Informe - Becas

Josseline Perdomo

Resumen

Documento donde se explica el proceso que se realizó con los datos, aplicando una serie de procedimientos para poder limpiar lo suficiente estos datos, es de suma importancia para la organización.

1. Introduccion

La principal tarea que se nos fue encomendada por la organización Becas Crema fue el de la limpieza de los datos para proceder hacer una vista minable. Un set de datos que se encuentra ordenado y de una manera mas limpia interna, es mucho mas manipulable que uno sin tratar, por ende esta tarea para cualquier set de datos es importante. Se explicará de manera breve las decisiones y el proceso que se tomó para poder limpiar y ordenar los datos de las renovaciones de Becas.

2. Eliminación de Características

Este proceso se realizó luego de analizar detenidamente el conjunto de datos y ver cuales eran los que tendrían mas relevancia sin caer en la redundancia de datos.

- Columna 1: Esta columna se eliminó porque en ningún momento en el proceso de tratar de deducir los datos se pudo saber que significaba.
- Columna 5: Esta columna se eliminó porque es la edad del estudiante y esta se puede calcular a través de la fecha de nacimiento, sin embargo, se utilizó el valor de dicha columna para calcular la fecha de nacimiento de aquellos estudiantes que no tenían una fecha válida, es decir, con el formato adecuado que a pesar del proceso de tratar la columna de fecha de nacimiento, no se podía lograr recuperar correctamente.
- Columna 22: Esta columna se eliminó porque podía ser generada a través de otra, que es la de cuantas veces ha inscrito TEG, sin embargo, se utilizo igualmente para corroborar la veracidad de la columna de cuantas veces inscribió TEG.
- Columna 28: Esta columna se eliminó porque era redundancia, ya que existe la columna 45 que es gastos de Residencia Estudiantil o Habitación Alquilada, sin embargo, se utilizó para verificar que los datos fueran correctos en ambas columnas.
- Columna 29: Esta columna se eliminó porque no es relevante tener la dirección de la residencia o habitación alquilada, o al menos no de esta manera.
- Columna 66: Esta columna se eliminó porque no es relevante tener las sugerencias de esta manera y además de que con la puntuación del estudiante que se encuentra en la columna 65 es suficiente para tomar decisiones.

3. Transformación de Datos Iniciales

Cada columna fue tratada de manera detallada y diferente pero con el fin de sacarle todo el provecho a estas.

3.1. Nombres

Se usaron nuevos nombres de características ya que con los anteriores era complicado el manejo de los datos, así como la indexación es mas sencillo de usar. Estos nombres pueden verse luego en el archivo .csv resultante.

3.2. Tipo de datos

Se transformó en cada columna el tipo de dato, dependiendo claro qué datos se encontraban en ella, esto para que el manejo fuera mas sencillo.

3.3. NaN

En los datos de las columnas de Ingresos y Egresos en donde se encontraba NaN se cambio a el valor cero, ya que era lo mas lógico dado el contexto del problema.

3.4. Periodo a Renovar

El periodo a renovar se separó en dos columnas porque basándonos en el paper Tidy Data en que dice que no

debería de haber dos características en una sola columna, en este caso era el año y el semestre del periodo a renovar.

3.5. Binarización

Columnas que solo tenían solamente dos valores posibles se procedió a binarizar.

3.6. Categorización

En columnas que hubieran valores finitos se procedió a categorizarlos, hubo columnas que se unieron posibles valores ya que representaban lo mismo.

4. Valores Faltantes

Como es de esperarse habían valores que no se encontraba, lo que complica cualquier estudio sobre los datos que se quiera realizar, por este motivo se procedió a imputar los valores faltantes.

4.1. Criterio

- Moda : Se uso la moda en los momentos en que los datos a imputar eran categóricos o llamados también finitos.
- Media : Se uso la media en los momentos en que los datos a imputar eran valores que podían ser reales.