El valor 3 para **Viudo (a).**

- **Sexo:** Los registros de esta columna se numerizaron. Para la numerización 1 a 1, se renombraron los valores de la siguiente manera:
  - El valor 0 para **Masculino.**
  - El valor 1 para **Femenino.**
- **Escuela:** Los registros de esta columna se numerizaron. Para la numerización 1 a 1, se renombraron los valores de la siguiente manera:
  - El valor 0 para **Enfermería.**
  - El valor 1 para **Bioanálisi.**
- **Modalidad de ingreso:** Los registros de esta columna se numerizaron. Para la numerización 1 a 1, se renombraron los valores de la siguiente manera:
  - El valor 0 para **los ingresados por OPSU.**
  - El valor 1 para **los ingresados por prueba interna.**
  - El valor 2 para **los ingresados por convenios internos.**
- **Semestre que cursa:** Para estandarizar esta columna se tomaron solo los números en los registros que indicaban el semestre que cursa el estudiante.
- **Ha cambiado de dirección:** Esta columna se unió con su consecuente "**De ser afirmativo indique el motivo**". De este modo, si la persona ha cambiado de dirección se coloca el valor del registro de la columna para indicar el motivo, en el registro pertinente de la columna que indica si ha cambiado de dirección o no. En caso de que no haya cambiado de dirección se colocó un cero(0).
- **Numero de materias aprobadas en el semestre o año anterior:** Para estandarizar esta columna solo se tomaron los enteros de cada registro con ayuda de la librería "re" para usar expresiones regulares en python.
- **Promedio ponderado aprobado:** En el proceso de estandarización de esta columna se convirtieron los valores para que todos tengan un estándar del tipo "00.000", ejemplo: 12.092 , 13.232. Concatenando en la tercera posición del registro un punto (.) en caso de que lo necesite. En algunos casos se eliminaros comas(,) que representaban el delimitador para los "miles" y en otros casos las comas se intercambiaron por puntos para delimitar los decimales.
- **Eficiencia:** La estandarización de esta columna se llevó a cabo de la siguiente manera: se concatenó al inicio de cada registro el string "0," en caso de que el valor del registro no fuera 1.
- **Estás realizando TEG o pasantías de grado:** Esta columna se unió con su consecuente "**Tesis, trabajo de grado o pasantías de grado**", en este caso y para los registros pertenecientes a ésa columna, los registros son llenados con el número en veces que se ha realizado TEG, ejemplo: primera vez, segunda vez. Por ende, si la persona ha realizado TEG o pasantías de grado se coloca el valor del registro de la columna para indicar "**las veces**", en el registro pertinente de la columna que indica si ha realizado TEG o pasantías de grado. En caso de que no haya realizando TEG se colocó un cero(0).

- **Ha solicitado algún beneficio a la universidad u otra institución:** Esta columna se unió con su consecuente "**En caso afirmativo, señale el año de la solicitud y motivo**". Si en la columna que dice si se realizo alguna solicitud de algún beneficio se introduce el valor en la celda respectiva que tiene la celda de la columna consecuente. En caso de que no haya realizando la solicitud se colocó un cero(0).
- **Se encuentra usted realizando alguna actividad que le genere ingresos:** Esta columna se unió con su consecuente "**En caso afirmativo, indique el tipo de actividad y su frecuencia**". Si en la columna que dice si se realiza algún actividad que genere ingresos se introduce el valor en la celda respectiva que tiene la celda de la columna consecuente. En caso de que no haya realizando la solicitud se colocó un cero(0).
- **Aporte mensual que brinda su responsable económico:** Se verificó que cada registro sea un número, en caso contrario se colocó un cero en registro correspondiente.
- **Aporte mensual que recibe de familiares o amigos:** Se verificó que cada registro sea un número, en caso contrario se colocó un cero en registro correspondiente.
- **Ingreso mensual que recibe por actividades a destajo o por horas:** Se verificó que cada registro sea un número, en caso contrario se colocó un cero en registro correspondiente.
- **Alimentación:** Se verificó que cada registro sea un número, en caso contrario se colocó un cero en registro correspondiente.
- **Transporte público:** Se verificó que cada registro sea un número, en caso contrario se colocó un cero en registro correspondiente.
- **Gastos médicos:** Se verificó que cada registro sea un número, en caso contrario se colocó un cero en registro correspondiente.
- **Gastos odontológicos:** Se verificó que cada registro sea un número, en caso contrario se colocó un cero en registro correspondiente.
- **Gastos personales:** Se verificó que cada registro sea un número, en caso contrario se colocó un cero en registro correspondiente.
- **Recreación:** Se verificó que cada registro sea un número, en caso contrario se colocó un cero en registro correspondiente.
- **Otros gastos:** Se verificó que cada registro sea un número, en caso contrario se colocó un cero en registro correspondiente.
- **Ingreso mensual de su responsable económico:** Esta columna se estandarizó de la siguiente manera: los valores en los registros con una especie de formato donde no exista punto o coma para delimitar los miles o millones, etc. En donde además el carácter para delimitar los enteros de los decimales es el punto ".", se dejaron igual. Por ende, todos los demás formatos de montos se cambiaron para que quede un formato como el especificado anterior mente, ejemplo: 12345.23, 33321.423.

- **Oros ingresos:** Para este caso se tomaron los valores numéricos de todos los registros. En los registros vacios se colocó un cero(0).
- **Vivienda:** Para este caso se tomaron los valores numéricos de todos los registros. En caso de no haber valores numéricos o si los registros se encuentran vacios se colocó un cero(0).
- **Alimentación:** Para este caso se tomaron los valores numéricos de todos los registros. En caso de no haber valores numéricos o si los registros se encuentran vacios se colocó un cero(0).
- **Transporte:** Para este caso se tomaron los valores numéricos de todos los registros. En caso de no haber valores numéricos o si los registros se encuentran vacios se colocó un cero(0).
- **Gastos médicos:** Para este caso se tomaron los valores numéricos de todos los registros. En caso de no haber valores numéricos o si los registros se encuentran vacios se colocó un cero(0).
- **Gastos odontológicos:** Para este caso se tomaron los valores numéricos de todos los registros. En caso de no haber valores numéricos o si los registros se encuentran vacios se colocó un cero(0).
- **Gastos educativos:** Para este caso se tomaron los valores numéricos de todos los registros. En caso de no haber valores numéricos o si los registros se encuentran vacios se colocó un cero(0).
- **Servicio publico:** Para este caso se tomaron los valores numéricos de todos los registros. En caso de no haber valores numéricos o si los registros se encuentran vacios se colocó un cero(0).
- **Condominio:** Para este caso se tomaron los valores numéricos de todos los registros. En caso de no haber valores numéricos o si los registros se encuentran vacios se colocó un cero(0).
- **Otros gastos:** Para este caso se tomaron los valores numéricos de todos los registros. En caso de no haber valores numéricos o si los registros se encuentran vacios se colocó un cero(0).

2- Eliminación de columnas pertinentes las cuales representaban redundancia o que a mi criterio no aportaban datos de interés para algún estudio, también se eliminaron columnas que sus registros fueran calculables. Las columnas eliminadas fueron las siguientes:

- **La primera columna:** Esta columna solo tenía en su registros números. Sin ningún encabezado que la identifique, por ende, no se puede saber que representa esa columna incluso estando claro de cual es el contexto del problema.
- Todas las columnas del tipo: **“De ser afirmativo indique el motivo o la referente a las veces que se ha realizado tesis”,** se eliminaron, ya que la columna es dependiente. Se tomó la información de la columna

dependiente y se imputó en la celda de la columna anterior en caso de ser afirmativa.

- **Dirección donde se encuentra ubicada la residencia o habitación:** Se tomó la decisión de eliminar esta columna ya que a mi criterio no aporta nada de información al estudio.
- **Contrajo matrimonio:** Esta columna fue eliminada ya que los valores son redundantes o se pueden deducir con el estado civil.
- **Ingreso mensual total:** Se tomó la decisión de eliminar esta columna ya que se puede deducir con la suma de todos los ingresos.
- **Residencia o habitación alquilada:** Esta columna es eliminada por que se une con la columna "**en caso de vivir en habitación alquilada o residencia estudiantil indique el monto**".
- **Totales de Egresos (del estudiante y el representante económico):** Estas columnas fueron eliminadas ya que se puede deducir sumando la cantidad de gastos de las columnas anteriores.
- **Sugerencias y recomendaciones para mejorar nuestra atención:** Esta columna fue eliminada ya que a mi criterio no aporta ningún dato interesante al problema.