Informe

Leonardo Santella

January 2016

Introducción

En el preprocesamiento de los datos proporcionados, se tomaron una serie de decisiones para la generación de una Vista Minable basada en los datos proporcionados. Estas decisiones seran respaldadas en este documento, además de incluir un breve análisis exploratorio de los datos.

Problema

El preprocesamiento de datos consiste en llevar a cabo una serie de pasos, para la transformación de dichos datos en una vista minable. Una vista minable, es un nuevo set de datos que mejora alguna de las características del análisis de los datos (Por ejemplo, la eficiencia de los algoritmos usados en el análisis). Los pasos llevados a cabo para transformar los datos en una vista minable, no estan preestablecidos, no existe un standard general de como realizarlos, sin embargo existen varias actividades o etapas que son definidas a través de los pasos. Limpieza de los datos y portabilidad de los datos, fueron las actividades implementadas en la generación de la Vista minable.

Cada una de las etapas de preprocesamiento tiene objetivos diferentes y por lo tanto tienen como finalidad resolver problemas diferentes.

Es importante destacar, que la generación de una Vista Minable, podria estar directamente asociada con la Tarea o 'Task' (o conjunto de estos) que se quiera realizar (Clusterización, Detección de valores anomalos, etc). En nuestro caso, no se provee una tarea para la cual seran utilizados los datos (Vista minable), por lo tanto muchas decisiones seran afectadas por lo anterior.

Preprocesamiento

Portabilidad de los datos

El objetivo de esta etapa del preprocesamiento de los datos es la transformación de los tipos de datos de las características de los individuos de la muestra. En nuestro caso, siguiendo los lineamientos establecidos por el documento 'Tidy Data', las características de los individuos son representadas como columnas.

La decisión que principalmente determina el resto de los pasos llevados implementados en esta actividad fue la transformación de la mayor cantidad de atributos, en atributos (o caracateristicas) numericas, ya que el uso de variables numericas, permite la utilización de las medidas de centralidad mas utilizadas, en ocasiones aumenta la eficiencia de los algoritmos, y existen un número considerable de librerias que implementan algoritmos de análisis de datos que funcionan con data de entrada numerica. Sin embargo, debido a que no se conoce la tarea (o conjunto de ellas) que se quiere realizar con dichos datos, se remite la conservación de algunas variables categóricas.

La transformación de los datos categóricos en el set de datos como por ejemplo E_Civil (Estado civil del individuo) fueron tranformados en variables numericas, a través de un proceso denominado Categorización. El proceso de categorización consiste en asignarle un número entero (tambien podria ser un número binario) a un valor categórico, y sustituirlo en el set de datos. Esto involucra una perdida de representatividad, pero tambien, una mejora en la eficiencia en memoria, ya que ocupa menos espacio en memoria un número entero, que una cadena de caracteres.

Limpieza de los datos

En esta etapa de preprocesamiento se toma las decisiones referentes los valores faltantes, valores inconsistentes y/o valores con errores, como tambien, la escala de representación para valores numericos (Por ejemplo, edad e ingresos). Las decisiones relacionadas con estos valores, varian entre la estimación o imputación y la eliminación de individuos o características (Filas o columnas) basada en ciertas metricas definidas por el experto u otra persona involucrada en el proceso de mineria de datos.

Al observar el set de datos, es posible apreciar que ciertos valores estan representados por la cadena 'NA', esta notación indica que el valor no esta disponible (Not Available) sin embargo, esta representación, en el entorno en que fue llevado a cabo el preprocesamiento de los datos, no es reconocido como un dato no disponible, sino como una cadena de caracteres, por lo tanto estos valores fueron sustituidos por el valor respectivo para la representación de datos no disponibles (numpy.nan, representa valores no numericos. 'Not a Number').

Luego de identificar los valores no disponibles, se tomaron decisiones en relación al tipo de característica (Categorica o Numerica). En las características numericas, los valores no disponibles fueron sustituidos por el número 0 debido a que se asumio que las personas dejaron dichos campos en blanco, en representación de que no existe dicha cantidad o número. En las características categóricas que fueron modificadas debido a que tenian valores faltantes (no dispobibles) se imputó la moda (Medida de centralidad que indica el valor que mas se repite en una serie de observaciones).

Ademas, en esta etapa de preprocesamiento, para la detección de valores erroneos, es posible aplicar tecnicas para la detección de valores anomalos, debido a que en ocasiones, estos pueden resultar ser valores erroneos. Esta decisiones relacionadas los valores anomalos son llevadas a cabo por algun experto que

forme parte del proceso de mineria de datos.

En las columnas numericas, existian valores que no tenian coherencia con la característica, como por ejemplo, la eficiencia (es un valor entre 0 y 1) existian valores mayores que 1.

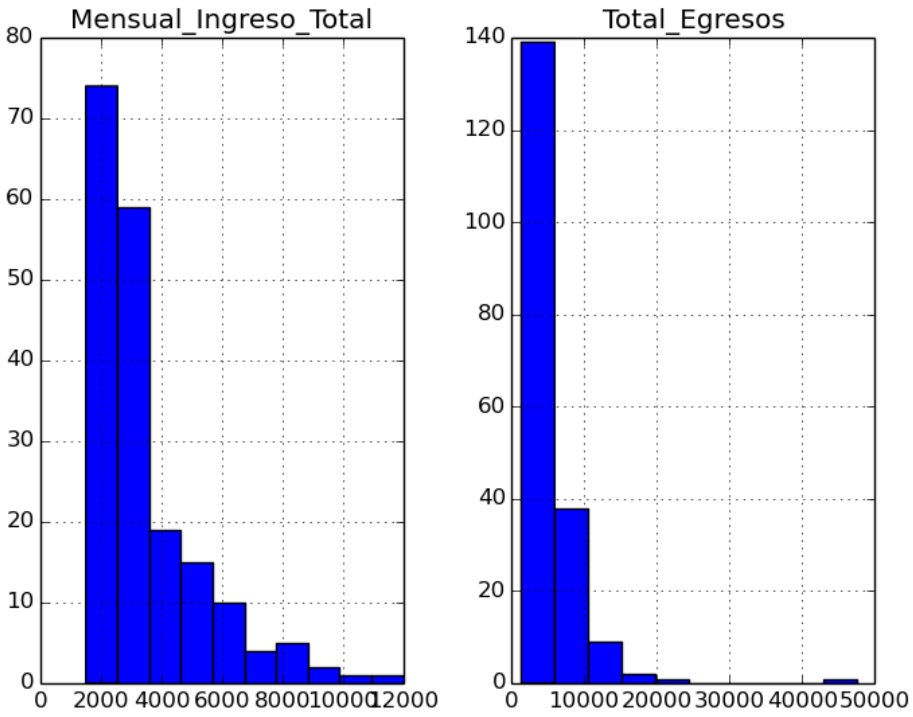
Obsevaciones

Mientras se llevaban a cabo las diferentes etapas del preprocesamiento, se intento implementar algunas tecnicas de reducción de dimensionalidad, debido a que el número de características es considerable, sin embargo no fue tomado en cuenta para la generación de la vista minable, debido a que los resultados no fueron los esperados como tambien el hecho de no conocer la finalidad de la vista minable. Las lineas de codigo que involucran la reducción de dimensionalidad se encuentran como comentarios y se encuentran en la parte final del archivo script.py.

Análisis Exploratorio de los datos

El análisis exploratorio de los datos tiene como finalidad el la visualización y reconocimiento de ciertas características de los datos del conjunto de datos, en muchas ocasiones, se dispone de un objetivo en concreto, sin embargo el unico objetivo de esta actividad, es el preprocesamiento de los datos para la generación de una vidta minable, desconocemos la finalidad de dichos datos, por lo tanto, solo se referira a ciertos aspectos que fueron considerados importantes

Index	Alimentacion	T_Publico	astos_Medico	astos_Odonto	stos_Personal	astos_Alquile	stos_Material	stos_Recreaci	Gastos_Otros
count	190	190	190	190	190	190	190	190	190
mean	1232.3684	608.08421	345	194.73684	569	300.89474	1040.8947	274.31579	326.5
std	2319.7774	588.28737	869.69176	443.52924	557.43607	1039.7516	702.272	446.95241	861.7946
min	0	0	0	0	0	0	100	0	0
25%	412.5	250	0	0	200	0	500	0	0
50%	1000	500	0	0	500	0	1000	0	200
75%	1500	750	500	0	800	0	1500	500	300
max	30000	3500	10000	2500	3000	6500	4000	2500	10900



Una Observación interesante es que existe un número considerable de personas que tiene gastos mayores a los ingresos

Index	Monto_Mensual_Beca	Mensual_Aporte_Resp	Mensual_Aporte_Amigos	Aporte_Destajo
count	190	190	190	190
mean	1495.2632	1618.9474	331.76316	107.89474
std	126.42284	1414.8715	718.4733	419.36055
min	200	0	0	0
25%	1500	600	0	0
50%	1500	1000	0	0
75%	1500	2000	500	0
max	2000	9000	5000	3000

Index	Ingreso_Responsable	Ingreso_Resp_Otros
count	190	112
mean	71903.406	3818.6561
std	649347.29	3494.1772
min	0	0
25%	6507.25	1425
50%	8450	3000
75%	12040.875	5605.5
max	8773759	17879

Index	Total_Egresos_Resp
count	190
mean	20072.183
std	96933.519
min	0
25%	7400
50%	10761
75%	16303.44
max	1344482