Informe donde se explican todas las decisiones que se tomó en cuenta a la hora de realizar la vista minable

A continuación se expondrá una descripción detallada de cada un de los cambios realizados en el archivo de entrada (data.csv).

Se aplicó numerización en 5 campos en los cuales se daban las condiciones para realizar esta técnica, esto con la finalidad de mejorar la visualización del dataset y optimizar posibles cálculos y análisis posteriores. Dentro de los campos que se le aplicó dicha técnica tenemos:

1. **Estado.civil:** este campo estaba compuesto por tres estados (soltero, casado y viudo), se realizo una numerización asignándole prioridad a los campos con mayor numero de repeticiones quedando en este orden: soltero: 1, Casado 2 , Viudo 3. Cabe destacar que se encontró un campo cuyo estado civil no era valido, el cual fue "Unido" por lógica se asumió que el encuestado erróneamente confundió el estado civil casado. Para corroborar el error se visualizo en el campo "Contrajo.Matrimonio" y efectivamente el valor era positivo. Por lo que se procedió a corregir el valor.

2. **Sexo:** En este campo se realizo la numerizacion tomando en cuenta que el sexo feminino tenia el mayor numero de repeticiones quedando en el orden: femenino 1, masculino 0.

3. **Escuela:** Se observo que solo se encuestaron estudiantes de la escuela de biología y enfermería por lo que se encontró conveniente numerizarlos, quedando: Bioanalisis 1, Enfermeria 0

4. **Modalidad.de.ingreso**: Dentro de este campo solo se observaron tres posibles valores los cuales fueron numerizados quedando: Prueba Interna y/o propedéutico 1, Asignado OPSU 2, Convenios 3.

5. **Estás.realizado.tesis.trabajo.de.grado.o.pasantías.de.grado**: llámese A este campo y B el siguiente inmediato. Se pudo notar que B es consecuente de A por lo que se observó conveniente realizar un fusion de ambos y aplicar la técnica de numerización. Se tomaron 4 casos en cuanto a si el estudiante ha realizado tesis, trabajo de grado o pasantías: No ha realizado tesis 0, primera vez 1, segunda vez 2, mas de dos 3.

Al analizar el dataset se encontró presencia de campos en los cuales se observó detalladamente cada uno de sus datos para tratarlos de la mejor forma, equilibrando la menor redundancia y la mayor cantidad de información. Dentro de este conjunto de campos tenemos:

1. **Edad**: este campo podría ser útil a la hora de realizar algún cálculo, pero debido a que tenemos un campo que nos provee la fecha de nacimiento del estudiante, se pudo observar que muchas de las edades no se encontraban actualizadas, razón por la cual se procedió a eliminarlo, tomando en cuenta que el mismo puede ser calculado a través del campo fecha de nacimiento.

2. **Contrajo.Matrimonio:** Este campo fue eliminado debido a que se ya se encuentra un campo Estado civil en el cual se puede visualizar si la persona contrajo o no matrimonio.

3. **Número.de.materias.inscritas.en.el.semestre.o.año.anterior:** este campo es redundante debido a que tenemos los campos: número de materias aprobadas, numero de materias retiradas y numero de materias reprobadas (las tres en el año anterior) y al sumarlas obtenemos el número de materias inscritas en el año anterior, por esta razón este campo fue eliminado.

4. **Ingreso.mensual.total**: este campo es redundante tanto en los ingresos del estudiante como en los ingresos del responsable económico debido a que el ingreso mensual total se puede obtener a través de la suma de todos los ingresos de los mismos, por lo que ambos campos fueron eliminados.

5. **Total.egresos:** este campo es redundante tanto en los egresos del estudiante como en los egresos del responsable económico debido a que el ingreso mensual total se puede obtener a través de la suma de todos los ingresos de los mismos, por lo que ambos campos fueron eliminados

6. **En.caso.de.vivir.en.habitación.alquilada.o.residencia.estudiantil..indique.el.monto.mensual.**: este campo es redundante ya que tenemos otro campo en la sección de egresos llamado "Residencia.o.hab.alquilada" el cual también indica el monto mensual de la residencia o habitación. En este caso se hizo un análisis detallado para unir ambas columnas dejando los datos correctos. Ya que por ejemplo en un campo teniamos "1 UT " mientras que en el otro "150", en dado caso mantuvimos el valor 150 con la finalidad de estandarizar el data set. Todos estos valores se mantuvieron siempre que en el campo de vivienda fuese igual a residencia o habitación estudiantil tal como lo indica la descripción del campo

7. **Período.a.renovar:** Este campo fue eliminado debido a que contenía muchos valores en los cuales se hacía imposible determinar que peŕiodo académico indicó el estudiante.

Cabe destacar que el primer campo no tiene una descripción, por lo que en un momento se llego a asumir que se trataba de un id de cada instancia ya que ninguno de los valores se repetían, pero al visualizar que mas adelante tenemos un campo Cedula.de.Identidad, se llega a la conclusión de que el campo sin descripción puede ser eliminado ya que podemos identificar unívocamente cada campo a través del campo "Cedula.de.Identidad"

Durante el análisis y revisión del dataset se encontraron diversos casos donde se observaban campos dependientes de otros, en tales casos realizamos una unión con la finalidad de ajustar la dimensionalidad del set de datos colocando la mayor cantidad de información. Dentro de este conjunto de pares de campos tenemos:

1. **Ha.cambiado.usted.de.dirección / de.ser.afirmativo.indique.motivo :** como podemos ver en la tupla el segundo campo es un condicional del primero. Si la condición del primer campo era verdadera, la respuesta del motivo (el valor ubicado en el campo 2) era copiado en el campo uno, de esta manera se reemplazan las respuestas afirmativas colocando directamente el motivo.

2. **Tipo.de.vivienda.donde.reside.mientras.estudia.en.la.universidad/ Dirección.donde.se.encuentra.ubicada.la.residencia.o.habitación.alquilada:** En este par de campos la desicion tomada fue unir ambos campos, se concatenó la dirección solamente en los casos donde vivienda era igual a residencia estudiantil o habitación alquilada. Se pudo observar que aunque no vivieran en residencia o habitacion alquilada los estudiantes ingresaron la dirección, es por ello que se realizó el filtro anteriormente descrito.

3. **X.Ha.solicitado.algún.otro.beneficio.a.la.Universidad.u.otra.Institución. / En.caso.afirmativo.señale..año.de.la.solicitud..institución.y.motivo:** como se pude observar al igual que en el punto numero 1, el segundo campo es un condicional del primero. Aplicamos la misma técnica. Unimos reemplazando el motivo solamentre cuando el primer campo es afirmativo.

4. **X.Se.encuentra.usted..realizando.alguna.actividad.que.le.genere.ingresos. / En.caso.de.ser.afirmativo..indique.tipo.de.actividad.y.su.frecuencia:** Al igual que el punto 1 y 3 unimos ambos campos y reemplazamos las respuestas afirmativas por los tipos de actividades expuestas.

Se encontraron incluidos en el data set campos donde no existía una estandardización en cuanto a los datos debido a que no seguían un estándar en especifico tal es el caso de:

1. **Fecha:** En este campo se observaron diversas fechas en distintos formatos por lo que se aplicó un proceso de estandardización tomando cada uno de las fechas de distintos formatos y convirtiéndolas en una fecha del tipo dd/mm/aa. Durante este proceso se observaron valores anómalos, los cuales en dos de los casos si se logró realizar la estandarización de manera exitosa, tales valores eran: 1051989 el cual se asumió como día: 10 mes: 5 y año 1989. y 21041994 el cual se asumió como día: 21, mes: 04 y año 1994. El ultimo caso si fue imposible llevarlo a este formato debido a que se observó alta inconsistencia, el cual es 19220485, por lo que se procedió a eliminar dicha instancia
2. **Año.que.cursa:** en este campo en todos los valores si se observó que seguían un estandar, de igual manera se tomo la decisión de llevarlo aun formato numérico para facilitar posibles cálculos posteriores.

3. **Número.de.materias.aprobadas.en.el.semestre.o.año.anterior**: en este campo se detecto un outlier en la fila 2, el cual era: mas de 10. Este valor no es un valor exacto y tomando en cuenta que el numero de el numero de materias inscritas eran 2, efectivamente si se trata de dato erróneo por lo que se procedió a eliminar dicha instancia.

4. **Promedio.ponderado.aprobado:** en este campo muchos de los valores no concuerdan con un promedio ya que eran valores enteros de 4 dígitos, por lo que se realizó modificaciones en dichos casos. Esto se llevo a cabo tomando los primeros dos dígitos de cada un de ellos, se colocaba un punto y luego se tomaba la otra parte del entero como parte decimal. No hubo necesidad de comparar si los primeros dos eran mayor que 20 debido a que se observo en el data set que cada uno de los primeros dos valores de cada numero y ninguno fue mayor a 20.

5. **Eficiencia**: se pudo observar que muchos datos no concordaban con una eficiencia ya que eran valores enteros de 4 dígitos y la eficiencia debe estar dentro del rango [0,1]. En estos casos, como el primer dígito siempre era mayor que uno, se le antepuso un cero para que estuviese acorde con un valor de eficiencia.

En los ingresos y egresos tanto del estudiante como los del responsable académico todos los valores descritos con NA y los datos que no contenían ningún valor se les asigno el valor por defecto 0, esto con la finalidad de estandarizar los campos.

En varios de los campos de ingresos y egresos del responsable académico se pudieron observar datos anómalos que no estaban acorde a el resto de los valores del campo, como por ejemplo valores que contenían la abreviatura “bs”, campos con comas y otros con puntos, entre otros. Todos estos valores fueron tratados y estandarizados quedando todos los valores en un mismo formato.