Universidad Central de Venezuela
Facultad de Ciencias
Escuela de Computación
Minería de datos
Andrea Tellería – CI: 20.614.114

Tarea 1: Preprocesamiento de datos

Informe

Primero que nada se pasó a leer la información, para esto se usó la estructura *DataFrame* provista por el paquete *pandas*, de esta forma se almacenó en una variable con la que después se manipularía la data.

Con ayuda de las funciones provistas por *pandas* se reemplazaron ciertas cadenas de caracteres de tal forma que fuese más fácil el almacenamiento. Siguiendo este esquema se decidió que cada escuela estaría representada por un número de esta forma:

Bioanálisis	0
Enfermería	1

Siguiendo esta misma decisión se reemplazaron otras cadenas, tales como el estado marital y el sexo. Parecido, se cambio el dato en el que expresaba el semestre que cursaba el estudiante, de tal forma que al final sólo quedase un dato numérico y no uno tipo *string*.

1er sem.	1
2do sem.	2
…	
10mo sem.	10

A la hora de manejar las fechas se decidió usar un dato tipo *datetime,* parte del paquete de *python* que está orientado al manejo de fechas y que tiene el mismo nombre que el tipo de dato que se buscaba obtener.

Para realizar el cambio, primero se buscó obtener las fechas mediante una función en la cual se validaban los formatos de fechas posibles que se podían tener. De esta forma pasar la fecha final a un dato tipo *datetime.* En el caso de encontrar fechas no válidas, o fechas vacías, se decidió que con ayuda de la columna de la Edad, se buscaría la edad de la persona, para evitar que este fuese mayor a lo que era y tomando en cuenta que no se tenía información de día o mes de nacimiento, se dejó el treinta y uno de diciembre como marca (31-12-XX).

De	**A**

19220485	12/12/199X

Con la eficiencia, ocurría que esta era mayor a uno (1), dejando ver que los estudiantes habían llenado este apartado, sin tomar en cuenta puntos (.), con lo cual el valor estaba en la mayorái de los casos distorsionado. Para evitar que los valores que si fueron escritos correctamente se viesen afectados, se creó una función que validara la eficiencia de tal forma que fuese un valor válido.

De	**A**
454612	0. 454612

Para el caso de columnas numéricas donde fueron escritos caracteres que no eran números, ya sean palabras o comas (,), se creó una función que validara dígitos, de tal forma de que si una palabra tenía caracteres que no fuesen numéricos, se eliminasen.

De	**A**
10,393,93	10393.93
15000 bs	15000

Errores de este tipo fueron comunes, en especial cuando se solicitaron montos.

Para el promedio ponderado, similar a lo que currió con la eficiencia, muchos valores no sonaban coherente, ya que estos superaban veinte (20), la máxima calificación existente. Después de notar esto, se decidió por crear análogo a la eficiencia de una función la cual solucionase esto. De tal forma la data fue modificada, donde:

De	**A**
7816	7.816
15893	15.893

Una vez se consideró que las columnas ya no tenían errores importantes, se pasó a tratar con los espacios vacíos y NAs, para esto se usaron las funciones de *panas*, tales como *fillna()* y *mean()*.

Se decidió que para los montos en el caso de no existir, se dejaría la media de la columna. Con lo cual no se afectaría la media final de la misma. Así pues, esta estrategia fue usada en la mayoría de las columnas cuyos tipos de datos, al ser de tipo numérico, lo permitieron (la mayoría de las columnas de ingresos y egresos, por ejemplo). Al mismo tiempo se decidió seguir otra estrategia en algunas columnas, tales como residencia y vivienda.

En el caso de estas dos, se tomó en cuenta que los estudiantes no habían señalado que vivían en residencia, de tal forma lucía ilógico escoger el valor de la media en estos casos. Para esta columna se decidió, pues, en vez de la media dejar estas celdas en cero (0).

Luego se procedió a eliminar columnas que no se consideraron necesarias a la hora de realizar un estudio, o por otras razones que se explicaran con mayor claridad a continuación.

En el caso de filas como: *X.Contrjo.Matrimonio* se decidió eliminar, ya que previamente se pregunta el estado civil, resulta redúndate saber si se ha casado. Esa información, ya se posee. La columna *unnamed (sin nombre)* encontrada, después de decidir que se trataba de una columna que indicaba números de filas, se pasó a eliminar.

Similar al análisis realizado cuando se eliminó *X.Contrjo.Matrimoni,* esta vez se pasó a eliminar edad, si ya tenemos la fecha de nacimiento, es solo un caso de calcular. Tenerla como tal no aporta nada nuevo. En el caso de la mayoría de explicaciones sobre, por qué se reprobaron materias, o sobre cómo mejorar el sistema, si bien son importantes, se prefirieron eliminar, ya que de realizar un estudio sobre becas, no importa mucho por qué un estudiante reprobó y con la evaluación del sistema es suficiente para saber si existe deficiencia o no.

Los totales de los montos también se eliminaron, esto en parte es para evitar incoherencias como varias de las que se vieron mientras se revisaba la data.

Ingreso Mensual del Responsable Legal	**Otros Ingresos**	**Ingreso Total del Responsable Legal**
19000	0	19152.14

Si se sabe que la información ya se tiene y disgregada, entonces la información de totales resulta redundante. Así pues este tipos de columnas fueron eliminadas.

Sí bien se decidió mantener la columna que informaba si un estudiante estaba en trabajo de grado, se decidió omitir la columna que expresaba si esta era la primera vez que lo hacía o no, ya que como tal, esta información no se consideró tenía importancia a la hora de decidir si ofrecer una beca o no. Análogo al método usado para ingresar en la universidad, sea cual sea este, más que ver con cómo se ingresó, un estudio sobre becas debería estar orientado a cuanto podría ayudar a un estudiante la ayuda que se ofrece y sobre la familia de esté, ya que con esta información se sabe si es merecedor de la ayuda o no.

Saber quién es el responsable económico, tampoco influye, basta con saber la carga familiar y la situación económica del mismo, más de saber si fue un padre, una madre o un hermano.

Personas con las que se vive, así como el lugar donde se queda un estudiante mientras estudia también son innecesarios, basta con saber el costo, en caso de vivir una residencia. Así pues, con el lugar de procedencia, el cual nos dice donde está la verdadera casa del estudiante, es suficiente.

Diferentes razones, llevaron a decidir eliminar la columna de semestre a renovar. En este caso fue debido a lo complicado que resultaba el auditar la data de forma que todos siguiesen el

mismo formato, debido a esto, si bien se planeó mantener esta columna, al final se decidió que mediante la fecha de ingreso y el semestre que cursaba el estudiante eran suficientes para saber el semestre a renovar.

Una vez hecho esto, se pasó a aplicar el algoritmo PCA a forma de reducir la dimensionalidad del problema. Al hacerlo, se descubrió que cualquier tipo que no fuese numérico producía errores en el método que se estaba usando, por lo cual, se optó por eliminar la columna de Fechas, la cuales al ser tipo *datetime* causaban el problema.