|  |  |  |  |
| --- | --- | --- | --- |
| Archivo | Encoding | Estudiantes | Columnas eliminadas |
| datasets/Rendimiento por estudiante 2010.csv | Utf-8 | 3.335.825 | ['SIT\_FIN\_R'] |
| datasets/Rendimiento por estudiante 2011.csv | Utf-8 | 3.326.746 | ['FEC\_ING\_ALU', 'SIT\_FIN\_R'] |
| datasets/Rendimiento por estudiante 2012.csv | Utf-8 | 3.308.477 | ['SIT\_FIN\_R'] |
| datasets/Rendimiento por estudiante 2013.csv | Utf-8 | 3.255.518 | ['GD\_ALU', 'COD\_TIP\_CUR', 'COD\_REG\_ALU', 'COD\_RAMA', 'SIT\_FIN\_R'] |
| datasets/Rendimiento por estudiante 2014.csv | latin | 3.227.534 | ['GD\_ALU', 'COD\_DEPE2', 'COD\_TIP\_CUR', 'COD\_REG\_ALU', 'COD\_RAMA'] |
| datasets/Rendimiento por estudiante 2015.csv | latin | 3.238.586 | ['COD\_DEPROV\_RBD', 'NOM\_DEPROV\_RBD', 'GD\_ALU', 'COD\_DEPE2', 'ESTADO\_ESTAB', 'COD\_TIP\_CUR', 'COD\_REG\_ALU', 'COD\_RAMA', 'SIT\_FIN\_R'] |
| datasets/Rendimiento por estudiante 2016.csv | Utf-8-sig | 3.226.943 | ['COD\_DEPROV\_RBD', 'NOM\_DEPROV\_RBD', 'COD\_DEPE2', 'ESTADO\_ESTAB', 'COD\_TIP\_CUR', 'COD\_DES\_CUR', 'COD\_REG\_ALU', 'COD\_RAMA', 'SIT\_FIN\_R'] |
| datasets/Rendimiento por estudiante 2017.csv | Utf-8 | 3.246.824 | ['COD\_DEPROV\_RBD', 'NOM\_DEPROV\_RBD', 'COD\_DEPE2', 'ESTADO\_ESTAB', 'COD\_TIP\_CUR', 'COD\_DES\_CUR', 'COD\_REG\_ALU', 'COD\_RAMA', 'SIT\_FIN\_R'] |
| datasets/Rendimiento por estudiante 2018.csv | Utf-8-sig | 3.293.750 | ['NOM\_REG\_RBD\_A', 'COD\_DEPROV\_RBD', 'NOM\_DEPROV\_RBD', 'COD\_DEPE2', 'ESTADO\_ESTAB', 'COD\_TIP\_CUR', 'COD\_DES\_CUR', 'COD\_REG\_ALU', 'COD\_RAMA', 'SIT\_FIN\_R'] |
| datasets/Rendimiento por estudiante 2019.csv | Utf-8-sig | 3.328.915 | ['NOM\_REG\_RBD\_A', 'COD\_DEPROV\_RBD', 'NOM\_DEPROV\_RBD', 'COD\_DEPE2', 'ESTADO\_ESTAB', 'COD\_TIP\_CUR', 'COD\_DES\_CUR', 'COD\_REG\_ALU', 'COD\_RAMA', 'COD\_MEN', 'SIT\_FIN\_R'] |

Encoding 2016, 2018, 2019 not working with “latin-1”

Primero que todo es necesario hacer un análisis exhaustivo de los datos de las tablas para agregar a nuestra propia base de datos y para limpiar los datos en el excel, por ende decidimos por un lado ver todos los datos de las columnas que contenía cada uno de los archivos por año en formato .csv, y por otro lado visualizar las columnas que contenía dicho archivo con la librería Pandas en Python.

Columnas funcionales que vamos a utilizar y que están disponibles en todos los años a analizar:

* AGNO
* RBD
* DGV\_RBD
* NOM\_RBD
* COD\_REG\_RBD
* COD\_COM\_RBD
* NOM\_COM\_RBD
* COD\_DEPE
* RURAL\_RBD
* COD\_ENSE
* COD\_GRADO
* LET\_CUR
* MRUN
* GEN\_ALU
* FEC\_NAC\_ALU
* COD\_COM\_ALU
* NOM\_COM\_ALU
* COD\_SEC
* COD\_ESPE
* PROM\_GRAL
* ASISTENCIA
* SIT\_FIN
* SIT\_FIN\_R

Columnas a eliminar definitivamente:

|  |  |
| --- | --- |
| Columna | Razón |
| LET\_RBD, NUM\_RBD | Dado que no pertenecen a ningún año para el análisis, sólo existen desde el año 2009 hacia atrás. |

Columnas a eliminar por análisis de descarte:

|  |  |
| --- | --- |
| Columna | Razón |
| NOM\_REG\_RBD\_A | Sólo aparece en los 2 últimos años y es poco significativo para el análisis de datos final. |
| COD\_DEPROV\_RBD, NOM\_DEPROV\_RBD | Estos datos sólo aparecen desde el año 2015 y creemos que es innecesario saber a que departamento provincial pertenece un establecimiento. |
| COD\_DEPE2 | Tenemos COD\_DEPE que es básicamente lo mismo, la diferencia es que la anterior aparece en todos los años y contiene 1 atributo más. En este caso que los datos sean agrupados, no cambia el resultado final. |
| ESTADO\_ESTAB | Sólo aparece desde el 2015. No es necesario saber en qué estado se encuentra un establecimiento para hacer el análisis de los datos. |
| COD\_TIP\_CUR | Sólo aparece desde el 2013. No es necesario saber si un curso es simple o combinado. Además en nuestro diagrama no aparece la entidad curso. |
| COD\_DES\_CUR | Dado que sólamente aplica a enseñanza media TP, es demasiado específico para un análisis general de los datos. Además los atributos de esta columna están más ligados a un establecimiento que a un curso en sí. |
| GD\_ALU | Sólo está disponible en los años 2013 a 2015, es muy poca cantidad de datos (menos de 1/3) considerando la totalidad general de los datos para un análisis. |
| COD\_REG\_ALU | Creemos que no es necesario saber a qué región pertenece un alumno si ya tenemos los datos a qué comuna pertenece un alumno, podemos obtener también los datos mediante el establecimiento. |
| COD\_RAMA | Los datos aparecen desde el 2013 y creemos que no es necesario saber que tipo de educación recibe un estudiante. |
| FEC\_ING\_ALU | Los datos sólo están disponibles en el año 2011 y como buenos diseñadores de bases de datos, sentimos que perjudicaría el agregarlo porque es redundante y el año de ingreso se puede repetir hasta mínimo 8 veces que es la cantidad de años que un estudiante se forma en un establecimiento educacional. |
| COD\_MEN | Son datos muy recientes y sólo está disponible en el último año en los datos. |

Columnas con posible eliminación (No estamos seguros si eliminar):

|  |  |
| --- | --- |
| Columna | Razón |
| COD\_ENSE2 | A diferencia de COD\_ENSE, este nos permite saber de forma agrupada qué tipo de enseñanza recibe el alumno, basica o media y es mucho más facil y rápido de visualizar en el análisis de datos, pero esta columna sólo está disponible desde el año 2014. |
| COD\_JOR | Pensamos eliminarlo porque los datos sólo están disponibles desde el año 2013 en adelante, pero un requerimiento de análisis en las preguntas pide saber a qué jornada asisten los alumnos. |
| EDAD\_ALU | Pensamos eliminarlo porque la verdad la mayoría de los alumnos tiene aproximadamente la misma edad y no es relevante para el análisis final. |
| INT\_ALU | Quizás si se quiere hacer un análisis más específico se podría obtener los datos de los alumnos con necesidades especiales, pero desde el año 2014 en adelante no se utilizan estos datos. |

Luego de finalizar el análisis anterior, decidimos guardar los datos de las columnas COD\_ENSE2 y COD\_JOR, dado que nos sirven para en análisis final y dar respuestas más concretas a las preguntas solicitadas por la universidad.

Para transformar los datos recopilamos decidimos crear nuestro propio diagrama de entidad relacionamiento detallando los datos contenidos en el PDF obtenido desde:

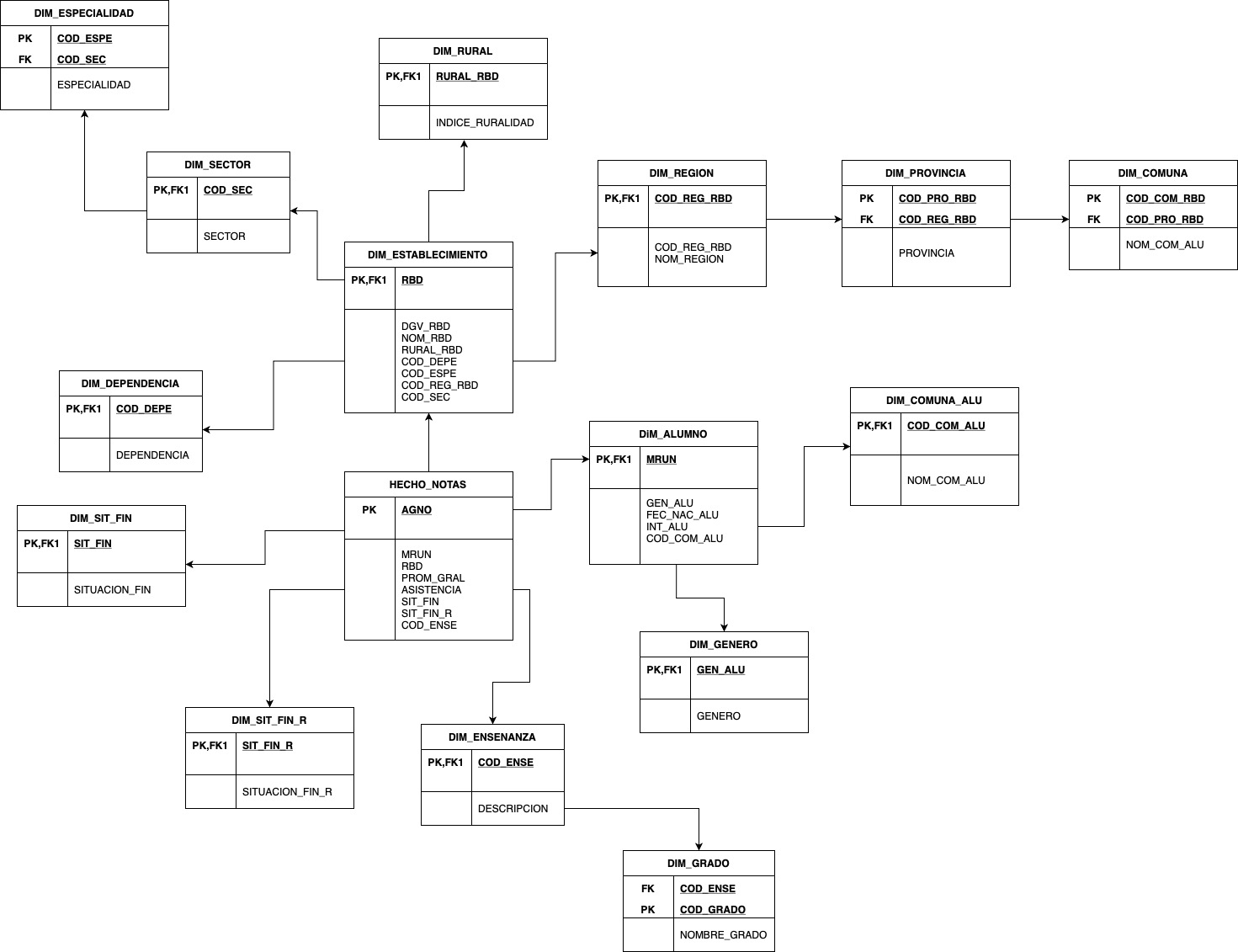


Diagrama de bases de datos multidimensional

Fuente: Elaboración de Equipo (draw.io)

Luego de hacer el diagrama nos dirigiremos a realizar el análisis del contenido de las tablas