Universidad Central de Venezuela

Facultad de Ciencias

Escuela de Computación

Materia: Minería de Datos

Alumno: Jefferson Santiago

Tarea #1: Preprocesamiento de datos.

Contexto del problema: Nos encontramos de primera mano con datos recolectados mediante las encuestas que se realizan para renovar la ayuda estudiantil que brinda la Organización de Becas Crema. De dichos datos, la Organización pide ayuda para realizar una limpieza de datos al conjunto de datos resultante a las encuestas realizadas.

A continuación se presentan las características o dimensiones asociadas al conjunto de datos:

En primer lugar se encuentra una primera columna numérica de la cual no se puede concluir nada, ya que no posee cabecera ni algún patrón que con el cual se pueda concluir o intuir el significado de dicha columna. Por ende, esa columna será eliminada del conjunto de datos.

El set de datos posee otras 64 columnas que si tienen su respectiva cabecera y se puede intuir el significado en el contexto del problema. Dichas columnas o dimensiones se explicarán a continuación:

Dimensiones:

Indique el período académico a renovar: En esta columna el estudiante le indica a la Organización Becas Crema, que periodo académico desea renovar. Por lo tanto contiene todos los periodos a renovar. Esta columna se estandarizará ya que los periodos no están escritos de la misma manera.

Cedula de identidad: Esta columna contiene el número de cedula de identidad de cada persona que se encuentra en el conjunto de datos. Todos los números de cédulas

que contiene la columna están escritos con el mismo formato y no necesita estandarización.

Fecha de Nacimiento, colocar solo datos numéricos: Esta columna hace referencia y contiene la fecha de nacimiento de las personas encontradas en el conjunto de datos. Dicha columna, además, especifica que los datos deben ser numéricos. Las fechas de nacimiento encontradas en esta columnas no cumplen todas un mismo formato, por lo tanto, es necesario estandarizarlas para que todas queden con un formato del tipo "dd/mm/aaaa".

Edad: Esta dimensión o columna simplemente almacena las edades de las personas almacenadas en el conjunto de datos. No todas las edades encontradas en esta dimensión poseen un estándar, ya que en algunos casos contiene la palabra "Años" la cual debe ser removida, de esta manera se estandarizará la columna para que todos los datos sean numéricos.

Estado civil: Esta columna hace referencia y se almacena el estado civil de cada persona almacenada en el conjunto de datos. El estado civil almacenado en esta columna se numerizará; Un proceso que será explicado con mayor detalle en el documento "informe.pdf".

Sexo: Es la columna que almacena el género de cada persona en el conjunto de datos. Se realizará una numeración que se especificará en el documento "informe.pdf".

Escuela: En esta columna se almacena la escuela a la que pertenece cada en el conjunto de datos. Los alumnos encuestados solo son de las escuelas de Enfermería y de Bioanálisis. En esta columna se realizará una numerización.

Año de ingreso a la UCV: Esta columna especifica el año de ingreso a la Universidad Central de Venezuela, de cada persona en el conjunto de datos. Como todos los datos de esta columna están bien estructurados, se dejará exactamente como estan.

Modalidad de ingreso a la UCV: Especifica el modo o manera por la cual el alumno ingresó a la Universidad Central de Venezuela. En esta columna se realizará una numerización, que se explicará con más detalle en el archivo "informe.pdf".

Semestre que cursa: En esta columna cada persona en el conjunto de datos, indica el semestre que está cursando en el momento de llenar la encuesta realizada por la Organización Becas Crema. Esta columna se estandarizará dejando solo los números que indica el semestre en curso.

Ha cambiado usted de dirección: Esta columna hace referencia al cambio de dirección de la persona. Se ha llenado en caso afirmativo con un "Si", en caso negativo con un "No". Esta columna se unirá con la siguiente "**De ser afirmativo indique el motivo**", el proceso será explicado con más detalle en el documento: "informe.pdf".

De ser afirmativo indique el motivo: Esta columna indica el motivo por el cual la persona cambió de dirección. En caso de no haber cambiado de dirección el registro queda vacío, simplemente no se llena o no aplica para esa persona que no ha cambiado de dirección.

Numero de materias inscritas en el semestre o año anterior: Como su nombre lo indica esta columna almacena el número de materias inscritas por estudiante en el semestre o año anterior. Esta columna está bien estructurada, en cada registro se encuentra un entero. Por lo cual no se realizará algún cambio.

Numero de materias aprobadas en el semestre o año anterior: Como su nombre lo indica esta columna almacena el número de materias aprobadas por estudiante en el semestre o año anterior. Se estructurará la columna dejando solo los números ingresados por cada estudiante, esto para tener el mismo tipo de dato en la columna.

Numero de materias retiradas en el semestre o año anterior: Como su nombre lo indica esta columna almacena el número de materias retiradas por estudiante en el semestre o año anterior. Esta columna está bien estructurada, en cada registro se encuentra un entero. Por lo cual no se realizará algún cambio.

Numero de materias reprobadas en el semestre o año anterior: Como su nombre lo indica esta columna almacena el número de materias reprobadas por estudiante en el semestre o año anterior. Esta columna está bien estructurada, en cada registro se encuentra un entero. Por lo cual no se realizará algún cambio.

Promedio ponderado aprobado: Esta columna almacena el promedio ponderado aprobado de cada persona en el conjunto de datos. Se estandarizará esta columna con un formato del tipo “00,0000” para evitar confusiones. El proceso se explicará con más detalle en el documento “informe.pdf”.

Eficiencia: Esta columna almacena Eficiencia de cada persona en el conjunto de datos. Se estandarizará esta columna con un formato del tipo “0,0000” para evitar confusiones. El proceso se explicará con más detalle en el documento “informe.pdf”.

Si reprobó una o más materias indique el motivo: Como se indica, en esta columna se coloca el motivo por el cual la persona reprobó una o más materias. En caso de que la persona no haya reprobado simplemente lo deja en blanco.

Numero de materias inscritas en el semestre o en curso: Como su nombre lo indica esta columna almacena el número de materias inscritas por estudiante en el semestre en curso. Esta columna está bien estructurada, en cada registro se encuentra un entero. Por lo cual no se realizará algún cambio.

Estás realizando tesis, trabajo de grado o pasantías de grado: En esta columna se almacena la información referente a si realiza o no TEG la persona en particular. Esta columna se unirá con la siguiente “**Tesis, trabajo de grado o pasantías de grado**” para disminuir la dimensionalidad del problema. Aunado a ello, se numerizará la columna. El proceso se explicará con más detalle en el documento “informe.pdf”.

Tesis, trabajo de grado o pasantías de grado: Esta columna representa o almacena la ocurrencia del evento, si está realizando trabajo de grado. Además indica si es primera vez, si el alumno lleva un segundo semestre con el TEG, o si el alumno lleva más de 2 semestres realizando el TEG.

Procedencia: Esta columna indica el lugar de procedencia de cada persona dentro del conjunto de datos.

Lugar donde reside mientras estudia en la universidad: Se refiere al lugar donde vive la persona ya sea un sitio temporal o no, mientras estudia en la universidad.

Personas con las cuales usted vive mientras estudia en la universidad: Como se indica, en esta columna el estudiante especifica con quien vive mientras estudia en la universidad.

Tipo de vivienda donde reside mientras estudia en la universidad: Como se indica, en esta columna especifica el tipo de vivienda que habita el estudiante.

En caso de vivir en habitación alquilada o residencia estudiantil indique el monto mensual: Esta columna especifica el monto a pagar por la residencia o el lugar donde vive dicho alumno. No está especificado si el monto es cancelado sólo por el estudiante o el responsable económico o ambos, si se requiere de este estudio se pueden consultar las columnas siguientes. Esta columna se unirá con la columna "**Residencia o habitación alquilada**" ya que poseen la misma información. Se unirán para disminuir la dimensionalidad del problema. Además se unirán y se encontrarán registros con números solo si el registro anterior de la columna de "**tipo de vivienda donde reside mientras estudia**" tiene habitación alquilada o residencia estudiantil.

Dirección donde se encuentra ubicada la residencia o habitación alquilada: Esta columna especifica la dirección sólo de las personas que poseen una habitación alquilada o residencia estudiantil.

Contrajo Matrimonio: Esta columna especifica si el estudiante ha contraído matrimonio. En caso de ser afirmativo se llena con el valor "Si", de lo contrario con "No".

Ha solicitado algún otro beneficio a la universidad u otra institución: Esta columna especifica si el alumno ha solicitado algún otro servicio a la universidad. Fue llenada con valores como "Si" o "No". Se unirá esta columna con la siguiente: "**En caso afirmativo señale el año de la solicitud, la institución y el motivo**", con el objetivo de disminuir la dimensionalidad.

En caso afirmativo señale el año de la solicitud, la institución y el motivo: Se refiere y se encuentran registros los cuales indican el año de la solicitud, la institución y el motivo de la solicitud del beneficio a cualquier otra institución.

Monto mensual de la beca: Esta columna indica el monto mensual que recibe cada estudiante como ayuda por la Organización Becas Crema. Esta columna contiene puros valores numéricos y no es necesario realizarle algún cambio en particular.

Aporte mensual que le brinda su responsable económico: Esta columna indica el monto mensual que recibe cada estudiante de su responsable económico si es que lo tiene. En caso contrario no aplica para determinados estudiantes. Esta columna se estandarizará para que en los datos faltantes o con valor "NA" se cambien a cero. El proceso de estandarización se explicará de una manera más detallada en el documento "informe.pdf".

Aporte mensual que recibe de familiares y amigos: Esta columna indica el monto mensual que recibe cada estudiante de familiares y amigos. Si ese no es el caso, no aplicará para

determinados estudiantes. Esta columna se estandarizará para que en los datos faltantes o con valor “NA” se cambien a cero. El proceso de estandarización se explicará de una manera más detallada en el documento “informe.pdf”.

Ingreso mensual que recibe por actividades a destajo o por horas: Esta columna indica el monto mensual que recibe cada estudiante mediante su trabajo a destajo o por hora. Si ese no es el caso, no aplicará para determinados estudiantes. Esta columna se estandarizará para que en los datos faltantes o con valor “NA” se cambien a cero. El proceso de estandarización se explicará de una manera más detallada en el documento “informe.pdf”.

Ingreso mensual total: Representa y en esta columna se almacena el ingreso mensual de manera totalizada de cada persona. Esta columna se estandarizará ya que no todos los registros se encuentran con la misma estructura de los montos.

Alimentación: Esta columna se refiere y almacena los gastos mensuales por motivos de alimentación de cada estudiante en particular si es pertinente. De no aplicar este gato para determinados alumnos, el valor en el registro es “NA”. Por lo tanto la columna se estandarizará.

Trasporte público: Esta columna se refiere y almacena los gastos mensuales por motivos de transporte de cada estudiante en particular si es pertinente. De no aplicar este gato para determinados alumnos, el valor en el registro es “NA”. Por lo tanto la columna se estandarizará.

Gastos médicos: Esta columna se refiere y almacena los gastos mensuales por motivos médicos de cada estudiante en particular si es pertinente. De no aplicar este gato para determinados alumnos, el valor en el registro es “NA”. Por lo tanto la columna se estandarizará.

Gastos odontologicos: Esta columna se refiere y almacena los gastos mensuales por motivos odontológicos de cada estudiante en particular si es pertinente. De no aplicar este gato para determinados alumnos, el valor en el registro es “NA”. Por lo tanto la columna se estandarizará.

Gastos personales: Esta columna se refiere y almacena los gastos mensuales personales de cada estudiante en particular si es pertinente. De no aplicar este gato para determinados alumnos, el valor en el registro es “NA”. Por lo tanto la columna se estandarizará.

Residencia o habitación alquilada: Esta columna se refiere y almacena los gastos mensuales por motivos de alquiler de residencia o habitación de cada estudiante en particular si es pertinente. De no aplicar este gato para determinados alumnos, el valor en el registro es “NA”. Por lo tanto la columna se estandarizará.

Materiales de estudio: Esta columna se refiere y almacena los gastos mensuales por materiales de estudio de cada estudiante en particular si es pertinente.

Recreación: Esta columna se refiere y almacena los gastos mensuales por motivos de recreación de cada estudiante en particular si es pertinente. De no aplicar este gato para determinados alumnos, el valor en el registro es “NA”. Por lo tanto la columna se estandarizará.

Otros gastos: Esta columna se refiere y almacena los gastos mensuales por otros motivos que no hayan sido tomados en cuenta anteriormente de cada estudiante en particular si es

pertinente. De no aplicar este gato para determinados alumnos, el valor en el registro es “NA”. Por lo tanto la columna se estandarizará.

Total de egresos: Esta columna representa el total de egresos mensual por cada alumno en particular.

Indique quien es su responsable económico: Esta columna indica quien es el responsable económico del estudiante.

Carga familiar: Esta columna indica la carga familiar que posee el responsable económico del alumno.

Ingreso mensual de su responsable económico: Como lo indica, esta columna guarda el ingreso mensual producto del trabajo del responsable económico del estudiante. Esta columna será estandarizada, el proceso será explicado en el documento “informe.pdf”.

Otros ingresos: Esta columna indica todos los ingresos mensuales distintos a los recibidos por el fruto del trabajo del responsable económico del estudiante. Esta columna será estandarizada, el proceso será explicado en el documento “informe.pdf”.

Total de ingresos: Esta columna indica el total de los ingresos mensuales del responsable económico del estudiante. Esta columna será estandarizada, el proceso será explicado en el documento “informe.pdf”.

Vivienda: Esta columna se refiere y almacena los gastos mensuales por motivos de vivienda de cada representante económico de un estudiante en particular si es pertinente. De no aplicar este gato para determinados alumnos, el valor en el registro es vacio. Por lo tanto la columna se estandarizará.

Alimentación: Esta columna se refiere y almacena los gastos mensuales por motivos de vivienda de cada representante económico de un estudiante en particular si es pertinente. De no aplicar este gato para determinados representantes de alumnos, el valor en el registro es vacío. Por lo tanto la columna se estandarizará.

Transporte: Esta columna se refiere y almacena los gastos mensuales por motivos de transporte de cada representante económico de un estudiante en particular si es pertinente. De no aplicar este gato para determinados representante de alumnos, el valor en el registro es vacío. Por lo tanto la columna se estandarizará.

Gastos odontológicos: Esta columna se refiere y almacena los gastos mensuales por motivos odontológicos de cada representante económico de un estudiante en particular si es pertinente. De no aplicar este gato para determinados representante de alumnos, el valor en el registro es vacío. Por lo tanto la columna se estandarizará.

Gastos educativos: Esta columna se refiere y almacena los gastos mensuales por motivos educativos de cada representante económico de un estudiante en particular si es pertinente. De no aplicar este gato para determinados representante de alumnos, el valor en el registro es “NA”. Por lo tanto la columna se estandarizará.

Servicios públicos, agua, luz, teléfono, gas: Esta columna se refiere y almacena los gastos mensuales por servicios públicos de cada representante económico de un estudiante en particular si es pertinente. De no aplicar este gato para determinados representante de alumnos, el valor en el registro es vacío. Por lo tanto la columna se estandarizará.

Condominio: Esta columna se refiere y almacena los gastos mensuales por motivos de condominio de cada representante económico de un estudiante en particular si es pertinente. De no aplicar este gato para determinados representante de alumnos, el valor en el registro es vacío. Por lo tanto la columna se estandarizará.

Otros gastos: Esta columna se refiere y almacena los gastos mensuales no indicados anteriormente de cada representante económico de un estudiante en particular si es pertinente. De no aplicar este gato para determinados representante de alumnos, el valor en el registro es "NA". Por lo tanto la columna se estandarizará.

Total egresos: Esta columna indica el total de los egresos mensuales del responsable económico del estudiante.

Opinión de los usuarios: Como se indica, en esta columna se encuentran las valoraciones que dan los estudiantes a la calidad del servicio de la Organización Becas Crema. Las valoraciones van de 1 a 5 donde 1 es la menor calidad de servicio según la opinión del estudiante o 5 es la mayor calidad de servicio, se podría decir que es como si se clasificara de la siguiente manera: muy malo, malo, ni malo ni bueno, bueno y muy bueno.

Sugerencias y recomendaciones: Como se indica, en esta columna se identifican las sugerencias y recomendaciones de los usuarios para mejorar el servicio de la Organización Becas Crema.