# Informe

En el presente informe se procede a dar detalles sobre todos los procedimientos utilizados al momento de generar la vista minable del set de datos.

En primera instancia, se busca estandarizar todas las columnas. Con esto se quiere decir que todos los datos de todas las columnas deben mostrarse de manera ordenada y siguiendo una especie de formato para cada una de ellas para maximizar la eficiencia.

En la mayoría de las columnas se busca numerizar las mismas puesto que es más eficiente su manejo de esta manera. Esto se realizó haciendo reemplazos en las columnas que tenían varios valores que, por ser iguales, podían representarse de manera numérica de forma que se pudieran agrupar e identificarlos con un único número que nos ayudara a diferenciar estos valores de manera más fácil y rápida (Por ejemplo, la columna de Sexo, se utilizo el número 0 para identificar el femenino y 1 para identificar el masculino. Así se realizó para esta columna y para una gran cantidad de las demás, dada la facilidad que nos otorga representarlas de esta manera.).

En algunas columnas como Eficiencia y Promedio_Pond, se hicieron varias multiplicaciones y divisiones para llevarlas a un formato común y como la mismas ya son numéricas no haría falta alguna otra modificación.

En el caso de las columnas que no pudieran ser numerizables, como las columnas que justificaciones y comentarios, se procedió a removerlas puesto que estos comentarios no son información relevante para el proceso de minería. Un ejemplo de esto sería la columna de Sugerencias. Las sugerencias, por ser comentarios, son extremadamente tediosos de numerizar y de igual manera no harían ningún aporte al momento de hacer procesos de minería, es más eficiente removerlos para que no entorpezcan el mismo.
Por otro lado la información de sugerencias podría ser relevante para algún analista posterior a este proceso, pero para este caso no lo es.
Por estos motivos se removieron las columnas de justificaciones y comentarios.

La columna de periodo tuvo que ser estandarizada de una manera distinta, si bien no es numerizable, se hicieron modificaciones para que quedaran en un mismo formato sencillo de entender. Este proceso se realizo buscando letras clave que indicaran el periodo y el año en que se encuentra el estudiante y se procedió a reemplazar cada fila por su representación en el formato adecuado.

En la columna de fecha de nacimiento se realizaron solo agregaciones en años que estaban en formatos distintos de manera que el resultado final fueron fechas estructuradas de igual manera. De igual forma se realizaron dos imputaciones de instancias a causa de esta columna, puesto que la información otorgada por el estudiante carecía de sentido, y siendo esta columna, información tan vital para el proceso se comenzó el proceso de imputación de instancias con estos dos casos especiales.

La columna edad fue removida porque la edad podría ser fácilmente calculada ya que estamos en posesión de la fecha de nacimiento del estudiante.

De igual forma, se realizaron imputaciones de columnas que poseían información redundante tales como los totales de algunos ingresos, egresos de los estudiantes y responsables económicos. Esta información puede ser calculada a partir de información que ya se tiene representada.

A medida que el código se desarrolla se van realizando gráficos que enseñan valores atípicos que se muestran a medida que se entandarizan las columnas. Se escogió para uno de los ejes la cédula del estudiante puesto que es un identificador único del mismo y el otro eje es el que empieza a variar dependiendo de la columna en la que se quieran ver outliers.

Para algunos outliers y valores ausentes se hicieron reemplazos con las medidas de moda y media para no hacer variar el set de datos y así conservar su integridad.