APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA LA PREDICCIÓN DE LA DESERCIÓN EN ESTUDIANTES DE PREGRADO DE LA UNIVERSIDAD PONTIFICIA BOLIVARIANA, SEDE CENTRAL MEDELLÍN

EDGAR ANTONIO GARCÍA OSPINA

UNIVERSIDAD PONTIFICIA BOLIVARIANA

ESCUELA INGENIERÍAS

FACULTAD DE INGENIERÍA EN TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIÓN

MAESTRÍA EN TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIÓN

MEDELLÍN - 2019

APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA LA PREDICCIÓN DE LA DESERCIÓN EN ESTUDIANTES DE PREGRADO DE LA UNIVERSIDAD PONTIFICIA BOLIVARIANA, SEDE CENTRAL MEDELLÍN

EDGAR ANTONIO GARCÍA OSPINA

Trabajo de grado para optar al título de Maestría en Tecnologías de la Información y la Comunicación

Asesor

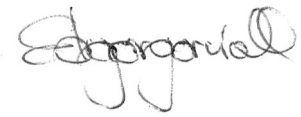
ANA ISABEL OVIEDO CARRASCAL

PhD. Ingeniería Electrónica

UNIVERSIDAD PONTIFICIA BOLIVARIANA
ESCUELA INGENIERÍAS
FACULTAD DE INGENIERÍA EN TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIÓN
MAESTRÍA EN TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIÓN
MEDELLÍN - 2019

DECLARACIÓN ORIGINALIDAD

"Declaro que esta tesis (o trabajo de grado) no ha sido presentada para optar a un título, ya sea en igual forma o con variaciones, en esta o cualquier otra universidad". Art. 82 Régimen Discente de Formación Avanzada, Universidad Pontificia Bolivariana.



FIRMA AUTOR

EDGAR ANTONIO GARCÍA OSPINA

C.C. 15.437.575

Medellín, 06 de diciembre de 2019

RESUMEN

La deserción en las instituciones educativas es un tema no sencillo de resolver ya que existen múltiples condiciones que pueden conducir a un estudiante a abandonar su proyecto de formación, con todas las implicaciones que esto conlleva. La aplicación de técnicas de minería de datos en el campo educativo facilita, entre otras, predecir una acción o un evento del entorno de aprendizaje, de forma que se tomen las acciones preventivas correspondientes en el momento oportuno. En este proyecto se pretende seguir una metodología de minería de datos para plantear, de acuerdo a la información disponible, un modelo que permita predecir la deserción en un caso de estudio relacionado con estudiantes de pregrado en la U.P.B., sede central. Se logró analizar de forma conjunta la información suministrada por Bienestar Universitario de una encuesta de caracterización realizada entre los años 2015 y 2016 complementada con datos del sistema central. Los modelos analíticos desarrollados son: 1) seleccionar los factores principales que influyen en la deserción según los datos disponibles, 2) perfilamiento de los estudiantes que desertan y los que no, 3) reglas de asociación entre los factores seleccionados, y por último 4) la búsqueda del mejor método predictivo de deserción. Se encontró que, los factores seleccionados, en su mayoría, son congruentes en los diferentes experimentos aplicados y que el árbol de decisión RandomForest se ajusta mejor para clasificar si un estudiante va a presentar deserción o no. Además, el conjunto de factores seleccionados puede servir para apoyar otros esfuerzos que se tengan en la universidad relacionados con la permanencia estudiantil.

PALABRAS CLAVE: Deserción, Educación superior, Minería de datos educativa, Predicción

ABSTRACT

Dropout in educational institutions is not a simple matter to solve since there are multiple conditions that can lead a student to abandon their training project, with all the implications that this entails. The application of data mining techniques in the educational field makes it possible, among others, to predict an action or event in the learning environment, so that the corresponding preventive actions could be taken at the appropriate time. This project aims to follow a data mining methodology to propose, according to the available information, a model that allows predicting dropout in a case study related to undergraduate students at the U.P.B., headquarters. It was possible to jointly analyze the information provided by University Welfare department corresponding to a characterization survey carried out between 2015 and 2016, complemented with data from the ERP system. The analytical models were carried out to: 1) select the main factors that influence dropout according to the available data, 2) student profiling of who decide to drop out and those who do not, 3) rules of association between the selected factors and finally, 4) selection for the best predictive method. It was found that, the selected factors are mostly congruent in the different experiments applied, and that the RandomForest decision tree fits better to classify whether a student is going to drop out or not. In addition, the set of selected factors can be used to support other efforts in the university related to student permanence.

KEY WORDS: Drop-out, Higher education, Educational datamining, Prediction

TABLA DE CONTENIDO

LISTA DE ILUSTRACIONES

LISTA DE TABLAS

INTRODUCCIÓN

El desarrollo del presente trabajo se encuentra en el ámbito educativo, aplicando minería de datos para la predicción, lo cual permitirá conocer detalles sobre variables y obtener conocimiento que a simple vista no es observado. El encontrar patrones que conduzcan en la predicción de posibles deserciones es de suma importancia, dado que permite enfocar esfuerzos tanto universitarios como del mismo estudiante para lograr que se avance en su proyecto de vida. Encontrar esos patrones y lograr la retención repercutirá de manera positiva tanto a nivel social, institucional y personal.

El trabajo se desarrolla inicialmente con datos obtenidos a partir de una encuesta llamada perfil integral, realizada en los años 2015 y 2016 a estudiantes de la Universidad Pontificia Bolivariana, Sede Central. Esta información se complementa con las condiciones académicas del estudiante en la universidad.

El desarrollo del presente trabajo consta de dos partes. En la primera parte se presenta la formulación del proyecto y se realiza la revisión literaria en la aplicación de métodos para establecer modelos de deserción estudiantil. En la segunda parte, se presenta el diseño de la solución y se da paso al desarrollo de la metodología de minería datos según CRISP-DM. Finalmente, se presentan algunas conclusiones y la posibilidad de realizar trabajos futuros.

PARTE 1. FORMULACIÓN DEL PROYECTO Y REVISIÓN LITERARIA

1. DESCRIPCIÓN DEL PROBLEMA

Para el Ministerio de Educación Nacional (MEN), un estudiante se considera desertor si abandona la Institución de Educación Superior (IES) durante dos períodos consecutivos. En la Universidad Pontificia Bolivariana, sede Medellín, la deserción oficial reportada ante el Sistema para la Prevención y Análisis de la Deserción en las Instituciones de Educación Superior (SPADIES) se ha venido comportado de la siguiente forma:

Tabla 1. Índice de deserción por cohorte 2006-2010

	2006	2007	2008	2009	2010
UPB Sede Central	49%	44%	40%	46%	49%
A nivel nacional	31%	31%	34%	35%	38%

Fuente: Ministerio de Educación Nacional – SPADIES. Consulta realizada el 28 de mayo de 2018

Aunque la cobertura nacional para acceder a la educación superior ha aumentado en los últimos años (MEN, 2016), también se evidencia que ha derivado cambios en las características socioeconómicas y académicas de los recién ingresados a la universidad. Estos cambios llevan a que los riesgos de deserción sean mayores dadas las nuevas vulnerabilidades, especialmente financieras y académicas (González Fiegehen, 2006; Torres & Zúñiga, 2012).

Entre los motivos de deserción evidenciados por la Dirección de Registro Universitario se encuentran las siguientes:

Tabla 2. Motivos de retiro Sede Central Medellín 2010-2016

Motivos de retiro	2010	2011	2012	2013	2014	2015	2016	Tendencia
Sede Central Medellín								
Académicos	129	174	143	138	280	302	243	**26.3**
Institucionales	245	163	232	48	215	165	125	**-13.3**
Socio-económicos	29	27	6	5	1	3	23	**-2.5**
Personales	96	155	209	124	124	145	102	**-3.1**

Fuente: Informe de autoevaluación institucional 2016 tomo tres – UPB.

Los motivos académicos incluyen: cambios de universidad, de carrera, cancelación de matrícula, y culminación de trabajo de grado; los motivos socioeconómicos se refieren a dificultades económicas y de ubicación laboral, mientras los personales hacen referencia a: viajes al exterior, cambio de ciudad, enfermedad, fallecimiento y dificultades familiares.

Como principal acción que toma la sede central para disminuir la deserción, es crear en el año 2015 el Programa Institucional de Permanencia, con el propósito claro de que los estudiantes terminen con éxito su proceso de formación. Entre las responsabilidades del programa están el diseño, desarrollo e implementación de estrategias que reduzcan las tasas de deserción trabajando en los componentes psicosocial, socioeconómico, académico, físico y espiritual.

Liderado por dicho programa, la Universidad está implementando una herramienta computacional para generación de alertas tempranas basada en las respuestas y la categorización que se le da a cada de ellas. Este producto tiene dos componentes básicos, por un lado, una encuesta de caracterización de 66 preguntas; que se ha aplicado, por disposición de Bienestar Universitario, a los estudiantes nuevos de los dos últimos semestres. Por el otro, resultados académicos que se extraen del sistema central.

Además, entre los años 2015 y 2016 la sede central, bajo el liderazgo de la Unidad de Bienestar Universitario, aplicó una encuesta a sus estudiantes de pregrado llamada Perfil Integral, la cual constó de aproximadamente 260 preguntas de caracterización sociodemográfica, familiar, económica, entre otras; a un importante número de estudiantes de pregrado.

Con todos estos esfuerzos se hace necesario realizar los análisis apropiados para aumentar la certeza que con la información que se está construyendo se pueda contribuir de manera positiva en la prevención de la deserción.

2. JUSTIFICACIÓN

Los datos recolectados a partir de los diferentes esfuerzos de caracterización mencionados anteriormente pueden contener información valiosa que, agregando otros componentes, pudieran servir para identificar variables relevantes que permitan tomar acciones en la mitigación de los riesgos de deserción. Esto por ahora no se sabe, dado que no se han realizado análisis de relacionamiento riguroso para tratar de descubrir información oculta entre esa nube de datos.

Se tiene entonces la necesidad de investigar acerca de los análisis que deben considerarse alrededor de los datos conseguidos. Esto nos permitirá, principalmente, responder de una manera más integral y proactiva respecto a la atención que los mismos estudiantes merecen. Por otro lado, se aumentará la precisión en los resultados haciendo más efectivos los programas de intervención a implementar. Derivado de todo lo anterior, se espera incrementar el nivel de confianza por parte de los actores interesados, para que las sugerencias o herramientas resultantes de este trabajo se hagan más fácil de sostener en el tiempo.

Una vez se logre conocer a los estudiantes con más precisión, en las dimensiones que resultasen ser relevantes; se podrá, por ejemplo, diseñar metodologías de enseñanzas más acordes a la forma en como dichos estudiantes aprenden. Esto aumentará necesariamente el desempeño académico, que a su vez se verá reflejado en mejoras en los rankings institucionales. Y no es sólo el desempeño académico, es mejorar las probabilidades de terminar un programa de estudio, que a futuro impactará de manera positiva las condiciones de vida de dichos profesionales.

Desde el punto de vista de la institución educativa en resultados tangibles, esas mejoras en desempeño académico, más unas áreas trabajando de forma articulada harán que; por un lado, se disminuya la deserción, lo que traerá consigo mejoras en

los ingresos por concepto de matrículas en el mediano y largo plazo y, por el otro, haya menos costes operativos.

Otro aspecto importante es que en los procesos de acreditación de alta calidad institucional y de los programas mismos se analizan los niveles de deserción. Por lo tanto, será importante mostrar los esfuerzos que se realizan en pro de este aspecto, dada las exigencias por parte del Consejo Nacional de Acreditación (CNA) (MEN, 2009).

Conocer mejor a los estudiantes, orientará a la Universidad en ofrecer lo que realmente necesitan, de la forma más adecuada y en el momento oportuno. Al aplicar técnicas de minería de datos se puede implementar un sistema que alerte a los diferentes actores acerca del riesgo de deserción de los estudiantes, de forma que se puedan tomar acciones proactivas más eficaces (Márquez Vera, Romero Morales, & Ventura Soto, 2012)

3. OBJETIVOS

Objetivo General

Desarrollar un modelo predictivo utilizando técnicas de minería de datos que permita anticiparse a la deserción en estudiantes de pregrado de la Universidad Pontificia Bolivariana.

Objetivos Específicos

Los objetivos específicos nombrados hacen referencia a las fases del proceso de minería de datos

- Preparar los datos relacionados con un caso de estudio de estudiantes de pregrado.
- Diseñar un modelo de predicción de deserción.
- Evaluar el modelo de predicción.
- Brindar recomendaciones respecto a variables determinantes de deserción según resultados obtenidos.

4. METODOLOGÍA

Para el desarrollo del proyecto se plantea una orientación metodológica principalmente cuantitativa, aunque también serán necesarias técnicas cualitativas. Con respecto a la primera parte, se recolectarán datos de los estudiantes provenientes de los sistemas de información, y lo que no se encuentre disponible y sea relevante para el cumplimiento de los objetivos, será necesario construir de manera propia. A esta parte cuantitativa se le aplicará técnicas de minería de datos, análisis estadístico, y se elegirán los algoritmos más apropiados de acuerdo al objetivo específico a cumplir. Por otro lado, se plantea el uso de posibles entrevistas a los actores que se consideren claves y nos permitan realizar validaciones pertinentes de acuerdo a los resultados esperados.

Se propone seguir las fases planteadas en CRISP-DM. En cada una de las seis fases están planteadas se sugieren ciertas tareas específicas, que se seguirán de acuerdo a la evolución del desarrollo del proyecto.

Comprensión del negocio: En esta fase se plantean realizar entrevistas para entender mejor los requerimientos planteados en el proyecto e identificar fuentes de información existentes.

Comprensión de los datos: Una vez se tengan concedidos los permisos de acceso a la información existente, se definirá el conjunto inicial y las características básicas. Se plantea utilizar herramientas de perfilamiento de datos como DQ Analyzer, a su vez que técnicas estadísticas.

Preparación de datos: Será necesario definir las variables relevantes de acuerdo a cada objetivo a cumplir. Se plantea entonces realizar análisis de correlación para seleccionar los datos apropiados al igual que la aplicación de técnicas de limpieza

de datos. No menos importante, se delimitará los periodos de tiempo con los cuales se trabajará la muestra de datos.

Modelamiento: Se implementarán herramientas de minería predictivas (técnicas supervisadas). Se aplicará por ejemplo árboles de decisión, K-means u otras, según sea el caso.

Evaluación: Se evaluarán los resultados de los modelos con los mismos datos existentes y en algunos casos podrá ser necesario realizar algún trabajo de campo para validar que una predicción fue la correcta; por ejemplo, encuestas a población desertora. Puede ser que en esta fase sea necesario volver a fases anteriores para refinar los datos y modelos.

Despliegue: Se instalarán los modelos a las personas que se definan como público objetivo, ya sea un grupo de docentes o personal administrativo. Será necesario hacer claridad en posibles parametrizaciones o ajustes para que los resultados tengan un nivel de confianza mínimamente aceptable.

5. MARCO TEÓRICO

5.1 Deserción

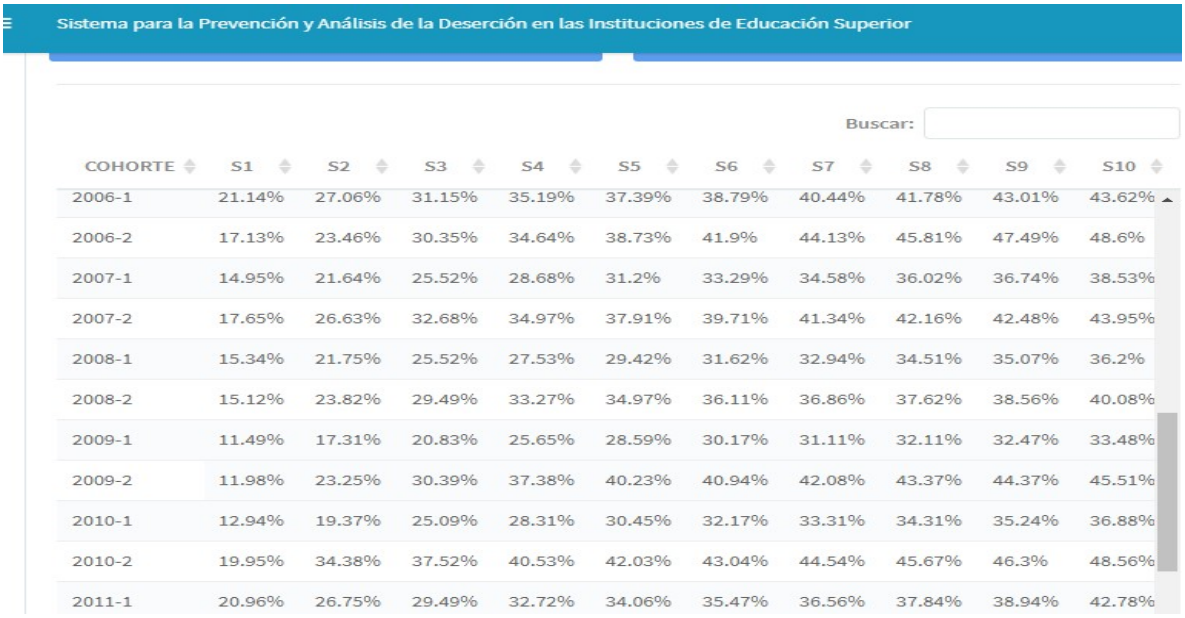
Para comenzar, se debe aclarar un poco más lo que significa deserción. Entre las diferentes definiciones se destaca la data por (Tinto, 1989), la que se analiza desde diferentes perspectivas: individual, institucional y gubernamental.

Deserción individual hace referencia a que un estudiante llega a la educación superior, pero por diferentes motivos abandona sus metas inicialmente propuestas. Según el autor, puede estar relacionada con la incapacidad de la institución universitaria para retener dicho estudiante. Desde la perspectiva institucional no es más que el número de estudiantes que dejan la universidad antes de obtener el título profesional. Y desde el lado gubernamental, se define como el abandono de todo el sistema educativo.

Para hablar en términos prácticos, y como objeto del presente proyecto, se trabajará en términos de deserción institucional, dado que los datos que se esperan analizar corresponderán a los estudiantes que están matriculados en el contexto UPB.

Otro concepto es la deserción por cohorte. Para esto se tiene en cuenta la definición dada por el gobierno colombiano (MEN, 2014), donde dice que se registra de manera acumulada el porcentaje de estudiantes que abandonaron en relación con el mismo periodo de inicio. Para ilustrar, se presenta una consulta directamente al sistema SPADIES en la tabla 3.

Tabla 3. Sistema SPADIES - Porcentaje de estudiantes que abandonan



Sistema para la Prevención y Análisis de la Deserción en las Instituciones de Educación Superior

Buscar:

COHORTE	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
2006-1	21.14%	27.06%	31.15%	35.19%	37.39%	38.79%	40.44%	41.78%	43.01%	43.62%
2006-2	17.13%	23.46%	30.35%	34.64%	38.73%	41.9%	44.13%	45.81%	47.49%	48.6%
2007-1	14.95%	21.64%	25.52%	28.68%	31.2%	33.29%	34.58%	36.02%	36.74%	38.53%
2007-2	17.65%	26.63%	32.68%	34.97%	37.91%	39.71%	41.34%	42.16%	42.48%	43.95%
2008-1	15.34%	21.75%	25.52%	27.53%	29.42%	31.62%	32.94%	34.51%	35.07%	36.2%
2008-2	15.12%	23.82%	29.49%	33.27%	34.97%	36.11%	36.86%	37.62%	38.56%	40.08%
2009-1	11.49%	17.31%	20.83%	25.65%	28.59%	30.17%	31.11%	32.11%	32.47%	33.48%
2009-2	11.98%	23.25%	30.39%	37.38%	40.23%	40.94%	42.08%	43.37%	44.37%	45.51%
2010-1	12.94%	19.37%	25.09%	28.31%	30.45%	32.17%	33.31%	34.31%	35.24%	36.88%
2010-2	19.95%	34.38%	37.52%	40.53%	42.03%	43.04%	44.54%	45.67%	46.3%	48.56%
2011-1	20.96%	26.75%	29.49%	32.72%	34.06%	35.47%	36.56%	37.84%	38.94%	42.78%

Al ubicarse en la última fila y realizar el cruce con la columna marcada como S10, se observa 42.78%. Se entenderá entonces que de los estudiantes que ingresaron por primera vez en el semestre 2011-1, y luego de transcurridos 10 semestres, ya habían abandonado el 42.78% de esos estudiantes.

También es importante la clasificación de acuerdo al momento en que se produce la deserción como se presenta en la ilustración número 1 tomada directamente del (MEN, 2009):

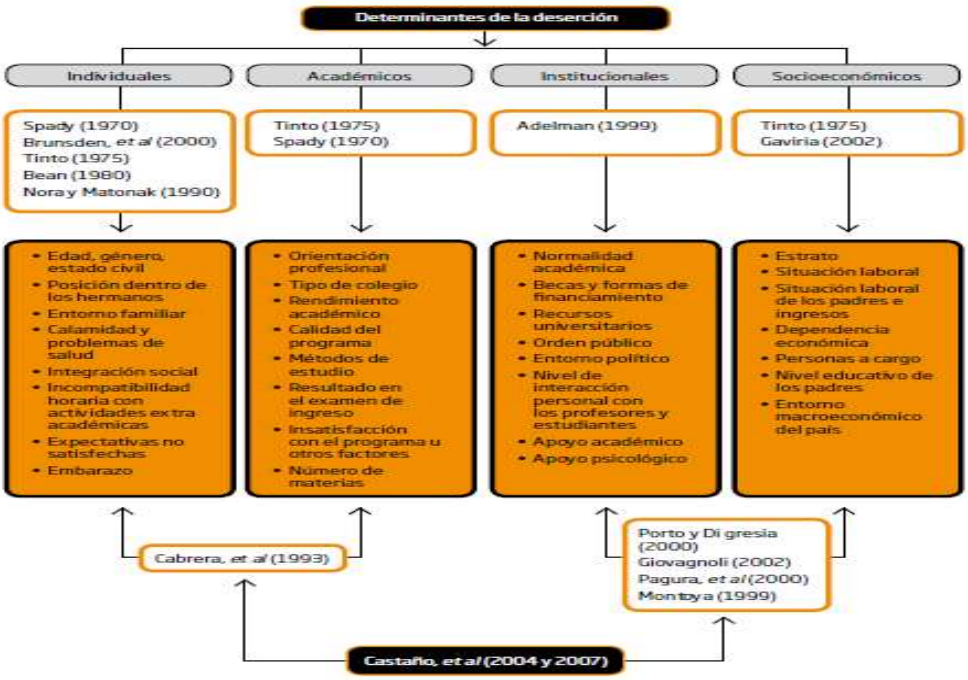
Ilustración 1. Tipos de deserción

Fuente: Adaptado de Castaño, *et al* (2004).

Si la deserción ocurre antes de iniciar a estudiar, se llamará deserción precoz, si ocurre hasta mitad del programa académico se llamará deserción temprana y si por el contrario después de la mitad, se llamará deserción tardía.

Haciendo un énfasis en las variables que deberían analizarse, existen estudios que hacen referencia a la siguiente categorización (MEN, 2009), como se puede ver en la siguiente ilustración.

Ilustración 2. Variables determinantes de deserción según el MEN



Más importante que la categorización mostrada por los diferentes autores, son las variables en sí mismas. Ellas dan una pista de los datos que se deben tener en cuenta y que son determinantes en la deserción.

5.2 Minería de datos

Pasando a otro concepto, se define la minería de datos como un conjunto de tareas no muy fáciles de entender, donde se trata de hacer descubrimientos importantes que no resaltan a simple vista. (Frawley, Piatetsky-Shapiro, & Matheus, 1992). Se dice que los datos por sí solos raramente muestran beneficios, el verdadero provecho se obtiene al analizarlos y que de alguna manera soporte la toma de decisiones (Riquelme, Ruiz, & Gilbert, 2006). Entonces de eso se trata la minería: tomar unos datos, entenderlos, depurarlos y aplicarles unas técnicas, de forma que se obtengan modelos que sean valiosos para tomar una acción con respecto a ellos.

Para aplicar las técnicas de minería es necesario identificar el tipo de análisis a realizar. Si lo que se busca es predecir el comportamiento futuro, se utilizan técnicas supervisadas; o si se quiere realizar una descripción de un conjunto de datos, se utilizan técnicas no supervisadas. Cada una de ellas se componen de diferentes grupos de algoritmos dependiendo de la técnica a utilizar (Han, Kamber, & Pei, 2011).

Las técnicas supervisadas, o de aprendizaje predictivo, precisan un conjunto de algoritmos que hacen necesario un conocimiento previo del comportamiento de los datos, de forma que los resultados de un modelo puedan ser validados de forma positiva o negativa utilizando unos datos de muestra con el cual el sistema aprende. Las técnicas más usuales son: árboles de decisión, redes neuronales, máquinas de soporte vectorial y métodos de regresión.

Las técnicas no supervisadas, por el contrario, no suponen un conocimiento previo de los datos. Se utilizan principalmente para clasificar o perfilar un conjunto de datos de acuerdo a características descubiertas en el proceso. Los modelos se crean a partir del reconocimiento de patrones. Las técnicas más usuales son: análisis de correlación, K-means y A-priori.

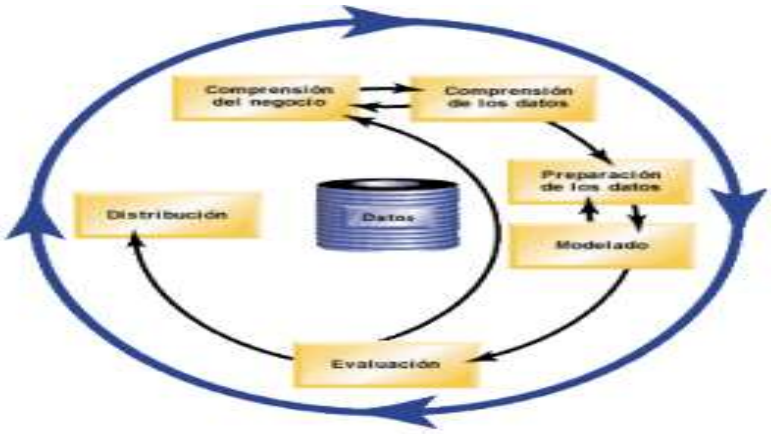
Lo anterior describe el concepto de minería de datos y algunas técnicas de acuerdo al conocimiento previo de los datos y el propósito del modelo a implementar. Qué técnica utilizar dependerá de las variables de entrada y del resultado esperado.

La aplicación de numerosas técnicas y aplicaciones de minería de datos en el campo educativo han dado paso a un concepto llamado minería de datos educacional. Inicialmente estas aplicaciones han apoyado propósitos operativos para las instituciones educativas, pero también pueden ser de gran provecho en los métodos de enseñanza (Baepler & Murdoch, 2010).

5.3 Metodología CRISP-DM

Ahora es importante conocer un poco acerca de una metodología que permite organizar el proceso en los diferentes momentos de tiempo de implementación de un proyecto de minería de datos. CRISP-DM, de las siglas en inglés CRoss-Industry Standard Process for Data Mining, ayuda en la orientación de la ejecución de trabajos de minería de datos (Chapman et al., 1999). Tiene un componente de metodología que define una serie de fases y tareas en cada fase. Las fases conforman el ciclo vital de un proyecto, que se esquematizan de la siguiente forma:

Ilustración 3. Metodología CRISP-DM



(IBM, 2012). Manual de CRISP-DM de IBM SPSS Modeler. Recuperado desde ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/es/CRISP-DM.pdf

Como se observa, el ciclo vital consiste de 6 fases, sin un orden específico entre ellas. Esto le da una flexibilidad para adaptarse a las necesidades específicas de cada modelo.

Tiene amplio uso, lo que la convierte en un estándar (Wirth & Hipp, 2002). Los seis pasos inician con la comprensión del negocio, su organización y las situaciones alrededor del problema; luego se entienden los datos y se les realiza análisis de calidad de los mismos; posteriormente se preparan, haciéndoles las limpiezas necesarias y aplicando transformaciones de formato según sea necesario; después se procede con la creación de un modelo que apunte al fin perseguido, se realizan las pruebas necesarias de acuerdo a las tareas definidas; ya con resultados se continúa con la evaluación para establecer si los resultados tienen razón de ser y en realidad ayudan en la solución del problema; dado que sí, se avanza con la fase de implementación (Chapman et al., 1999)

6. MARCO LEGAL

De acuerdo con la Carta Magna colombiana (Constitución Política, 1991), se establece en el artículo 69, la autonomía que tienen las universidades para crear directivas y estatutos siempre y cuando estén bajo la ley. Esto quiere decir que las universidades tienen la potestad y autonomía para regular los procesos académicos y administrativos, y llevando esto a nuestro contexto, están en libre posición para implementar mecanismos tendientes a garantizar la permanencia estudiantil y los planes de acción relacionados. En otras palabras, no se podrían esperar que se implementen leyes donde se establezcan cifras máximas de deserción. Sólo se podrán sugerir modelos de apoyo o guías tendientes a aumentar la permanencia del estudiante en la institución educativa.

Lo anteriormente dicho se soporta en la Ley 30 de 1992, Art. 4, en donde en sus apartados finales refiere: "… Por ello, la Educación Superior se desarrollará en un marco de libertades de enseñanza, de aprendizaje, de investigación y de cátedra".

También se esclarece un poco más el tema diciendo que el Estado, la sociedad y la familia deben trabajar de forma conjunta por la calidad de la educación, y que el Estado debe atender en forma permanente los factores que favorecen la calidad y el mejoramiento de la educación, Ley 115 de 1994, Art. 4.

Posteriormente se exige que la educación superior tenga un nivel de calidad de acuerdo con el mercado. Para esto se implementa el Sistema de Aseguramiento de la Calidad de la Educación (SACES) donde le compete al Ministerio de Educación Nacional operarlo de común acuerdo con las instituciones, docentes, estudiantes, pares académicos, científicos y demás organismos privados y oficiales, Ley 1188 de 2008.

En el año 2010, por medio del Decreto 1295, se reglamentan los registros calificados para los programas de educación superior en Colombia. Esto quiere decir que para un programa impartir un programa de educación en Colombia es requisito contar

con registro calificado otorgado por el MEN. En el artículo 6, del mismo decreto hace referencia a que deben existir directrices documentadas a nivel institucional donde se describan mecanismos y criterios para selección, permanencia, promoción y evaluación de docentes y estudiantes, de acuerdo a lo previsto en la Constitución y la ley.

Para conectar lo anterior a nuestro contexto, la U.P.B. fue establecida en la Arquidiócesis de Medellín mediante Decreto Arzobispal No. 124 del 15 de septiembre de 1936 y le fue reconocida la personería jurídica civil mediante Resolución Ejecutiva No. 48 el 22 de febrero de 1937 y por Resolución No. 021 del 21 de abril de 1959 del Ministerio de Trabajo, fue reconocida como establecimiento sin ánimo de lucro.

Como se establece en los Estatutos Generales universitarios del día 2 de septiembre de 2013, en su artículo 1, es una institución católica con el ejercicio de una misión pastoral que propicia el avance científico, mediante la investigación y la enseñanza. Como lo faculta la ley, desarrolla de forma autónoma sus programas académicos y proceso administrativos. Con todo lo anterior se ratifica que no existe reglamentación de ley alguna que obligue al cumplimiento específico en lo relacionado a indicadores de permanencia.

Sin embargo, y como un instrumento destacado, en el año 2015 fue publicado la Guía para Implementación de Modelo de Permanencia y Graduación Estudiantil en Instituciones de Educación Superior; en la que básicamente se dan recomendaciones para que se posicionen en las agendas de cada institución políticas, fases, componentes y herramientas tendientes a garantizar la permanencia de los estudiantes. Este modelo fue el resultado de trabajo conjunto entre diferentes IES oficiales y privadas, Ministerio de Educación Nacional, pares académicos y miembros del Consejo Nacional de Acreditación (CNA).

Dado que la población universitaria puede estar conformada en sus periodos iniciales por estudiantes menores de edad, se debe considerar los principios, derechos y obligaciones planteados en el Código de la Infancia y Adolescencia. En

especial se debe hace énfasis en que las instituciones públicas o privadas no podrán imponer medidas; que, de alguna manera, afecten a los estudiantes tanto física como psicológicamente (Congreso de Colombia, 2006).

Por parte de la protección de datos personales, se hace importante respetar las disposiciones que en esta materia hace referencia la conocida ley del Hábeas Data (Congreso de Colombia, 2008). En dicha ley se debe considerar los principios establecidos y términos bajo los cuales circula la información personal, de forma que no se cause perjuicio alguno a los estudiantes.

En términos de objetivos sociales y económicos, se recibe un informe por parte de la Organización para la Cooperación y el Desarrollo Económicos (OCDE) en cual destaca que Colombia debería reducir la deserción escolar en la educación secundaria argumentando que la enseñanza debe reorientarse en términos prácticos del mercado laboral involucrando el sector productivo en este proceso (*Estudios Económicos de la OCDE Colombia www.oecd.org/eco/surveys/economic-survey-colombia.htm*, 2017).

En dicho informe también se recalca que Colombia debería realizar mayores esfuerzos en términos de presupuesto para la educación. Se recalca que, al realizar una mayor inversión en la educación primaria, se pueden reducir sustancialmente las tasas de abandono en la educación secundaria, y de paso se mejora los resultados en el rendimiento estudiantil (Heckman, 2008).

Aunque no se hace alusión directamente en el informe de la OCDE a la deserción estudiantil en nivel superior, un joven que no culmine sus estudios secundarios no podrá acceder a un nivel educativo en nivel terciario y por lo tanto están directamente relacionados. Es decir, las recomendaciones que hace la OCDE en términos de deserción en los niveles de primaria y secundaria también impactan la no graduación en instituciones de educación superior.

7. ESTADO DEL ARTE

7.1 Estudios de deserción a nivel internacional

La predicción de la deserción en instituciones de educación superior y los factores que influyen en la misma es un tema que se viene estudiando a nivel mundial desde los años 70s (Reason, 2009). Uno de los autores destacados para esa década (Astin, 1975) presenta las condiciones que hacen que un estudiante aumente las probabilidades de culminar los estudios, tomando como factores clave las características al momento de ingresar a una institución de educación superior tales como género, edad y lugar de residencia, así como las características de la institución, básicamente: tipo, ubicación y proceso de admisión.

Otro autor plantea, en esa misma década, (Tinto, 1975) un modelo teórico para incrementar la permanencia basado en la integración de los componentes social y académico del estudiante con su institución. Hace especial énfasis en el cuidado que se debe tener en las posibles variables que pueden determinar la decisión de deserción por parte del estudiante y los retiros impartidos por las mismas instituciones.

Años más tarde se vinieron incluyendo otros tipos de variables. Por ejemplo, el promedio de notas de los estudios secundarios y el desempeño en el examen de admisión a la universidad mostraban ser valiosos para predecir la permanencia de los estudiantes (Astin, Korn, & Green, 1987). Estos investigadores entrevistaron a aproximadamente 8,000 estudiantes, cruzando variables propias del estudiante con los resultados académicos en la secundaria y resultados en el examen de admisión. Usaron análisis de regresión para concluir que estas dos últimas variables eran sus predictores de mayor peso. Esta conclusión se seguía afirmando a finales de la década de los 90's después de diferentes estudios (Levitz, Noel, & Richter, 1999).

El uso de minería de datos en el campo educativo aparece en el año 2003, cuando se aplican diferentes técnicas de aprendizaje de máquina para predecir la deserción en un curso impartido a distancia (Kotsiantis, Pierrakeas, & Pintelas, 2003). Los

investigadores utilizaron variables tanto de caracterización individual, socioeconómicas, académicas, como de asistencia a las tutorías presenciales. Los algoritmos predictivos fueron entrenados con un conjunto de datos y probados con otros grupos diferentes. Se observaron mejores resultados con algunos algoritmos a medida que se incluía más información llegando a niveles de precisión del 83%.

Existen determinantes académicos que también pueden ser importantes a tener en cuenta. Muestra de ello es un proyecto llevado a cabo en la Universidad de Purdue, USA, donde, al igual que en ejercicio anterior, se tuvo en cuenta la variable asistencia a clases. El proyecto tuvo ciertos inconvenientes en ser implementado dada la dinámica de disponibilidad inmediata de los datos y el acceso a información que pudiera ser sensible por parte de los estudiantes, pero se muestra internacionalmente como un caso exitoso para aumentar la permanencia (Jisc, 2013). Se realizó con análisis predictivo mostrando alertas tempranas basado en semáforos.

Otro avance, en ese mismo sentido, se muestra cuando (Bayer, Bydžovs, Eryk, Aš Obšívač, & Popelínsk, 2012) enriquecieron la caracterización básica de 775 estudiantes con datos provenientes del comportamiento en redes sociales. Para ello utilizaron técnicas de minería de texto para buscar interrelaciones entre correos electrónicos, archivos, contenidos de cursos, etc. con temas como amistades, temas de conversación, comentarios, gustos, entre otros.

Esto les permitió identificar su red de amigos, representado por un sociograma, y las características de los mismos. Trabajaron con diferentes algoritmos como árboles de decisión, máquinas de soporte vectorial, clasificador bayesiano, entre otros. En definitiva, lo que obtuvieron fue una mayor precisión en la predicción de la deserción.

Saurabh Pal en la India (Pal, 2012), también utilizó el algoritmo con técnicas bayesianas para clasificar si un estudiante universitario pudiera desertar o no, teniendo en cuenta que este algoritmo es especialmente adecuado cuando se tiene una alta dimensionalidad de variables de entrada.

Con respecto a las aplicaciones técnicas de minería de datos en el campo de la deserción estudiantil se encuentran otros diferentes trabajos alrededor mundo, especialmente en instituciones que ofrecen educación a distancia, dado sus altos riesgos en este sentido. Para esa metodología de enseñanza se hace esencial contar con este tipo de análisis (Kotsiantis, Patriarcheas, & Xenos, 2010; Yang, Sinha, Adamson, & Rose, 2017).

Estos últimos autores analizaron las cohortes medidas en semanas de unos cuantos cursos en línea abierto masivamente (MOOC) de Coursera. Comenzaron aplicando análisis de redes sociales y terminaron detectando de forma exploratoria que, las primeras cohortes eran las menos sensibles a desertar. Justificaron que los estudiantes que iniciaban el curso en cohortes tardías tenían dificultades para participar en las comunidades de discusión.

Un caso referenciado a nivel latinoamericano, son las aplicaciones de técnicas de minería de datos tanto a nivel descriptivo como predictivo, en la Pontificia Universidad Católica de Valparaíso, Chile; donde los investigadores Olaya Ocaranza y Mónica Quiroz desarrollaron todo un programa de apoyo con metodología integradora para favorecer la adecuada inserción de los estudiantes a la vida universitaria con un consecuente aumento en la permanencia (Ocaranza & Quiroz, 2006). Se basa en estrategias para disminuir la deserción temprana, partiendo de una caracterización para todos los estudiantes de primer semestre, inducción a los novatos, tutorías, becas y apoyos psicoeducativos. Tuvieron en cuenta variables como promedio de notas en la educación media, tipo de colegio, nivel educativo y situación económica de los padres. Lograron un nivel de acierto del 87%.

7.2 Estudios de deserción a nivel local

A nivel colombiano se encuentran varios estudios, entre ellos está uno realizado en la Universidad de Nariño (Timarán Pereira, 2013), el cual se encuentra referenciado a nivel internacional, donde muestra el uso de diferentes técnicas sobre los datos

recolectados durante 15 años. Aunque el objetivo era mostrar que la herramienta propia utilizada era fiable, se describe cómo aplicaron árboles de decisión y otras técnicas de clasificación para llegar a un modelo de patrones de deserción.

Otro estudio local, aunque no precisamente de deserción, se realizó en la Universidad Popular del Cesar (Oñate Bowen, 2016) para predecir bloqueo académico para estudiantes los primeros cuatro semestres. Se concluyó que el algoritmo K-Means modelaba mejor las características de los estudiantes y, por el otro lado, los árboles de decisión presentaron un mejor desempeño en comparación con las redes bayesianas al momento de predecir qué estudiantes presentarán bloqueo académico para el siguiente semestre.

En términos propiamente de la U.P.B., se han realizado análisis exploratorios de deserción (Vélez Martínez, 2016), pero no específicamente utilizando técnicas de minería de datos. Se resalta el reconocimiento que le hace el M.E.N. a la seccional Bucaramanga con su Programa de Acompañamiento Académico – P.A.C., donde destaca sus líneas de acción (MEN, 2009), pero ninguna de ellas muestra que aplique algoritmos o técnicas de minería descriptiva o predictiva. Este programa fue creado en el año 2001 como respuesta a facilitar el tránsito de los estudiantes del colegio a sus dos primeros semestres de vida universitaria. Inicialmente se realizan intervenciones grupales al inicio del semestre y otra después de exámenes parciales. De los estudiantes que se identifiquen como de alto riesgo, se citan para realizar acompañamiento individual y así prevenir una posible deserción.

7.3 Discusión

En el campo de minería de datos para propósitos educativos, se encontró que básicamente es un área relativamente nueva, donde tanto a nivel internacional como nacional se ha explorado y aplicado en diferentes instituciones de educación superior. Se visualiza, que hay un espacio amplio para la investigación en este campo de la minería en el ámbito colombiano.

Lo que se observó en términos prácticos es la aplicación, en la mayoría de los casos, de técnicas supervisadas para que, utilizando un conjunto de datos, el modelo

pueda aprender y logre una buena clasificación como predicción. Es importante seleccionar unas buenas variables predictoras y aplicarles las técnicas apropiadas.

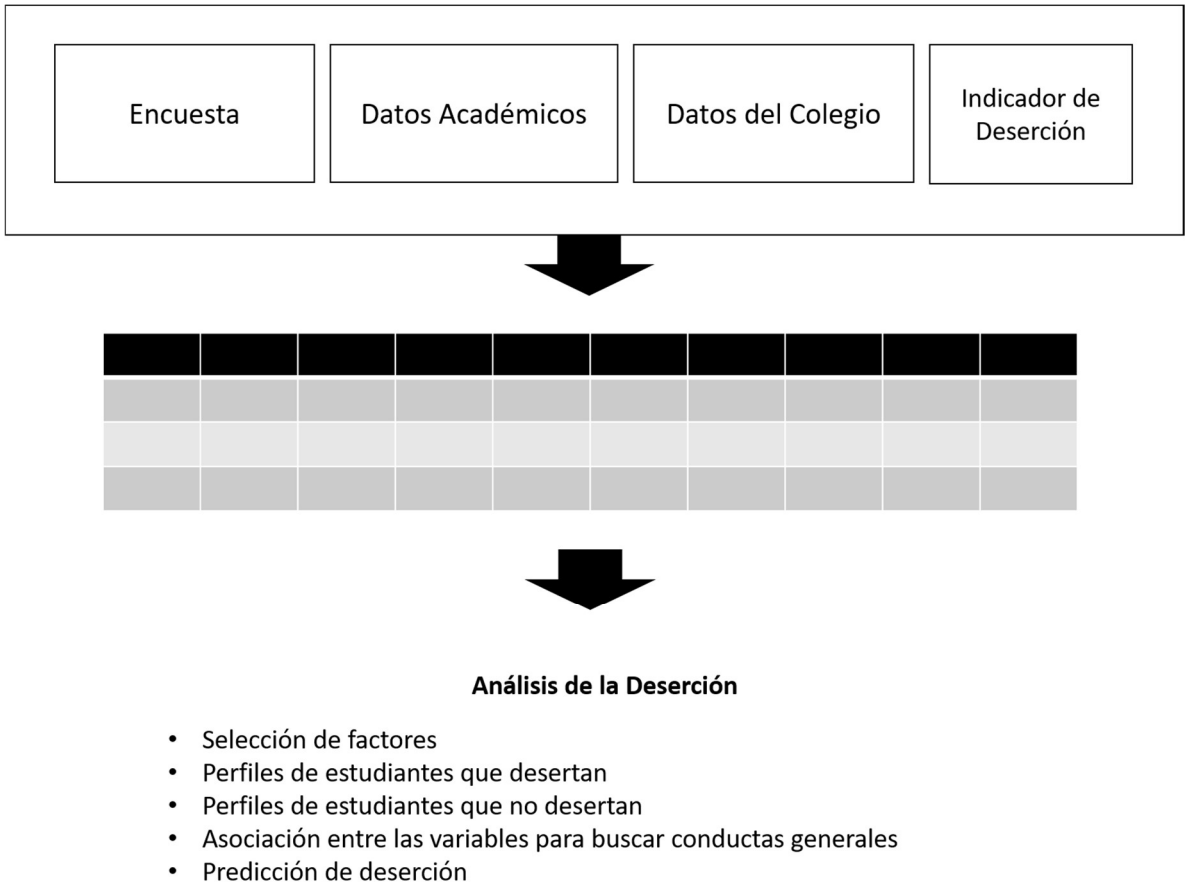
La necesidad de predecir la deserción es determinante para tomar acciones pertinentes y poder disminuir sus índices. Aunque como lo dice Tinto, conocer los predictores para una posible deserción no es suficiente; la institución debe plantear mecanismos para que se creen ambientes que hagan que los estudiantes no sólo permanezcan, sino que sobresalgan.

PARTE 2. SOLUCION Y DESARROLLO DE LA METODOLOGÍA CRISP-DM

8. DISEÑO DE LA SOLUCIÓN

Con el objetivo de desarrollar un modelo predictivo utilizando técnicas de minería de datos, que permita anticiparse a la deserción en estudiantes de pregrado de la Universidad Pontificia Bolivariana; se diseña la siguiente solución analítica, donde se tienen diferentes fuentes de datos: encuesta de perfil integral, datos académicos, datos del colegio de procedencia y un indicador de deserción.

Ilustración 4. Diseño de la solución



Mediante las fuentes de datos, se construye una sábana unificada de datos, la cual es usada para crear varios modelos analíticos que permiten analizar la deserción. En los capítulos siguientes se describe a detalle la creación de dichos modelos mediante la aplicación de la metodología CRISP-DM.

9. DESCRIPCIÓN DEL NEGOCIO

9.1 Universidad Pontificia Bolivariana

La Universidad Pontificia Bolivariana (UPB), es una institución educativa colombiana que declara en su misión la formación integral de personas. Fundada en el año 1936, con presencia nacional en programas de extensión y con atención permanente en sus seccionales de Bucaramanga, Montería, Palmira y sede central Medellín. Se visualiza como una institución de excelencia educativa donde prima el liderazgo humano, científico, empresarial y social al servicio del país. (UPB, 2018)

Actualmente la sede central Medellín se encuentra acreditada en alta calidad por parte del Ministerio de Educación Nacional (MEN), y como universidad cuenta con acreditación multicampus. Para efectos del presente trabajo, se trabajará sólo en la sede central.

Según los datos de estudiantes, se encuentra una población como la que indica la siguiente tabla:

Tabla 4. Composición de estudiantes UPB Sede Central Medellín 2010-2016

Composición del estamento estudiantil								
Nivel	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**Tendencia**
Sede Central Medellín								
Pregrado	10.830	10.983	11.451	12.251	12.316	13.194	13.297	**453.1**
Especialización	1.312	1.414	1.301	1.378	1.219	1.340	1.283	**-11.3**
Maestría	382	529	490	642	1.061	1.220	1.263	**164.1**
Doctorado	72	102	148	159	163	225	271	**30.6**
Subtotal	**12.596**	**13.028**	**13.390**	**14.430**	**14.759**	**15.979**	**16.114**	**636.6**

Fuente: Informe de autoevaluación institucional 2016 tomo tres – UPB.

Y con respecto al contexto de programas académicos, la universidad presenta una organización académica en diferentes niveles de formación:

Tabla 5. Organización académica UPB Sede Central Medellín a 2016

UPB	Escuelas	Facultades	Programas			
			Pregrado	Especialización	Maestría	Doctorado
Sede Central Medellín	8	26	46	103	50	9

Fuente: Informe de autoevaluación institucional 2016 tomo tres – UPB.

9.2 Caso de estudio: estudiantes de pregrado desertores 2015-2016

Bienestar Universitario de la Universidad Pontificia Bolivariana ejecutó, en los periodos 2015-2, 2016-1 y 2016-2, una encuesta dirigida a todos los estudiantes de pregrado en la seccional Medellín, con el ánimo de obtener una caracterización detallada de los mismos, se llamó encuesta de perfil integral. Se obtiene acceso a estos resultados y también a la información académica que se encuentra en el sistema central.

En los periodos posteriores a 2016-2 no se realizó la encuesta de perfil integral, dado que se planeó la adquisición de un software para este fin.

Tanto los resultados de la encuesta, como la información académica, serán usados en la predicción de deserción en el desarrollo de este trabajo.

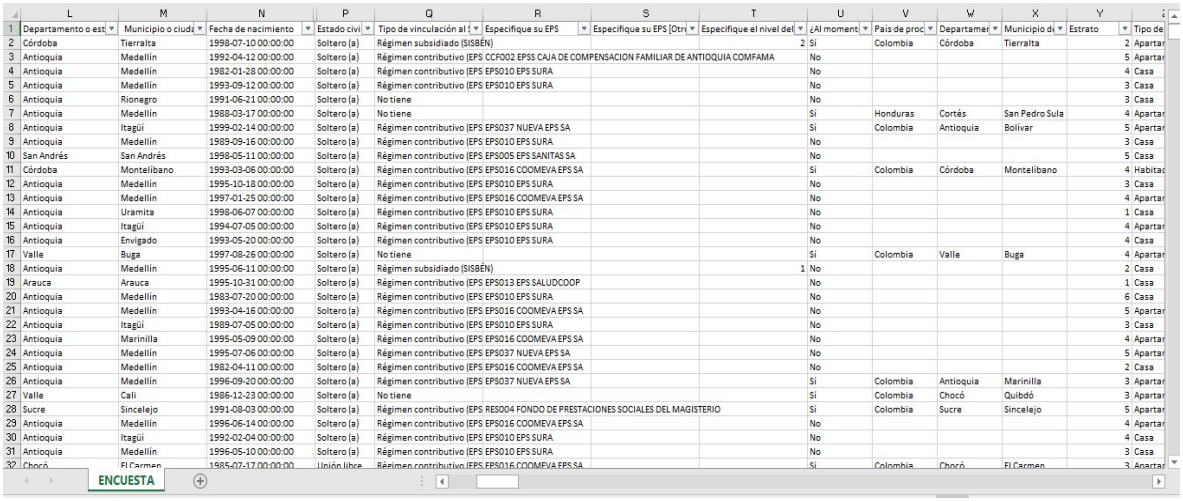
10. COMPRENSIÓN DE LOS DATOS

Como ya se había mencionado, el presente trabajo toma los datos de la encuesta de perfil integral y datos académicos que se pueden obtener para cada estudiante en un momento determinado del tiempo.

10.1 Encuesta de perfil integral

Se toma inicialmente el archivo en Excel con los datos de la encuesta de perfil integral realizada por Bienestar Universitario. En la siguiente ilustración se muestran algunas columnas de dicho archivo.

Ilustración 5. Algunas variables de la encuesta de perfil integral



	L	M	N	P	Q	R	S	T	U	V	W	X	Y
1	Departamento o est	Municipio o ciud	Fecha de nacimiento	Estado civi	Tipo de vinculación al	Especifique su EPS	Especifique su EPS [Otr	Especifique el nivel del	¿Al moment	País de proc	Departamer	Municipio d	Estrato
2	Córdoba	Tierralta	1998-07-10 00:00:00	Soltero (a)	Régimen subsidiado (SISBÉN)			2	Sí	Colombia	Córdoba	Tierralta	2
3	Antioquia	Medellín	1992-04-12 00:00:00	Soltero (a)	Régimen contributivo (EPS	CCF002 EPSS CAJA DE COMPENSACION FAMILIAR DE ANTIOQUIA COMFAMA			No				5
4	Antioquia	Medellín	1982-01-28 00:00:00	Soltero (a)	Régimen contributivo (EPS	EPS010 EPS SURA			No				4
5	Antioquia	Medellín	1993-09-12 00:00:00	Soltero (a)	Régimen contributivo (EPS	EPS010 EPS SURA			No				3
6	Antioquia	Rionegro	1991-06-21 00:00:00	Soltero (a)	No tiene				No				3
7	Antioquia	Medellín	1988-03-17 00:00:00	Soltero (a)	No tiene				Sí	Honduras	Cortés	San Pedro Sula	4
8	Antioquia	Itagüí	1999-02-14 00:00:00	Soltero (a)	Régimen contributivo (EPS	EPS037 NUEVA EPS SA			Sí	Colombia	Antioquia	Bolivar	5
9	Antioquia	Medellín	1989-09-16 00:00:00	Soltero (a)	Régimen contributivo (EPS	EPS010 EPS SURA			No				3
10	San Andrés	San Andrés	1998-05-11 00:00:00	Soltero (a)	Régimen contributivo (EPS	EPS005 EPS SANITAS SA			No				5
11	Córdoba	Montelíbano	1993-03-06 00:00:00	Soltero (a)	Régimen contributivo (EPS	EPS016 COOMEVA EPS SA			Sí	Colombia	Córdoba	Montelíbano	4
12	Antioquia	Medellín	1995-10-18 00:00:00	Soltero (a)	Régimen contributivo (EPS	EPS010 EPS SURA			No				3
13	Antioquia	Medellín	1997-01-25 00:00:00	Soltero (a)	Régimen contributivo (EPS	EPS016 COOMEVA EPS SA			No				4
14	Antioquia	Uramita	1998-06-07 00:00:00	Soltero (a)	Régimen contributivo (EPS	EPS010 EPS SURA			No				1
15	Antioquia	Itagüí	1994-07-05 00:00:00	Soltero (a)	Régimen contributivo (EPS	EPS010 EPS SURA			No				4
16	Antioquia	Envigado	1993-05-20 00:00:00	Soltero (a)	Régimen contributivo (EPS	EPS010 EPS SURA			No				4
17	Valle	Buga	1997-08-26 00:00:00	Soltero (a)	No tiene				Sí	Colombia	Valle	Buga	4
18	Antioquia	Medellín	1995-06-11 00:00:00	Soltero (a)	Régimen subsidiado (SISBÉN)			1	No				2
19	Arauca	Arauca	1995-10-31 00:00:00	Soltero (a)	Régimen contributivo (EPS	EPS013 EPS SALUDCOOP			No				1
20	Antioquia	Medellín	1983-07-20 00:00:00	Soltero (a)	Régimen contributivo (EPS	EPS010 EPS SURA			No				6
21	Antioquia	Medellín	1993-04-16 00:00:00	Soltero (a)	Régimen contributivo (EPS	EPS016 COOMEVA EPS SA			No				5
22	Antioquia	Itagüí	1989-07-05 00:00:00	Soltero (a)	Régimen contributivo (EPS	EPS010 EPS SURA			No				3
23	Antioquia	Marinilla	1995-05-09 00:00:00	Soltero (a)	Régimen contributivo (EPS	EPS016 COOMEVA EPS SA			No				4
24	Antioquia	Medellín	1995-07-06 00:00:00	Soltero (a)	Régimen contributivo (EPS	EPS037 NUEVA EPS SA			No				5
25	Antioquia	Medellín	1982-04-11 00:00:00	Soltero (a)	Régimen contributivo (EPS	EPS016 COOMEVA EPS SA			No				2
26	Antioquia	Medellín	1996-09-20 00:00:00	Soltero (a)	Régimen contributivo (EPS	EPS037 NUEVA EPS SA			Sí	Colombia	Antioquia	Marinilla	3
27	Valle	Cali	1986-12-23 00:00:00	Soltero (a)	No tiene				Sí	Colombia	Chocó	Quibdó	3
28	Sucre	Sincelejo	1991-08-03 00:00:00	Soltero (a)	Régimen contributivo (EPS	RES004 FONDO DE PRESTACIONES SOCIALES DEL MAGISTERIO			Sí	Colombia	Sucre	Sincelejo	5
29	Antioquia	Medellín	1996-06-14 00:00:00	Soltero (a)	Régimen contributivo (EPS	EPS016 COOMEVA EPS SA			No				4
30	Antioquia	Itagüí	1992-02-04 00:00:00	Soltero (a)	Régimen contributivo (EPS	EPS010 EPS SURA			No				4
31	Antioquia	Medellín	1996-05-10 00:00:00	Soltero (a)	Régimen contributivo (EPS	EPS010 EPS SURA			No				3
32	Chocó	El Carmen	1985-07-17 00:00:00	Unión libre	Régimen contributivo (EPS	EPS016 COOMEVA EPS SA			Sí	Colombia	Chocó	El Carmen	3

ENCUESTA

10.1.1 Descripción de los datos

El archivo consta de 12935 registros y 277 columnas. A continuación, se nombran las columnas:

Tabla 6. Columnas encuesta de perfil integral

Columna	Nombre Columna
1	ID de respuesta (Consecutivo asignado por el sistema de encuestas)
2	Fecha y hora de envío (Terminación de la encuesta)
3	Última página de respuesta (8 páginas para todos los casos)
4	Lenguaje inicial (Español para todos los casos)

Columna	Nombre Columna
5	Fecha y hora de inicio de la encuesta
6	Fecha de la última acción en la encuesta
7	Dirección IP desde la cual se realizó la encuesta
8	URL de referencia
9	Acepto los términos y condiciones (Sí para todos los casos)
10	ID del Estudiante en el sistema de información
11	País de nacimiento
12	Departamento o estado de nacimiento
13	Municipio o ciudad de nacimiento
14	Fecha de nacimiento
15	Estado civil
16	Tipo de vinculación al Sistema General de Seguridad Social en Salud
17	Especifique su EPS
18	Especifique su EPS [Otro] (Si no se encuentra en la lista)
19	Especifique el nivel del SISBÉN
20	¿Al momento de iniciar sus estudios en la universidad se trasladó de Ciudad o Municipio-
21	País de procedencia
22	Departamento o estado de procedencia
23	Municipio de procedencia
24	Estrato Socioeconómico
25	Tipo de residencia (Casa, Apartamento, Habitación)
26	Tenencia de la residencia (Arriendo, Propia, Familiar, Residencia)
27	Zona de la residencia (Urbana, Rural)
28	Estado actual del padre (Vivo, Fallecido, NS/NR)
29	¿Convive con el padre-
30	Si no convive con él es debido a:
31	Si no convive con él es debido a: [Otro] (Si no está en la lista)
32	Nivel educativo del padre
33	Ocupación del padre
34	Promedio de ingresos del padre (En rango de salarios mínimos)
35	Estado actual de la madre (Viva, Fallecida, NS/NR)
36	¿Convive con la madre-
37	Si no convive con ella es debido a:
38	Si no convive con ella es debido a: [Otro] (Si no está en la lista)
39	Nivel educativo de la madre
40	Ocupación de la madre
41	Promedio de ingresos de la madre (En rango de salarios mínimos)
42	¿Quién es su acudiente-
43	¿Convive con el acudiente-

Columna	Nombre Columna
44	¿Usted tiene hermanos-
45	¿Cuántos hermanos tiene-
46	¿Cuántos hermanos tiene- [Otro] (Si no está en la lista)
47	¿Cuál es el lugar que ocupa entre sus hermanos-
48	Cantidad de hermanos del estudiante que han culminado estudios de pregrado o postgrado (Técnico, Tecnológica, Universitaria, Licenciatura, Especialización, Maestría, Doctorado)
49	¿Vive con alguno de ellos-
50	¿Con cuántos hermanos vive-
51	¿Vive con otros miembros de la familia-
52	¿ Con cuántos miembros de la familia vive-
53	¿Con cuáles miembros de la familia vive- [Abuelo(a)]
54	¿Con cuáles miembros de la familia vive- [Tío(a)]
55	¿Con cuáles miembros de la familia vive- [Primo(a)]
56	¿Con cuáles miembros de la familia vive- [Suegro(a)]
57	¿Con cuáles miembros de la familia vive- [Otro]
58	¿Convive con otras personas que no sean de su familia-
59	¿Con cuántas personas que no sean de la familia convive-
60	¿Actualmente tiene personas a cargo-
61	¿Número de personas a cargo-
62	¿Usted tiene hijos-
63	Número de hijos
64	¿En su barrio se presentan algunas de estas situaciones o fenómenos- [Barreras invisibles]
65	¿En su barrio se presentan algunas de estas situaciones o fenómenos- [Conflicto armado]
66	¿En su barrio se presentan algunas de estas situaciones o fenómenos- [Extorsión]
67	¿En su barrio se presentan algunas de estas situaciones o fenómenos- [Hurto]
68	¿En su barrio se presentan algunas de estas situaciones o fenómenos- [Consumo de sustancias psicoactivas]
69	¿En su barrio se presentan algunas de estas situaciones o fenómenos- [Ninguna de las anteriores]
70	¿En su barrio se presentan algunas de estas situaciones o fenómenos- [Otro]
71	¿Ha sido víctima del conflicto armado-
72	¿Pertenece a alguna etnia o resguardo indígena-
73	Especifique el resguardo al cual pertenece
74	Especifique el resguardo al cual pertenece [Otro] (Si no está en la lista)
75	¿Quién representa la figura de autoridad en la familia-
76	¿Quién representa la figura de autoridad en la familia- [Otro] (Si no está en la lista)
77	¿Quién toma las decisiones en su familia- [Papá]

Columna	Nombre Columna
78	¿Quién toma las decisiones en su familia- [Mamá]
79	¿Quién toma las decisiones en su familia- [Hermano(a)]
80	¿Quién toma las decisiones en su familia- [Abuelo(a)]
81	¿Quién toma las decisiones en su familia- [Tío(a)]
82	¿Quién toma las decisiones en su familia- [Pareja]
83	¿Quién toma las decisiones en su familia- [Suegro(a)]
84	¿Quién toma las decisiones en su familia- [Usted]
85	¿Quién toma las decisiones en su familia- [Otro]
86	¿Le satisface el apoyo que recibe de su familia cuando tiene algún problema y/o necesidad-
87	De qué manera enfrenta su familia las crisis usualmente: [Entre los miembros de la familia se piden apoyo]
88	De qué manera enfrenta su familia las crisis usualmente: [Tratan de solucionar la situación hablando asertivamente]
89	De qué manera enfrenta su familia las crisis usualmente: [Cada uno se va por su lado evadiendo la situación]
90	De qué manera enfrenta su familia las crisis usualmente: [Buscan ayuda profesional para orientarlos]
91	De qué manera enfrenta su familia las crisis usualmente: [Todos son indiferentes]
92	De qué manera enfrenta su familia las crisis usualmente: [Todos se alteran y se empeora la situación]
93	¿En su familia han vivido alguno de los siguientes acontecimientos- [Secuestro]
94	¿En su familia han vivido alguno de los siguientes acontecimientos- [Asesinato de uno de los miembros de la familia]
95	¿En su familia han vivido alguno de los siguientes acontecimientos- [Violación]
96	¿En su familia han vivido alguno de los siguientes acontecimientos- [Extorsión]
97	¿En su familia han vivido alguno de los siguientes acontecimientos- [Suicidio]
98	¿En su familia han vivido alguno de los siguientes acontecimientos- [Desplazamiento forzado]
99	¿En su familia han vivido alguno de los siguientes acontecimientos- [Desaparición]
100	¿En su familia han vivido alguno de los siguientes acontecimientos- [Ninguno]
101	¿En su familia han vivido alguno de los siguientes acontecimientos- [Otro] (Diferente a los anteriores)
102	¿Actualmente en su familia presentan alguna de las siguientes situaciones- [Malas Relaciones intrafamiliares]
103	¿Actualmente en su familia presentan alguna de las siguientes situaciones- [Fallecimiento de algún familiar (primer grado consanguinidad)]
104	¿Actualmente en su familia presentan alguna de las siguientes situaciones- [Violencia intrafamiliar]
105	¿Actualmente en su familia presentan alguna de las siguientes situaciones- [Abuso o violencia sexual]
106	¿Actualmente en su familia presentan alguna de las siguientes situaciones- [Enfermedad crónica de algún pariente]

Columna	Nombre Columna
107	¿Actualmente en su familia presentan alguna de las siguientes situaciones-[Separación de los padres]
108	¿Actualmente en su familia presentan alguna de las siguientes situaciones-[Alcoholismo o adicción a sustancias]
109	¿Actualmente en su familia presentan alguna de las siguientes situaciones-[Desplazamiento forzado]
110	¿Actualmente en su familia presentan alguna de las siguientes situaciones-[Dificultades económicas de la familia]
111	¿Actualmente en su familia presentan alguna de las siguientes situaciones-[Ninguna de las anteriores]
112	¿Actualmente en su familia presentan alguna de las siguientes situaciones-[Otro] (Diferente a los anteriores)
113	Posee Necesidades Educativas Especiales (NEE) como: [Discapacidad Intelectual]
114	Posee Necesidades Educativas Especiales (NEE) como: [Déficit de atención con Hiperactividad (TDAH)]
115	Posee Necesidades Educativas Especiales (NEE) como: [Hipoacusia o baja audición]
116	Posee Necesidades Educativas Especiales (NEE) como: [Sordera profunda]
117	Posee Necesidades Educativas Especiales (NEE) como: [Baja Visión diagnosticada]
118	Posee Necesidades Educativas Especiales (NEE) como: [Ceguera]
119	Posee Necesidades Educativas Especiales (NEE) como: [Trastornos del lenguaje (opción de cual)]
120	Posee Necesidades Educativas Especiales (NEE) como: [Discapacidad motora o motriz (opción de cuál)]
121	Posee Necesidades Educativas Especiales (NEE) como: [Autismo]
122	Posee Necesidades Educativas Especiales (NEE) como: [Asperger]
123	Posee Necesidades Educativas Especiales (NEE) como: [Parálisis cerebral]
124	Posee Necesidades Educativas Especiales (NEE) como: [Lesión neuromuscular]
125	Posee Necesidades Educativas Especiales (NEE) como: [Síndrome de Down]
126	Posee Necesidades Educativas Especiales (NEE) como: [Ninguna]
127	Posee Necesidades Educativas Especiales (NEE) como: [Otro] (Diferente a las anteriores)
128	Padece alguna de las siguientes enfermedades: [Alergias]
129	Padece alguna de las siguientes enfermedades: [Insuficiencia Renal]
130	Padece alguna de las siguientes enfermedades: [Artritis, lupus]
131	Padece alguna de las siguientes enfermedades: [Diabetes / Hipertensión / Insuficiencia cardíaca]
132	Padece alguna de las siguientes enfermedades: [Hemofilia]
133	Padece alguna de las siguientes enfermedades: [Cáncer]
134	Padece alguna de las siguientes enfermedades: [Epilepsia]
135	Padece alguna de las siguientes enfermedades: [Deficiencia cardiaca]
136	Padece alguna de las siguientes enfermedades: [Migraña]
137	Padece alguna de las siguientes enfermedades: [Fibromialgia]

Columna	Nombre Columna
138	Padece alguna de las siguientes enfermedades: [Colon irritable / Gastritis]
139	Padece alguna de las siguientes enfermedades: [Síndrome de fatiga crónica]
140	Padece alguna de las siguientes enfermedades: [Disautonomía]
141	Padece alguna de las siguientes enfermedades: [Trastorno de ansiedad / Depresión / Trastorno Bipolar]
142	Padece alguna de las siguientes enfermedades: [Síndrome de hiperactividad]
143	Padece alguna de las siguientes enfermedades: [Ninguna de las anteriores]
144	Padece alguna de las siguientes enfermedades: [Otro] (Diferente a las anteriores)
145	¿Actualmente trabaja-
146	Cargo o actividad que desempeña
147	Empresa o lugar donde labora
148	Tipo de contrato
149	Número de horas semanales que labora
150	Promedio de ingresos mensuales o por labor realizada
151	¿Cuántas personas aportan al sostenimiento económico de la familia-
152	¿Quién es el principal proveedor económico de su familia-
153	¿Quién es el principal proveedor económico de su familia- [Otro]
154	¿Quién es el principal proveedor económico de su familia- [Otro] - (Columna repetida)
155	¿Cuánto suman los ingresos mensuales de su familia- (En rango de salarios mínimos)
156	¿Con los ingresos que actualmente la familia recibe, logran cubrir sus necesidades básicas-
157	Especifique el tipo de necesidades que se quedan sin cubrir [Servicios básicos]
158	Especifique el tipo de necesidades que se quedan sin cubrir [Transporte]
159	Especifique el tipo de necesidades que se quedan sin cubrir [Alimentación]
160	Especifique el tipo de necesidades que se quedan sin cubrir [Ocio]
161	Especifique el tipo de necesidades que se quedan sin cubrir [Otro]
162	Forma de pago de la matrícula [De contado (Efectivo)]
163	Forma de pago de la matrícula [Entidad Bancaria]
164	Forma de pago de la matrícula [ICETEX]
165	Forma de pago de la matrícula [Beneficiario de beca]
166	Forma de pago de la matrícula [Fondo EPM]
167	Forma de pago de la matrícula [Tarjeta de crédito]
168	Forma de pago de la matrícula [Otro] (Diferente a las anteriores)
169	Especifique el tipo de beca
170	¿Usted considera que requiere ayuda financiera para el pago de la matrícula-
171	¿Recibe algún tipo de apoyo por parte de la Universidad-
172	Especifique el tipo de apoyo que recibe
173	Ha recibido alguno o todos de los siguientes apoyos: [Alimentación]
174	Ha recibido alguno o todos de los siguientes apoyos: [Fotocopias]

Columna	Nombre Columna
175	Ha recibido alguno o todos de los siguientes apoyos: [Transporte]
176	Ha recibido alguno o todos de los siguientes apoyos: [Ninguno]
177	Considera usted que requiere de otros apoyos como: [Alimentación]
178	Considera usted que requiere de otros apoyos como: [Fotocopias]
179	Considera usted que requiere de otros apoyos como: [Transporte]
180	Considera usted que requiere de otros apoyos como: [Materiales para la elaboración de sus trabajos]
181	Considera usted que requiere de otros apoyos como: [Beca]
182	Considera usted que requiere de otros apoyos como: [Crédito]
183	Considera usted que requiere de otros apoyos como: [Ninguno]
184	Considera usted que requiere de otros apoyos como: [Otro]
185	¿Durante su educación media presentó alguna de las siguientes situaciones- [Dificultad en adaptación social]
186	¿Durante su educación media presentó alguna de las siguientes situaciones- [Dificultades de salud]
187	¿Durante su educación media presentó alguna de las siguientes situaciones- [Pérdida de años]
188	¿Durante su educación media presentó alguna de las siguientes situaciones- [Dificultades académicas]
189	¿Durante su educación media presentó alguna de las siguientes situaciones- [Ninguno]
190	¿Durante su educación media presentó alguna de las siguientes situaciones- [Otro]
191	¿Cuál fue el principal motivo que lo llevó a elegir la Universidad Pontificia Bolivariana- [Calidad académica]
192	¿Cuál fue el principal motivo que lo llevó a elegir la Universidad Pontificia Bolivariana- [Reconocimiento en el medio]
193	¿Cuál fue el principal motivo que lo llevó a elegir la Universidad Pontificia Bolivariana- [Formación integral]
194	¿Cuál fue el principal motivo que lo llevó a elegir la Universidad Pontificia Bolivariana- [Por decisión de sus padres]
195	¿Cuál fue el principal motivo que lo llevó a elegir la Universidad Pontificia Bolivariana- [No pasó a otra institución de educación superior]
196	¿Cuál fue el principal motivo que lo llevó a elegir la Universidad Pontificia Bolivariana- [Otro] (Diferente a los anteriores)
197	Motivo por el cual eligió la carrera [Sugerencia recibida por orientación profesional]
198	Motivo por el cual eligió la carrera [Interés propio]
199	Motivo por el cual eligió la carrera [Reconocimiento en el medio]
200	Motivo por el cual eligió la carrera [Presión familiar]
201	Motivo por el cual eligió la carrera [Presión social]
202	Motivo por el cual eligió la carrera [No sabía qué estudiar]
203	Motivo por el cual eligió la carrera [Interés económico]
204	Motivo por el cual eligió la carrera [Otro]
205	¿Realizó orientación vocacional antes de ingresar a la universidad-

Columna	Nombre Columna
206	¿Considera que tiene dificultad en alguno de los siguientes aspectos- [Trabajar contenidos matemáticos y numéricos]
207	¿Considera que tiene dificultad en alguno de los siguientes aspectos- [Concentrarse y prestar atención]
208	¿Considera que tiene dificultad en alguno de los siguientes aspectos- [Comprensión lectora]
209	¿Considera que tiene dificultad en alguno de los siguientes aspectos- [Hablar en público]
210	¿Considera que tiene dificultad en alguno de los siguientes aspectos- [Las actividades artísticas y manuales]
211	¿Considera que tiene dificultad en alguno de los siguientes aspectos- [Memorizar]
212	¿Considera que tiene dificultad en alguno de los siguientes aspectos- [Ver bien o escuchar bien en el aula]
213	¿Considera que tiene dificultad en alguno de los siguientes aspectos- [Ninguno]
214	¿Considera que tiene dificultad en alguno de los siguientes aspectos- [Otro]
215	¿Con cuáles de los siguientes recursos NO cuenta para el estudio en casa- [No cuenta con computador para estudiar y realizar sus trabajos académicos]
216	¿Con cuáles de los siguientes recursos NO cuenta para el estudio en casa- [No cuenta con acceso a internet como herramienta de consulta]
217	¿Con cuáles de los siguientes recursos NO cuenta para el estudio en casa- [No cuenta con libros para estudiar]
218	¿Con cuáles de los siguientes recursos NO cuenta para el estudio en casa- [No tiene espacio para estudiar]
219	¿Con cuáles de los siguientes recursos NO cuenta para el estudio en casa- [Ninguno]
220	¿Con cuáles de los siguientes recursos NO cuenta para el estudio en casa- [Otro]
221	Califique el espacio que tiene en su casa para estudiar
222	¿Cuáles de los siguientes métodos de estudio utiliza usualmente- [Subrayar]
223	¿Cuáles de los siguientes métodos de estudio utiliza usualmente- [Sacar ideas principales]
224	¿Cuáles de los siguientes métodos de estudio utiliza usualmente- [Apuntes]
225	¿Cuáles de los siguientes métodos de estudio utiliza usualmente- [Mapas conceptuales, resumen, lluvia de ideas, etc.]
226	¿Cuáles de los siguientes métodos de estudio utiliza usualmente- [Memorización mecánica (sin comprender lo que estudio)]
227	¿Cuáles de los siguientes métodos de estudio utiliza usualmente- [Repasar después de cada clase]
228	¿Cuáles de los siguientes métodos de estudio utiliza usualmente- [Estudiar minutos antes del examen o exposición]
229	¿Cuáles de los siguientes métodos de estudio utiliza usualmente- [Otro]
230	Cómo se siente con los métodos de estudio que utiliza usualmente:
231	Participa en grupos o realiza actividades relacionadas con: [Arte]
232	Participa en grupos o realiza actividades relacionadas con: [Deporte]
233	Participa en grupos o realiza actividades relacionadas con: [Formación personal]
234	Participa en grupos o realiza actividades relacionadas con: [Espiritual]

Columna	Nombre Columna
235	Participa en grupos o realiza actividades relacionadas con: [Sociales]
236	Participa en grupos o realiza actividades relacionadas con: [Proyección comunitaria]
237	Participa en grupos o realiza actividades relacionadas con: [Emprendimiento]
238	Participa en grupos o realiza actividades relacionadas con: [Ninguno]
239	Participa en grupos o realiza actividades relacionadas con: [Otro]
240	¿Qué tipo de actividad realiza con mayor frecuencia en su tiempo libre- [Ver T.V]
241	¿Qué tipo de actividad realiza con mayor frecuencia en su tiempo libre- [Escuchar música]
242	¿Qué tipo de actividad realiza con mayor frecuencia en su tiempo libre- [Dormir]
243	¿Qué tipo de actividad realiza con mayor frecuencia en su tiempo libre- [Jugar video Juegos]
244	¿Qué tipo de actividad realiza con mayor frecuencia en su tiempo libre- [Deporte]
245	¿Qué tipo de actividad realiza con mayor frecuencia en su tiempo libre- [Compartir con los amigos]
246	¿Qué tipo de actividad realiza con mayor frecuencia en su tiempo libre- [Navegar en internet]
247	¿Qué tipo de actividad realiza con mayor frecuencia en su tiempo libre- [Leer]
248	¿Qué tipo de actividad realiza con mayor frecuencia en su tiempo libre- [Ninguna]
249	¿Qué tipo de actividad realiza con mayor frecuencia en su tiempo libre- [Otro]
250	¿Realiza actividad física-
251	Frecuencia por semana
252	Especifique el tipo de actividad física realiza [Actividades al aire libre]
253	Especifique el tipo de actividad física realiza [Gimnasio]
254	Especifique el tipo de actividad física realiza [Fútbol]
255	Especifique el tipo de actividad física realiza [Natación]
256	Especifique el tipo de actividad física realiza [Artes Marciales]
257	Especifique el tipo de actividad física realiza [Baloncesto.]
258	Especifique el tipo de actividad física realiza [Voleibol]
259	Especifique el tipo de actividad física realiza [Otro]
260	¿Realiza alguna actividad cultural o artística-
261	Especifique el tipo de actividad cultural o artística que realiza [Teatro]
262	Especifique el tipo de actividad cultural o artística que realiza [Baile]
263	Especifique el tipo de actividad cultural o artística que realiza [Música (Instrumento o canto)]
264	Especifique el tipo de actividad cultural o artística que realiza [Cuentería, expresión oral o corporal]
265	Especifique el tipo de actividad cultural o artística que realiza [Clases de artes plásticas (Manualidades, dibujo, pintura, escultura)]
266	Especifique el tipo de actividad cultural o artística que realiza [Ninguna]
267	Especifique el tipo de actividad cultural o artística que realiza [Otro] (Diferente a las anteriores)
268	¿Qué tipo de espectáculos o actividades disfruta- [Conciertos]

Columna	Nombre Columna
269	¿Qué tipo de espectáculos o actividades disfruta- [Shows de baile]
270	¿Qué tipo de espectáculos o actividades disfruta- [Cuentería, expresión oral o corporal]
271	¿Qué tipo de espectáculos o actividades disfruta- [Demostraciones de arte]
272	¿Qué tipo de espectáculos o actividades disfruta- [Cine]
273	¿Qué tipo de espectáculos o actividades disfruta- [Ninguno]
274	¿Qué tipo de espectáculos o actividades disfruta- [Otro] (Diferente a los anteriores)
275	Tiempo total para la realización de la encuesta (Segundos)
276	CODIGO
277	ARCHIVO (1,2)

10.1.2 Cálculo de nuevas variables

Tomando los datos de la encuesta de perfil integral, se agregan las siguientes dos columnas.

Tabla 7. Nuevas variables desde la encuesta de perfil integral

Columna	Descripción
Edad al momento de la encuesta	Se toma la fecha de nacimiento y se calcula la edad al momento de la encuesta.
Periodo	Con la fecha de terminación de la encuesta se calcula si corresponde al semestre 2015-2, 2016-1 o 2016-2.

10.2 Datos académicos

Previo a la obtención de los datos académicos, se realizó reunión con representantes del Programa de Permanencia de la Sede Central, donde se concluyó que antes de predecir la deserción, es importante conocer qué estudiantes están rezagados, dado que es muy probable que esta condición esté estrechamente relacionada con la probabilidad de deserción. El criterio dado para definir rezago es un estudiante que tenga dos o más semestres de atraso según su semestre de inicio. Por ejemplo, si un estudiante inició un programa hace cinco semestres, pero

al día de hoy su ubicación es segundo semestre, significa que estaría rezagado en tres semestres.

10.2.1 Descripción de los datos

Los siguientes datos académicos se obtienen del sistema de información central (Banner) para los periodos académicos 2015-2, 2016-1 y 2016-2; correspondientes a todos los estudiantes en programas de pregrado, que tengan asignados una ubicación en un semestre académico válido y no sean estudiantes de doble programa. Lo anterior se complementa con la información obtenida del sitio FTP del ICFES (ftp.icfes.gov.co), donde se encuentra la clasificación de los colegios de acuerdo a los resultados de las pruebas Saber11.

Tabla 8. Columnas datos académicos

Columna	Descripción
ID	Identificación del estudiante en el sistema de información
Periodo	Periodo académico. Incluye el año y semestre correspondiente para el cual se presentan los datos siguientes.
Programa	Programa académico de pregrado.
Créditos Programa	Créditos totales del programa académico.
Semestres Programa	Número de semestres que tiene el programa académico.
Edad inicio programa	Edad del estudiante al momento de iniciar el programa académico.
Ubicación	Ubicación semestral de acuerdo al sistema de información para el periodo académico.
Estado académico	Estado académico según el sistema de información en el periodo académico.
Código del Colegio	Código del colegio de procedencia según el Ministerio de Educación de Colombia.
Colegio	Nombre del colegio de procedencia
Naturaleza Colegio	Si el colegio de procedencia es Oficial o No Oficial
Género Colegio	Tipo de población que maneja el colegio de procedencia. Femenino, Masculino, Mixto.
Calendario Colegio	Calendario del colegio. Puede ser A, B, F
Categoría Colegio	Categoría que se le da al colegio según resultados de pruebas Saber11. Puede ser desde Muy Superior hasta Muy Inferior.
Créditos inscritos	También llamados créditos intentados. Incluye total créditos cancelados, aprobados y reprobados de todos los cursos que haya matriculado hasta el periodo académico

Columna	Descripción
Créditos aprobados	Total créditos aprobados hasta el periodo académico.
Año inicio	Año académico de ingreso al programa académico de pregrado
Semestre inicio	Semestre correspondiente de ingreso al programa (1 ó 2)
Promedio notas	Promedio académico general que incluye las notas de los cursos aprobados y reprobados hasta el periodo académico
Periodo de corte	Dos semestres después del periodo académico. Con este periodo de corte se calculan las variables involucradas en la deserción.
Año última matrícula	Último año de matrícula de cursos al periodo de corte.
Semestre última matrícula	Semestre correspondiente a la última matrícula de cursos (1 ó 2) al periodo de corte.
Créditos aprobados totales	Total de créditos aprobados desde el inicio del programa hasta el periodo de corte.
Periodo grado	Periodo en el cual se le haya otorgado grado para el programa académico hasta el periodo de corte.

10.2.2 Cálculo de nuevas variables

Tomando los datos académicos descritos se adicionan las siguientes columnas.

Tabla 9. Nuevas variables desde datos académicos

Columna	Descripción
% Aprobados / Intentados	Relación entre el número de créditos aprobados y el número de créditos inscritos al periodo académico.
% avance programa	Relación entre el número de créditos aprobados totales al periodo de corte y el número de créditos del programa.
# semestres desde inicio programa	Número de semestres transcurridos desde el inicio del programa académico hasta el periodo académico.
Semestres atraso	Diferencia entre el número de semestres transcurridos desde el inicio del programa y la ubicación semestral dada por el sistema de información.
# semestres desde última matrícula	Número de semestres transcurridos desde la última matrícula de cursos y el periodo de corte.
Rezagado	Toma el valor SI cuando para los estudiantes que llevan como mínimos dos semestres de atraso entre la ubicación del sistema y el número de semestres desde el inicio del programa académico.
Desertor	Toma el valor SI cuando un estudiante lleve más de dos semestres sin matricularse al periodo de corte, que no haya terminado su pensum

Columna	Descripción
	académico (porcentaje de avance del programa menor al 100%) y que no se haya graduado. Esta se convierte en la variable objetivo.

10.3 Resumen de datos

Ahora se muestra de una manera concisa los datos a analizar y algunos criterios básicos.

Tabla 10. Resumen de datos a analizar

Encuesta de perfil integral	Datos académicos	Datos del colegio de procedencia	Deserción
Realizada en los periodos: 2015-2, 2016-1 y 2016-2	Tomados del mismo periodo de tiempo donde se aplicó la encuesta	Complementan el set de datos académicos al mismo periodo de tiempo.	Se calcula como no matricularse durante 2 semestres consecutivos posteriores a la realización de la encuesta

11. PREPARACIÓN DE LOS DATOS

11.1 Integración de los datos

Una vez se tienen las dos fuentes anteriores con todas sus columnas, se realiza la integración de las mismas por los campos ID del estudiante y periodo. Con esto se obtiene una base de datos que conserva la integralidad de la información, donde cada registro corresponde al resultado de la encuesta de perfil integral complementado con datos académicos.

Lo anterior quiere decir que a las 277 columnas de la encuesta de perfil integral se le agregan las siguientes 28 columnas:

Tabla 11. Columnas integradas a la encuesta de perfil integral

Columna
Programa
Créditos del Programa
Ubicación Semestral
Código del Colegio
Colegio
Naturaleza Colegio
Género Colegio
Calendario Colegio
Categoría Colegio
Créditos inscritos
Créditos aprobados
Año inicio
Semestre de Inicio
Promedio notas
Periodo de corte

Columna
Año última matrícula
Semestre última matrícula
Créditos aprobados totales
Periodo grado
Estado académico
Edad al iniciar el programa
% Aprobados / Intentados
% Avance programa
Semestres desde inicio programa
Semestres atraso
Semestres desde última matrícula
Rezagado
Desertor (variable objetivo)

11.2 Variables irrelevantes

Las siguientes variables se eliminan dado que corresponden a datos de control de la encuesta, las cuales no influyen de ninguna manera en los resultados de la predicción.

Tabla 12. Columnas irrelevantes - datos de control

Columna
ID de respuesta (Consecutivo asignado por el sistema de encuestas)
Fecha y hora de envío (Terminación de la encuesta)
Última página de respuesta (8 páginas para todos los casos)
Lenguaje inicial (Español para todos los casos)
Fecha y hora de inicio de la encuesta
Fecha de la última acción en la encuesta
Dirección IP desde la cual se realizó la encuesta

Columna
URL de referencia
Acepto los términos y condiciones (Sí para todos los casos)
Tiempo total para la realización de la encuesta (Segundos)
CODIGO
ARCHIVO (1,2)

Las siguientes variables corresponden a textos abiertos y por lo tanto no se pueden categorizar fácilmente, se decide eliminarlas.

Tabla 13. Columnas irrelevantes - textos abiertos

Columna
Especifique su EPS [Otro]
Si no convive con él (padre) es debido a: [Otro]
Si no convive con ella (madre) es debido a: [Otro]
Cuántos hermanos tiene- [Otro]
Con cuáles miembros de la familia vive- [Otro]
En su barrio se presentan algunas de estas situaciones o fenómenos- [Otro]
Especifique el resguardo al cual pertenece [Otro]
Quién representa la figura de autoridad en la familia- [Otro]
Quién toma las decisiones en su familia- [Otro]
En su familia han vivido alguno de los siguientes acontecimientos- [Otro]
Actualmente en su familia presentan alguna de las siguientes situaciones- [Otro]
Posee Necesidades Educativas Especiales (NEE) como: [Otro]
Padece alguna de las siguientes enfermedades: [Otro]
Cargo o actividad que desempeña
Empresa o lugar donde labora
Promedio de ingresos mensuales o por labor realizada
Número de horas semanales que labora
Cuántas personas aportan al sostenimiento económico de la familia-

Columna
Quién es el principal proveedor económico de su familia- [Otro]
Especifique el tipo de necesidades que se quedan sin cubrir [Otro]
Forma de pago de la matrícula [Otro]
Especifique el tipo de beca
Especifique el tipo de apoyo que recibe
Considera usted que requiere de otros apoyos como: [Otro]
Durante su educación media presentó alguna de las siguientes situaciones- [Otro]
Cuál fue el principal motivo que lo llevó a elegir la Universidad Pontificia Bolivariana- [Otro]
Motivo por el cual eligió la carrera [Otro]
Considera que tiene dificultad en alguno de los siguientes aspectos- [Otro]
Con cuáles de los siguientes recursos NO cuenta para el estudio en casa- [Otro]
Cuáles de los siguientes métodos de estudio utiliza usualmente- [Otro]
Participa en grupos o realiza actividades relacionadas con: [Otro]
Qué tipo de actividad realiza con mayor frecuencia en su tiempo libre- [Otro]
Especifique el tipo de actividad física realiza [Otro]
Especifique el tipo de actividad cultural o artística que realiza [Otro]
Qué tipo de espectáculos o actividades disfruta- [Otro]

Las siguientes variables sólo tienen contenido dependiendo del resultado de otras preguntas de la encuesta, se decide eliminarlas.

Tabla 14. Columnas irrelevantes - contenido dependiente de otras columnas

Columna	Dependiente de
Especifique su EPS	Tipo de vinculación al Sistema General de Seguridad Social en Salud
Especifique el nivel del SISBÉN	Tipo de vinculación al Sistema General de Seguridad Social en Salud
País de procedencia	Al momento de iniciar sus estudios en la universidad se trasladó de Ciudad o Municipio-

Columna	Dependiente de
Departamento o estado de procedencia	Al momento de iniciar sus estudios en la universidad se trasladó de Ciudad o Municipio-
Municipio de procedencia	Al momento de iniciar sus estudios en la universidad se trasladó de Ciudad o Municipio-
Si no convive con él es debido a:	Convive con el padre-
Si no convive con ella es debido a:	Convive con la madre-
Con cuántos hermanos vive-	Vive con alguno de ellos-
Con cuántos miembros de la familia vive-	Vive con otros miembros de la familia-
Con cuántas personas que no sean de la familia convive-	Convive con otras personas que no sean de su familia-
Número de personas a cargo-	Actualmente tiene personas a cargo-
Número de hijos	Usted tiene hijos-
Especifique el resguardo al cual pertenece	Pertenece a alguna etnia o resguardo indígena-
Tipo de contrato	Actualmente trabaja-
Especifique el tipo de necesidades que se quedan sin cubrir [Servicios básicos]	Con los ingresos que actualmente la familia recibe, logran cubrir sus necesidades básicas-
Especifique el tipo de necesidades que se quedan sin cubrir [Transporte]	Con los ingresos que actualmente la familia recibe, logran cubrir sus necesidades básicas-
Especifique el tipo de necesidades que se quedan sin cubrir [Alimentación]	Con los ingresos que actualmente la familia recibe, logran cubrir sus necesidades básicas-
Especifique el tipo de necesidades que se quedan sin cubrir [Ocio]	Con los ingresos que actualmente la familia recibe, logran cubrir sus necesidades básicas-
Frecuencia por semana	Realiza actividad física?
Especifique el tipo de actividad física realiza [Actividades al aire libre]	Realiza actividad física?
Especifique el tipo de actividad física realiza [Gimnasio]	Realiza actividad física?
Especifique el tipo de actividad física realiza [Fútbol]	Realiza actividad física?
Especifique el tipo de actividad física realiza [Natación]	Realiza actividad física?
Especifique el tipo de actividad física realiza [Artes Marciales]	Realiza actividad física?
Especifique el tipo de actividad física realiza [Baloncesto]	Realiza actividad física?
Especifique el tipo de actividad física realiza [Voleibol]	Realiza actividad física?
Especifique el tipo de actividad cultural o artística que realiza [Teatro]	Realiza alguna actividad cultural o artística?

Columna	Dependiente de
Especifique el tipo de actividad cultural o artística que realiza [Baile]	Realiza alguna actividad cultural o artística?
Especifique el tipo de actividad cultural o artística que realiza [Música (Instrumento o canto)]	Realiza alguna actividad cultural o artística?
Especifique el tipo de actividad cultural o artística que realiza [Cuenteria, expresión oral o corporal]	Realiza alguna actividad cultural o artística?
Especifique el tipo de actividad cultural o artística que realiza [Clases de artes plásticas (Manualidades, dibujo, pintura, escultura)]	Realiza alguna actividad cultural o artística?
Con cuáles miembros de la familia vive-[Abuelo(a)]	Vive con otros miembros de la familia?
Con cuáles miembros de la familia vive-[Tío(a)]	Vive con otros miembros de la familia?
Con cuáles miembros de la familia vive-[Primo(a)]	Vive con otros miembros de la familia?
Con cuáles miembros de la familia vive-[Suegro(a)]	Vive con otros miembros de la familia?
Cantidad de hermanos del estudiante que han culminado estudios de pregrado o postgrado (Técnico, Tecnológica, Universitaria, Licenciatura, Especialización, Maestría, Doctorado)	Cuántos hermanos tiene?
Convive con el acudiente?	Quién es su acudiente?

Las siguientes variables se utilizaron para realizar la integración de los datos y por lo tanto no son relevantes, se decide eliminarlas.

Tabla 15. Columnas irrelevantes - integración de datos

Columna	Justificación
Periodo Académico	Se usó para realizar la integración de los datos de la encuesta con los datos académicos.
Código Colegio	Se usó para obtener datos atributos del colegio.
Nombre del colegio de procedencia	No es relevante
Créditos inscritos	Se usó para calcular % Aprobados / Intentados
Créditos aprobados	Se usó para calcular % Aprobados / Intentados

Columna	Justificación
Año inicio del programa	Permitió calcular el número de semestres desde inicio programa
Semestre inicio	Permitió calcular el número de semestres desde inicio programa
Periodo de corte	Usada en el cálculo de la variable objetivo que es la Deserción
Año última matrícula	Permitió calcular el número de semestres desde última matrícula
Semestre última matrícula	Permitió calcular el número de semestres desde última matrícula
Créditos aprobados totales	Permitió calcular el porcentaje avance en el programa académico
Periodo grado	Se usó para determinar si un estudiante se había graduado o no
% avance programa	Usada en el cálculo de la variable objetivo que es la Deserción
# Semestres desde inicio programa	Usada para calcular la variable rezagado
Semestres atraso	Usada para calcular la variable rezagado
# Semestres desde última matrícula	Usada en el cálculo de la variable objetivo que es la Deserción

11.3 Variables redundantes

Las dos variables siguientes no tienen relevancia dado que otras tienen contenido redundante, se decide eliminarlas.

Tabla 16. Variables redundantes

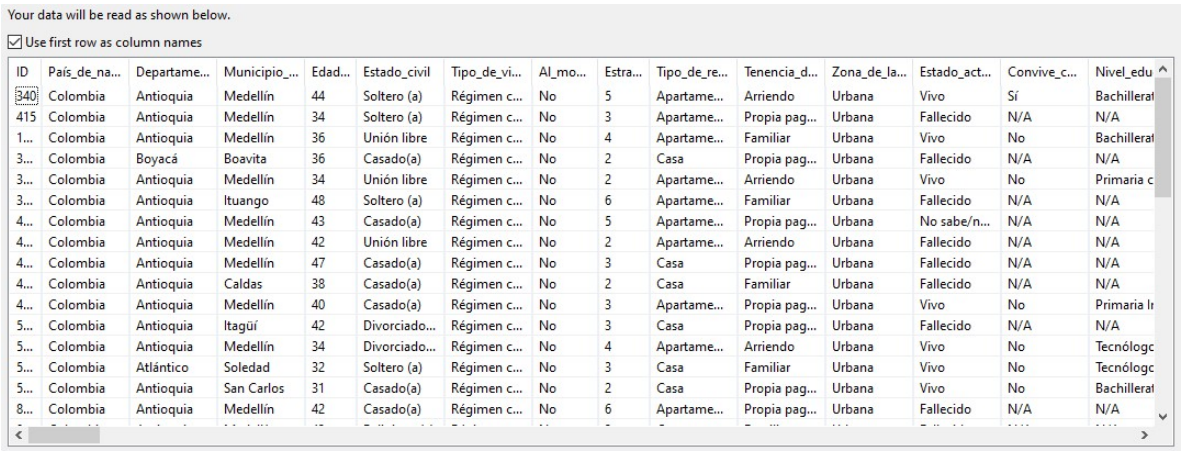
Columna	Redundante con
Fecha de Nacimiento	Edad
Usted tiene hermanos?	Cuántos hermanos tiene?

Después de eliminadas las anteriores variables se cuenta con un set de datos de 204 columnas y 12935 registros.

11.4 Descripción estadística de los datos

Se utiliza la herramienta de perfilamiento de datos llamada DQ-Analyzer para tener un panorama general de cómo está la calidad de los datos; en cuanto a duplicidad, consistencia y completitud.

Ilustración 6. Uso de DQ Analyzer para descripción estadística de los datos



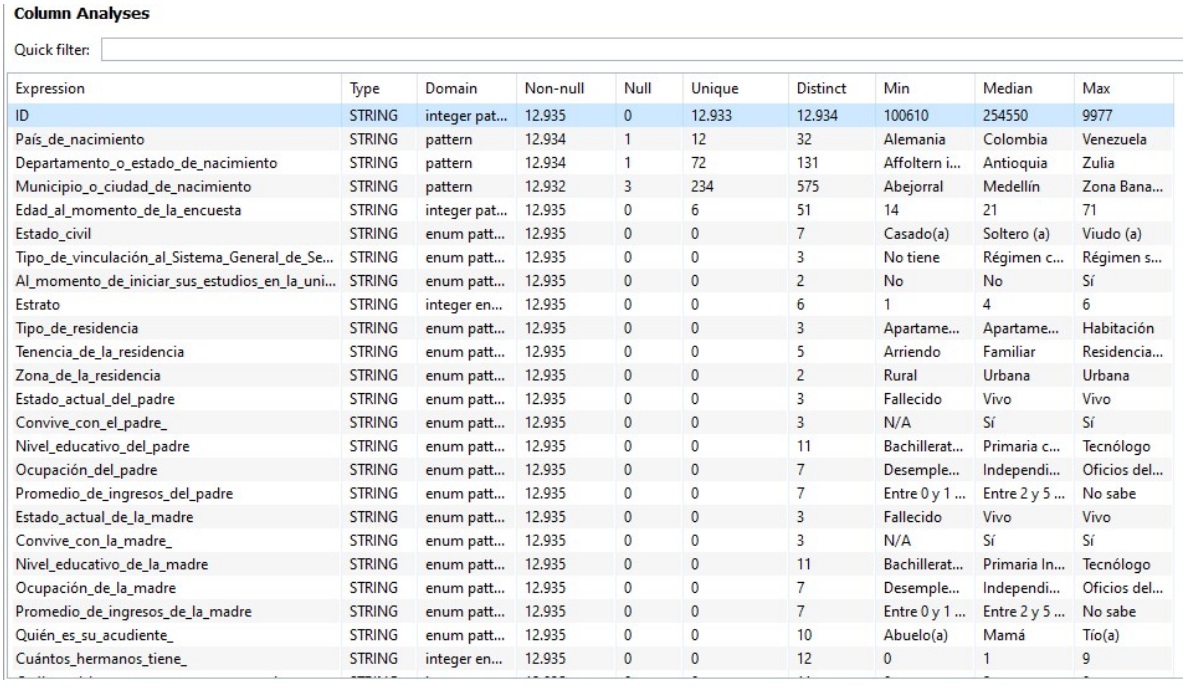
Your data will be read as shown below.

Use first row as column names

ID	País_de_na...	Departame...	Municipio_...	Edad...	Estado_civil	Tipo_de_vi...	Al_mo...	Estra...	Tipo_de_re...	Tenencia_d...	Zona_de_la...	Estado_act...	Convive_c...	Nivel_edu
340	Colombia	Antioquia	Medellín	44	Soltero (a)	Régimen c...	No	5	Apartame...	Arriendo	Urbana	Vivo	Sí	Bachillerat
415	Colombia	Antioquia	Medellín	34	Soltero (a)	Régimen c...	No	3	Apartame...	Propia pag...	Urbana	Fallecido	N/A	N/A
1...	Colombia	Antioquia	Medellín	36	Unión libre	Régimen c...	No	4	Apartame...	Familiar	Urbana	Vivo	No	Bachillerat
3...	Colombia	Boyacá	Boavita	36	Casado(a)	Régimen c...	No	2	Casa	Propia pag...	Urbana	Fallecido	N/A	N/A
3...	Colombia	Antioquia	Medellín	34	Unión libre	Régimen c...	No	2	Apartame...	Arriendo	Urbana	Vivo	No	Primaria c
3...	Colombia	Antioquia	Ituango	48	Soltero (a)	Régimen c...	No	6	Apartame...	Familiar	Urbana	Fallecido	N/A	N/A
4...	Colombia	Antioquia	Medellín	43	Casado(a)	Régimen c...	No	5	Apartame...	Propia pag...	Urbana	No sabe/n...	N/A	N/A
4...	Colombia	Antioquia	Medellín	42	Unión libre	Régimen c...	No	2	Apartame...	Arriendo	Urbana	Fallecido	N/A	N/A
4...	Colombia	Antioquia	Medellín	47	Casado(a)	Régimen c...	No	3	Casa	Propia pag...	Urbana	Fallecido	N/A	N/A
4...	Colombia	Antioquia	Caldas	38	Casado(a)	Régimen c...	No	2	Casa	Familiar	Urbana	Fallecido	N/A	N/A
4...	Colombia	Antioquia	Medellín	40	Casado(a)	Régimen c...	No	3	Apartame...	Propia pag...	Urbana	Vivo	No	Primaria In
5...	Colombia	Antioquia	Itagüí	42	Divorciado...	Régimen c...	No	3	Casa	Propia pag...	Urbana	Fallecido	N/A	N/A
5...	Colombia	Antioquia	Medellín	34	Divorciado...	Régimen c...	No	4	Apartame...	Arriendo	Urbana	Vivo	No	Tecnólogc
5...	Colombia	Atlántico	Soledad	32	Soltero (a)	Régimen c...	No	3	Casa	Familiar	Urbana	Vivo	No	Tecnólogc
5...	Colombia	Antioquia	San Carlos	31	Casado(a)	Régimen c...	No	2	Casa	Propia pag...	Urbana	Vivo	No	Bachillerat
8...	Colombia	Antioquia	Medellín	42	Casado(a)	Régimen c...	No	6	Apartame...	Propia pag...	Urbana	Fallecido	N/A	N/A

Se obtiene una vista preliminar de los resultados de la siguiente forma:

Ilustración 7. Vista preliminar resultados DQ Analyzer



Column Analyses

Quick filter:

Expression	Type	Domain	Non-null	Null	Unique	Distinct	Min	Median	Max
ID	STRING	integer pat...	12.935	0	12.933	12.934	100610	254550	9977
País_de_nacimiento	STRING	pattern	12.934	1	12	32	Alemania	Colombia	Venezuela
Departamento_o_estado_de_nacimiento	STRING	pattern	12.934	1	72	131	Affoltern i...	Antioquia	Zulia
Municipio_o_ciudad_de_nacimiento	STRING	pattern	12.932	3	234	575	Abejorral	Medellín	Zona Bana...
Edad_al_momento_de_la_encuesta	STRING	integer pat...	12.935	0	6	51	14	21	71
Estado_civil	STRING	enum patt...	12.935	0	0	7	Casado(a)	Soltero (a)	Viudo (a)
Tipo_de_vinculación_al_Sistema_General_de_Se...	STRING	enum patt...	12.935	0	0	3	No tiene	Régimen c...	Régimen s...
Al_momento_de_iniciar_sus_estudios_en_la_uni...	STRING	enum patt...	12.935	0	0	2	No	No	Sí
Estrato	STRING	integer en...	12.935	0	0	6	1	4	6
Tipo_de_residencia	STRING	enum patt...	12.935	0	0	3	Apartame...	Apartame...	Habitación
Tenencia_de_la_residencia	STRING	enum patt...	12.935	0	0	5	Arriendo	Familiar	Residencia...
Zona_de_la_residencia	STRING	enum patt...	12.935	0	0	2	Rural	Urbana	Urbana
Estado_actual_del_padre	STRING	enum patt...	12.935	0	0	3	Fallecido	Vivo	Vivo
Convive_con_el_padre_	STRING	enum patt...	12.935	0	0	3	N/A	Sí	Sí
Nivel_educativo_del_padre	STRING	enum patt...	12.935	0	0	11	Bachillerat...	Primaria c...	Tecnólogo
Ocupación_del_padre	STRING	enum patt...	12.935	0	0	7	Desemple...	Independi...	Oficios del...
Promedio_de_ingresos_del_padre	STRING	enum patt...	12.935	0	0	7	Entre 0 y 1 ...	Entre 2 y 5 ...	No sabe
Estado_actual_de_la_madre	STRING	enum patt...	12.935	0	0	3	Fallecido	Vivo	Vivo
Convive_con_la_madre_	STRING	enum patt...	12.935	0	0	3	N/A	Sí	Sí
Nivel_educativo_de_la_madre	STRING	enum patt...	12.935	0	0	11	Bachillerat...	Primaria In...	Tecnólogo
Ocupación_de_la_madre	STRING	enum patt...	12.935	0	0	7	Desemple...	Independi...	Oficios del...
Promedio_de_ingresos_de_la_madre	STRING	enum patt...	12.935	0	0	7	Entre 0 y 1 ...	Entre 2 y 5 ...	No sabe
Quién_es_su_acudiente_	STRING	enum patt...	12.935	0	0	10	Abuelo(a)	Mamá	Tío(a)
Cuántos_hermanos_tiene_	STRING	integer en...	12.935	0	0	12	0	1	9

Los resultados se presentan en la siguiente tabla:

Tabla 17. Resultados DQ Analyzer

Variable	% Nulos	Distintos	Únicos
ID	0.00%	12,934	12,933
País_de_nacimiento	0.01%	32	12
Departamento_o_estado_de_nacimiento	0.01%	131	72
Municipio_o_ciudad_de_nacimiento	0.02%	575	234
Edad_al_momento_de_la_encuesta	0.00%	51	6
Estado_civil	0.00%	7	0
Tipo_de_vinculación_al_Sistema_General_de_Seguridad_Social_en_Salud	0.00%	3	0
Al_momento_de_iniciar_sus_estudios_en_la_universidad_se_trasladó_de_Ciudad_o_Municipio_	0.00%	2	0
Estrato	0.00%	6	0
Tipo_de_residencia	0.00%	3	0
Tenencia_de_la_residencia	0.00%	5	0
Zona_de_la_residencia	0.00%	2	0
Estado_actual_del_padre	0.00%	3	0
Convive_con_el_padre_	0.00%	3	0
Nivel_educativo_del_padre	0.00%	11	0
Ocupación_del_padre	0.00%	7	0
Promedio_de_ingresos_del_padre	0.00%	7	0
Estado_actual_de_la_madre	0.00%	3	0
Convive_con_la_madre_	0.00%	3	0
Nivel_educativo_de_la_madre	0.00%	11	0
Ocupación_de_la_madre	0.00%	7	0
Promedio_de_ingresos_de_la_madre	0.00%	7	0
Quién_es_su_acudiente_	0.00%	10	0
Cuántos_hermanos_tiene_	0.00%	12	0
Cuál_es_el_lugar_que_ocupa_entre_sus_hermanos_	0.00%	11	0
Vive_con_alguno_de_ellos___Hermanos_	0.01%	3	0
Vive_con_otros_miembros_de_la_familia_	0.00%	2	0
Convive_con_otras_personas_que_no_sean_de_su_familia_	0.00%	2	0
Actualmente_tiene_personas_a_cargo_	0.00%	2	0
Usted_tiene_hijos_	0.00%	2	0
En_su_barrio_se_presentan_algunas_de_estas_situaciones_o_fenómenos___Barreras_invisibles_	0.00%	2	0
En_su_barrio_se_presentan_algunas_de_estas_situaciones_o_fenómenos___Conflicto_armado_	0.00%	2	0

Variable	% Nulos	Distintos	Únicos
En_su_barrio_se_presentan_algunas_de_estas_situaciones_o_fenómenos___Extorsión_	0.00%	2	0
En_su_barrio_se_presentan_algunas_de_estas_situaciones_o_fenómenos___Hurto_	0.00%	2	0
En_su_barrio_se_presentan_algunas_de_estas_situaciones_o_fenómenos___Consumo_de_sustancias_psicoactivas_	0.00%	2	0
En_su_barrio_se_presentan_algunas_de_estas_situaciones_o_fenómenos___Ninguna_de_las_anteriores_	0.00%	2	0
Ha_sido_víctima_del_conflicto_armado_	0.00%	2	0
Pertenece_a_alguna_etnia_o_resguardo_indígena_	0.00%	2	0
Quién_representa_la_figura_de_autoridad_en_la_familia_	0.00%	9	0
Quién_toma_las_decisiones_en_su_familia___Papá_	0.00%	2	0
Quién_toma_las_decisiones_en_su_familia___Mamá_	0.00%	2	0
Quién_toma_las_decisiones_en_su_familia___Hermano_a__	0.00%	2	0
Quién_toma_las_decisiones_en_su_familia___Abuelo_a__	0.00%	2	0
Quién_toma_las_decisiones_en_su_familia___Tío_a__	0.00%	2	0
Quién_toma_las_decisiones_en_su_familia___Pareja_	0.00%	2	0
Quién_toma_las_decisiones_en_su_familia___Suegro_a__	0.00%	2	0
Quién_toma_las_decisiones_en_su_familia___Usted_	0.00%	2	0
Le_satisface_el_apoyo_que_recibe_de_su_familia_cuando_tiene_algún_problema_y_o_necesidad_	0.00%	5	0
De_qué_manera_enfrenta_su_familia_las_crisis_usualmente___Entre_los_miembros_de_la_familia_se_piden_apoyo_	0.00%	2	0
De_qué_manera_enfrenta_su_familia_las_crisis_usualmente___Tratan_de_solucionar_la_situación_hablando_asertivamente_	0.00%	2	0
De_qué_manera_enfrenta_su_familia_las_crisis_usualmente___Cada_uno_se_va_por_su_lado_evadiendo_la_situación_	0.00%	2	0
De_qué_manera_enfrenta_su_familia_las_crisis_usualmente___Buscan_ayuda_profesional_para_orientarlos_	0.00%	2	0
De_qué_manera_enfrenta_su_familia_las_crisis_usualmente___Todos_son_indiferentes_	0.00%	2	0
De_qué_manera_enfrenta_su_familia_las_crisis_usualmente___Todos_se_alteran_y_se_empeora_la_situación_	0.00%	2	0
En_su_familia_han_vivido_alguno_de_los_siguientes_acontecimientos___Secuestro_	0.00%	2	0
En_su_familia_han_vivido_alguno_de_los_siguientes_acontecimientos___Asesinato_de_uno_de_los_miembros_de_la_familia_	0.00%	2	0

Variable	% Nulos	Distintos	Únicos
En_su_familia_han_vivido_alguno_de_los_siguientes_acontecimientos___Violación_	0.00%	2	0
En_su_familia_han_vivido_alguno_de_los_siguientes_acontecimientos___Extorsión_	0.00%	2	0
En_su_familia_han_vivido_alguno_de_los_siguientes_acontecimientos___Suicidio_	0.00%	2	0
En_su_familia_han_vivido_alguno_de_los_siguientes_acontecimientos___Desplazamiento_forzado_	0.00%	2	0
En_su_familia_han_vivido_alguno_de_los_siguientes_acontecimientos___Desaparición_	0.00%	2	0
En_su_familia_han_vivido_alguno_de_los_siguientes_acontecimientos___Ninguno_	0.00%	2	0
Actualmente_en_su_familia_presentan_alguna_de_las_siguientes_situaciones___Malas_Relaciones_intrafamiliares_	0.00%	2	0
Actualmente_en_su_familia_presentan_alguna_de_las_siguientes_situaciones___Fallecimiento_de_algún_familiar__primer_grado_consanguinidad__	0.00%	2	0
Actualmente_en_su_familia_presentan_alguna_de_las_siguientes_situaciones___Violencia_intrafamiliar_	0.00%	2	0
Actualmente_en_su_familia_presentan_alguna_de_las_siguientes_situaciones___Abuso_o_violencia_sexual_	0.00%	2	0
Actualmente_en_su_familia_presentan_alguna_de_las_siguientes_situaciones___Enfermedad_crónica_de_algún_pariente_	0.00%	2	0
Actualmente_en_su_familia_presentan_alguna_de_las_siguientes_situaciones___Separación_de_los_padres_	0.00%	2	0
Actualmente_en_su_familia_presentan_alguna_de_las_siguientes_situaciones___Alcoholismo_o_adicción_a_sustancias_	0.00%	2	0
Actualmente_en_su_familia_presentan_alguna_de_las_siguientes_situaciones___Desplazamiento_forzado_	0.00%	2	0
Actualmente_en_su_familia_presentan_alguna_de_las_siguientes_situaciones___Dificultades_económicas_de_la_familia_	0.00%	2	0
Actualmente_en_su_familia_presentan_alguna_de_las_siguientes_situaciones___Ninguna_de_las_anteriores_	0.00%	2	0
Posee_Necesidades_Educativas_Especiales__NEE__como___Discapacidad_Intelectual_	0.00%	2	0
Posee_Necesidades_Educativas_Especiales__NEE__como___Déficit_de_atención_con_Hiperactividad__TDAH__	0.00%	2	0
Posee_Necesidades_Educativas_Especiales__NEE__como___Hipoacusia_o_baja_audición_	0.00%	2	0
Posee_Necesidades_Educativas_Especiales__NEE__como___Sordera_profunda_	0.00%	2	0
Posee_Necesidades_Educativas_Especiales__NEE__como___Baja_Visión_diagnosticada_	0.00%	2	0

Variable	% Nulos	Distintos	Únicos
Posee_Necesidades_Educativas_Especiales__NEE__como___Ceguera_	0.00%	2	0
Posee_Necesidades_Educativas_Especiales__NEE__como___Trastornos_del_lenguaje_	0.00%	2	0
Posee_Necesidades_Educativas_Especiales__NEE__como___Discapacidad_motora_o_motriz_	0.00%	2	0
Posee_Necesidades_Educativas_Especiales__NEE__como___Autismo_	0.00%	2	0
Posee_Necesidades_Educativas_Especiales__NEE__como___Asperger_	0.00%	2	0
Posee_Necesidades_Educativas_Especiales__NEE__como___Parálisis_cerebral_	0.00%	2	0
Posee_Necesidades_Educativas_Especiales__NEE__como___Lesión_neuromuscular_	0.00%	2	0
Posee_Necesidades_Educativas_Especiales__NEE__como___Síndrome_de_Down_	0.00%	2	0
Posee_Necesidades_Educativas_Especiales__NEE__como___Ninguna_	0.00%	2	0
Padece_alguna_de_las_siguientes_enfermedades___Alergias_	0.00%	2	0
Padece_alguna_de_las_siguientes_enfermedades___Insuficiencia_Renal_	0.00%	2	0
Padece_alguna_de_las_siguientes_enfermedades___Artritis__lupus_	0.00%	2	0
Padece_alguna_de_las_siguientes_enfermedades___Diabetes___Hipertensión___Insuficiencia_cardíaca_	0.00%	2	0
Padece_alguna_de_las_siguientes_enfermedades___Hemofilia_	0.00%	2	0
Padece_alguna_de_las_siguientes_enfermedades___Cáncer_	0.00%	2	0
Padece_alguna_de_las_siguientes_enfermedades___Epilepsia_	0.00%	2	0
Padece_alguna_de_las_siguientes_enfermedades___Deficiencia_cardiaca_	0.00%	2	0
Padece_alguna_de_las_siguientes_enfermedades___Migraña_	0.00%	2	0
Padece_alguna_de_las_siguientes_enfermedades___Fibromialgia_	0.00%	2	0
Padece_alguna_de_las_siguientes_enfermedades___Colon_irritable___Gastritis_	0.00%	2	0
Padece_alguna_de_las_siguientes_enfermedades___Síndrome_de_fatiga_crónica_	0.00%	2	0
Padece_alguna_de_las_siguientes_enfermedades___Disautonomía_	0.00%	2	0
Padece_alguna_de_las_siguientes_enfermedades___Trastorno_de_ansiedad___Depresión___Trastorno_Bipolar_	0.00%	2	0

Variable	% Nulos	Distintos	Únicos
Padece_alguna_de_las_siguientes_enfermedades___Síndrome_de_hiperactividad_	0.00%	2	0
Padece_alguna_de_las_siguientes_enfermedades___Ninguna_de_las_anteriores_	0.00%	2	0
Actualmente_trabaja_	0.00%	2	0
Quién_es_el_principal_proveedor_económico_de_su_familia_	0.00%	9	0
Cuánto_suman_los_ingresos_mensuales_de_su_familia_	0.00%	6	0
Con_los_ingresos_que_actualmente_la_familia_recibe__logran_cubrir_sus_necesidades_básicas_	0.00%	2	0
Forma_de_pago_de_la_matrícula__De_contado__Efectivo__	0.00%	2	0
Forma_de_pago_de_la_matrícula__Entidad_Bancaria_	0.00%	2	0
Forma_de_pago_de_la_matrícula__ICETEX_	0.00%	2	0
Forma_de_pago_de_la_matrícula__Beneficiario_de_beca_	0.00%	2	0
Forma_de_pago_de_la_matrícula__Fondo_EPM_	0.00%	2	0
Forma_de_pago_de_la_matrícula__Tarjeta_de_crédito_	0.00%	2	0
Usted_considera_que_requiere_ayuda_financiera_para_el_pago_de_la_matrícula_	0.00%	2	0
Recibe_algún_tipo_de_apoyo_por_parte_de_la_Universidad_	0.00%	2	0
Ha_recibido_alguno_o_todos_de_los_siguientes_apoyos___Alimentación_	0.00%	2	0
Ha_recibido_alguno_o_todos_de_los_siguientes_apoyos___Fotocopias_	0.00%	2	0
Ha_recibido_alguno_o_todos_de_los_siguientes_apoyos___Transporte_	0.00%	2	0
Ha_recibido_alguno_o_todos_de_los_siguientes_apoyos___Ninguno_	0.00%	2	0
Considera_usted_que_requiere_de_otros_apoyos_como___Alimentación_	0.00%	2	0
Considera_usted_que_requiere_de_otros_apoyos_como___Fotocopias_	0.00%	2	0
Considera_usted_que_requiere_de_otros_apoyos_como___Transporte_	0.00%	2	0
Considera_usted_que_requiere_de_otros_apoyos_como___Materiales_para_la_elaboración_de_sus_trabajos_	0.00%	2	0
Considera_usted_que_requiere_de_otros_apoyos_como___Beca_	0.00%	2	0
Considera_usted_que_requiere_de_otros_apoyos_como___Crédito_	0.00%	2	0
Considera_usted_que_requiere_de_otros_apoyos_como___Ninguno_	0.00%	2	0

Variable	% Nulos	Distintos	Únicos
Durante_su_educación_media_presentó_alguna_de_las_siguientes_situaciones___Dificultad_en_adaptación_social_	0.00%	2	0
Durante_su_educación_media_presentó_alguna_de_las_siguientes_situaciones___Dificultades_de_salud_	0.00%	2	0
Durante_su_educación_media_presentó_alguna_de_las_siguientes_situaciones___Pérdida_de_años_	0.00%	2	0
Durante_su_educación_media_presentó_alguna_de_las_siguientes_situaciones___Dificultades_académicas_	0.00%	2	0
Durante_su_educación_media_presentó_alguna_de_las_siguientes_situaciones___Ninguno_	0.00%	2	0
Cuál_fue_el_principal_motivo_que_lo_llevó_a_elegir_la_Universidad_Pontificia_Bolivariana___Calidad_académica_	0.00%	2	0
Cuál_fue_el_principal_motivo_que_lo_llevó_a_elegir_la_Universidad_Pontificia_Bolivariana___Reconocimiento_en_el_medio_	0.00%	2	0
Cuál_fue_el_principal_motivo_que_lo_llevó_a_elegir_la_Universidad_Pontificia_Bolivariana___Formación_integral_	0.00%	2	0
Cuál_fue_el_principal_motivo_que_lo_llevó_a_elegir_la_Universidad_Pontificia_Bolivariana___Por_decisión_de_sus_padres_	0.00%	2	0
Cuál_fue_el_principal_motivo_que_lo_llevó_a_elegir_la_Universidad_Pontificia_Bolivariana___No_pasó_a_otra_institución_de_educación_superior_	0.00%	2	0
Motivo_por_el_cual_eligió_la_carrera__Sugerencia_recibida_por_orientación_profesional_	0.00%	2	0
Motivo_por_el_cual_eligió_la_carrera__Interés_propio_	0.00%	2	0
Motivo_por_el_cual_eligió_la_carrera__Reconocimeinto_en_el_medio_	0.00%	2	0
Motivo_por_el_cual_eligió_la_carrera__Presión_familiar_	0.00%	2	0
Motivo_por_el_cual_eligió_la_carrera__Presión_social_	0.00%	2	0
Motivo_por_el_cual_eligió_la_carrera__No_sabía_qué_estudiar_	0.00%	2	0
Motivo_por_el_cual_eligió_la_carrera__Interés_económico_	0.00%	2	0
Realizó_orientación_vocacional_antes_de_ingresar_a_la_universidad_	0.00%	2	0
Considera_que_tiene_dificultad_en_alguno_de_los_siguientes_aspectos___Trabajar_contenidos_matemáticos_y_numéricos_	0.00%	2	0
Considera_que_tiene_dificultad_en_alguno_de_los_siguientes_aspectos___Concentrarse_y_prestar_atención_	0.00%	2	0
Considera_que_tiene_dificultad_en_alguno_de_los_siguientes_aspectos___Comprensión_lectora_	0.00%	2	0

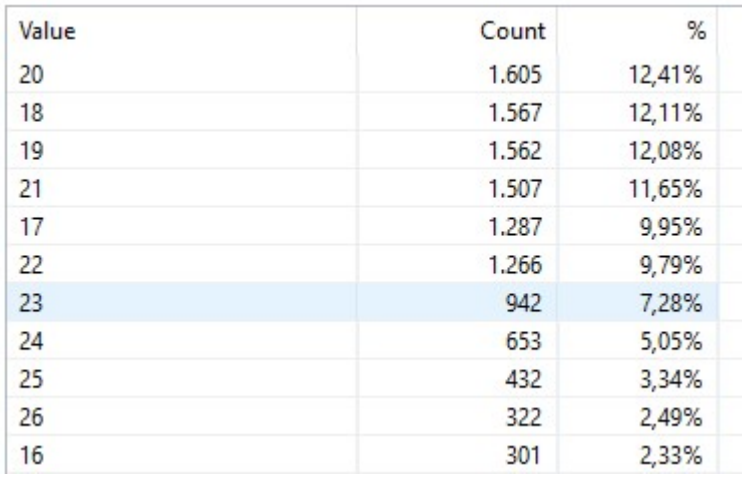
Variable	% Nulos	Distintos	Únicos
Considera_que_tiene_dificultad_en_alguno_de_los_siguientes_aspectos___Hablar_en_público_	0.00%	2	0
Considera_que_tiene_dificultad_en_alguno_de_los_siguientes_aspectos___Las_actividades_artísticas_y_manuales_	0.00%	2	0
Considera_que_tiene_dificultad_en_alguno_de_los_siguientes_aspectos___Memorizar_	0.00%	2	0
Considera_que_tiene_dificultad_en_alguno_de_los_siguientes_aspectos___Ver_bien_o_escuchar_bien_en_el_aula_	0.00%	2	0
Considera_que_tiene_dificultad_en_alguno_de_los_siguientes_aspectos___Ninguno_	0.00%	2	0
Con_cuáles_de_los_siguientes_recursos_NO_cuenta_para_el_estudio_en_casa___No_cuenta_con_computador_para_estudiar_y_realizar_sus_trabajos_académicos_	0.00%	2	0
Con_cuáles_de_los_siguientes_recursos_NO_cuenta_para_el_estudio_en_casa___No_cuenta_con_acceso_a_internet_como_herramienta_de_consulta_	0.00%	2	0
Con_cuáles_de_los_siguientes_recursos_NO_cuenta_para_el_estudio_en_casa___No_cuenta_con_libros_para_estudiar_	0.00%	2	0
Con_cuáles_de_los_siguientes_recursos_NO_cuenta_para_el_estudio_en_casa___No_tiene_espacio_para_estudiar_	0.00%	2	0
Con_cuáles_de_los_siguientes_recursos_NO_cuenta_para_el_estudio_en_casa___Ninguno_	0.00%	2	0
Califique_el_espacio_que_tiene_en_su_casa_para_estudiar	0.00%	6	0
Cuáles_de_los_siguientes_métodos_de_estudio_utiliza_usualmente___Subrayar_	0.00%	2	0
Cuáles_de_los_siguientes_métodos_de_estudio_utiliza_usualmente___Sacar_ideas_principales_	0.00%	2	0
Cuáles_de_los_siguientes_métodos_de_estudio_utiliza_usualmente___Apuntes_	0.00%	2	0
Cuáles_de_los_siguientes_métodos_de_estudio_utiliza_usualmente___Mapas_conceptuales__resumen__lluvia_de_ideas__etc_	0.00%	2	0
Cuáles_de_los_siguientes_métodos_de_estudio_utiliza_usualmente___Memorización_mecánica__sin_comprender_lo_que_estudio__	0.00%	2	0
Cuáles_de_los_siguientes_métodos_de_estudio_utiliza_usualmente___Repasar_después_de_cada_clase_	0.00%	2	0
Cuáles_de_los_siguientes_métodos_de_estudio_utiliza_usualmente___Estudiar_minutos_antes_del_examen_o_exposición_	0.00%	2	0
Cómo_se_siente_con_los_métodos_de_estudio_que_utiliza_usualmente_	0.00%	4	0

Variable	% Nulos	Distintos	Únicos
Participa_en_grupos_o_realiza_actividades_relacionadas_con___Arte_	0.00%	2	0
Participa_en_grupos_o_realiza_actividades_relacionadas_con___Deporte_	0.00%	2	0
Participa_en_grupos_o_realiza_actividades_relacionadas_con___Formación_personal_	0.00%	2	0
Participa_en_grupos_o_realiza_actividades_relacionadas_con___Espiritual_	0.00%	2	0
Participa_en_grupos_o_realiza_actividades_relacionadas_con___Sociales_	0.00%	2	0
Participa_en_grupos_o_realiza_actividades_relacionadas_con___Proyección_comunitaria_	0.00%	2	0
Participa_en_grupos_o_realiza_actividades_relacionadas_con___Emprendimiento_	0.00%	2	0
Participa_en_grupos_o_realiza_actividades_relacionadas_con___Ninguno_	0.00%	2	0
Qué_tipo_de_actividad_realiza_con_mayor_frecuencia_en_su_tiempo_libre___Ver_TV_	0.00%	2	0
Qué_tipo_de_actividad_realiza_con_mayor_frecuencia_en_su_tiempo_libre___Escuchar_música_	0.00%	2	0
Qué_tipo_de_actividad_realiza_con_mayor_frecuencia_en_su_tiempo_libre___Dormir_	0.00%	2	0
Qué_tipo_de_actividad_realiza_con_mayor_frecuencia_en_su_tiempo_libre___Jugar_video_Juegos_	0.00%	2	0
Qué_tipo_de_actividad_realiza_con_mayor_frecuencia_en_su_tiempo_libre___Deporte_	0.00%	2	0
Qué_tipo_de_actividad_realiza_con_mayor_frecuencia_en_su_tiempo_libre___Compartir_con_los_amigos_	0.00%	2	0
Qué_tipo_de_actividad_realiza_con_mayor_frecuencia_en_su_tiempo_libre___Navegar_en_internet_	0.00%	2	0
Qué_tipo_de_actividad_realiza_con_mayor_frecuencia_en_su_tiempo_libre___Leer_	0.00%	2	0
Qué_tipo_de_actividad_realiza_con_mayor_frecuencia_en_su_tiempo_libre___Ninguna_	0.00%	2	0
Realiza_actividad_física_	0.00%	2	0
Realiza_alguna_actividad_cultural_o_artística_	0.00%	2	0
Qué_tipo_de_espectáculos_o_actividades_disfruta___Conciertos_	0.00%	2	0
Qué_tipo_de_espectáculos_o_actividades_disfruta___Shows_de_baile_	0.00%	2	0
Qué_tipo_de_espectáculos_o_actividades_disfruta___Cuentería__expresión_oral_o_corporal_	0.00%	2	0
Qué_tipo_de_espectáculos_o_actividades_disfruta___Demostraciones_de_arte_	0.00%	2	0
Qué_tipo_de_espectáculos_o_actividades_disfruta___Cine_	0.00%	2	0
Qué_tipo_de_espectáculos_o_actividades_disfruta___Ninguno_	0.00%	2	0

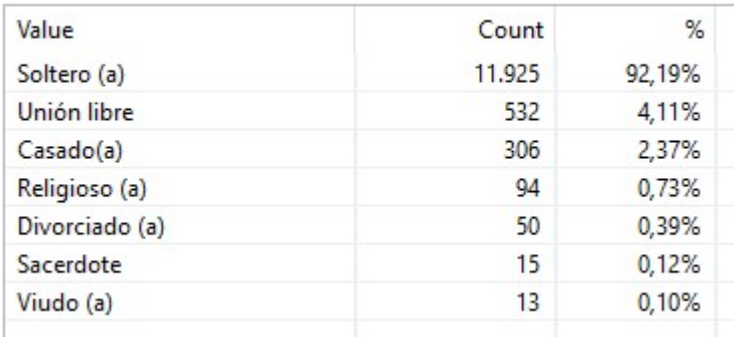
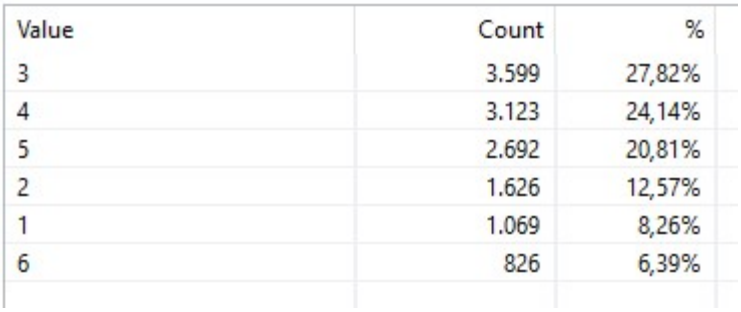
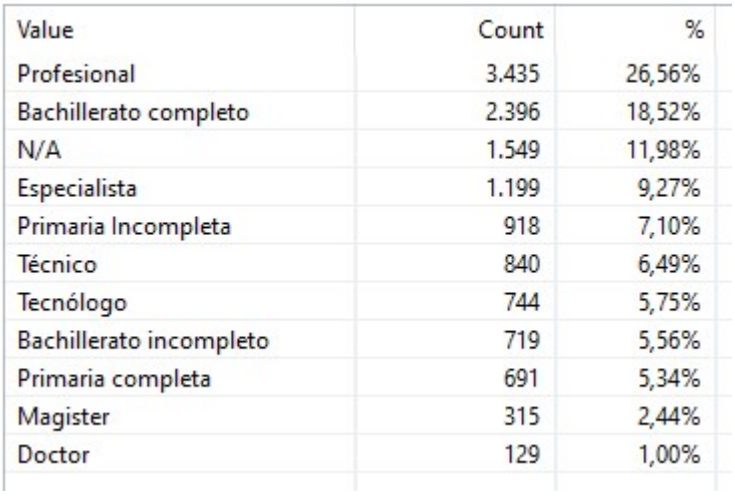
Variable	% Nulos	Distintos	Únicos
Programa_Académico	15.07%	50	6
Créditos_Programa	15.07%	23	1
Nro_Semestres_Programa	15.07%	6	0
Estado_Académico	15.07%	7	0
Edad_al_iniciar_el_programa	15.07%	42	9
Ubicación_Semestral	15.07%	13	0
Calendario_Colegio	17.92%	3	0
Género_Población_Colegio	17.92%	3	0
Naturaleza_Colegio	17.92%	2	0
Categoría_Colegio	17.92%	7	0
Porcentaje_Aprobados___Intentados	15.08%	2,944	162
Promedio_Notas	15.42%	10,159	9,606
Rezagado	15.07%	3	0
Desertor	15.07%	2	0

A continuación, se presenta la descripción estadística de algunas variables en cuanto a análisis de frecuencias.

Tabla 18. Descripción estadística - análisis de frecuencias

Variable	Resultado
Edad al momento de la encuesta	

Value	Count	%
20	1.605	12,41%
18	1.567	12,11%
19	1.562	12,08%
21	1.507	11,65%
17	1.287	9,95%
22	1.266	9,79%
23	942	7,28%
24	653	5,05%
25	432	3,34%
26	322	2,49%
16	301	2,33%

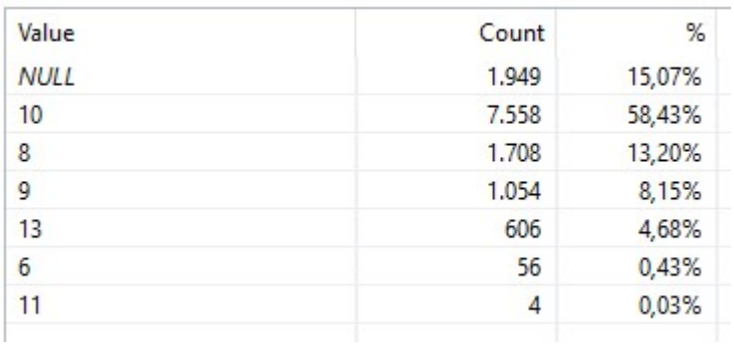
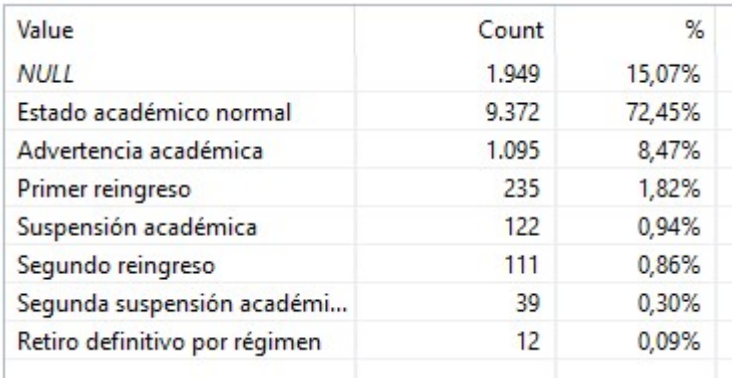
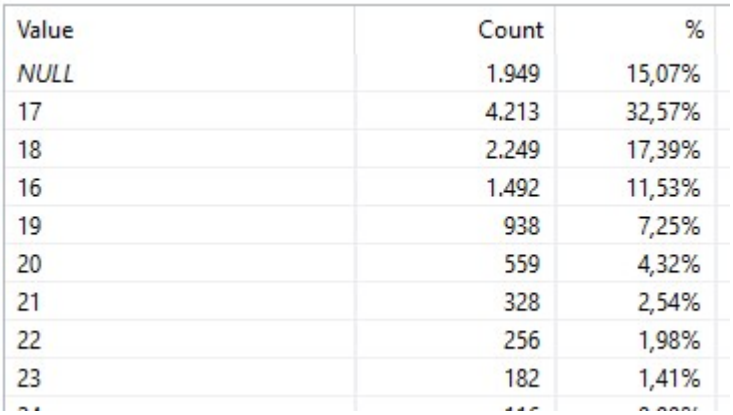
Variable	Resultado
Estado civil	
Estrato	
Nivel educativo del padre	

Value	Count	%
Soltero (a)	11.925	92,19%
Unión libre	532	4,11%
Casado(a)	306	2,37%
Religioso (a)	94	0,73%
Divorciado (a)	50	0,39%
Sacerdote	15	0,12%
Viudo (a)	13	0,10%

Value	Count	%
3	3.599	27,82%
4	3.123	24,14%
5	2.692	20,81%
2	1.626	12,57%
1	1.069	8,26%
6	826	6,39%

Value	Count	%
Profesional	3.435	26,56%
Bachillerato completo	2.396	18,52%
N/A	1.549	11,98%
Especialista	1.199	9,27%
Primaria Incompleta	918	7,10%
Técnico	840	6,49%
Tecnólogo	744	5,75%
Bachillerato incompleto	719	5,56%
Primaria completa	691	5,34%
Magister	315	2,44%
Doctor	129	1,00%

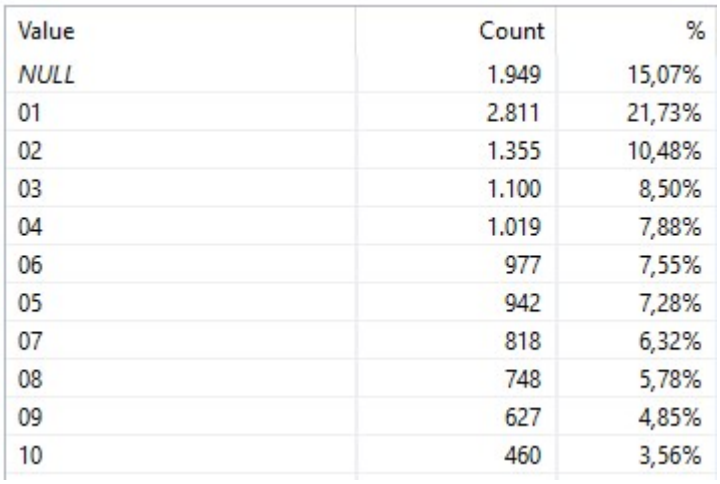
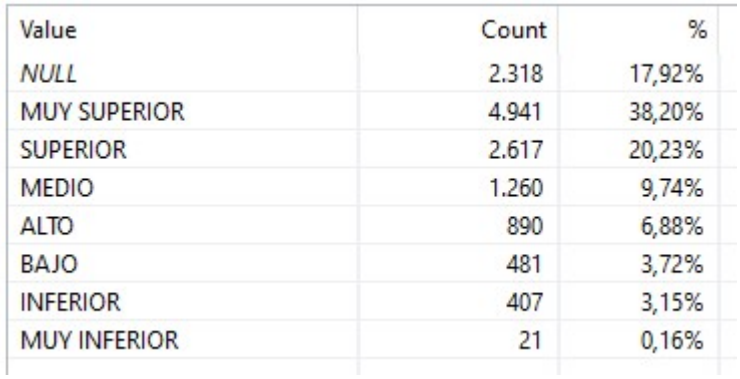
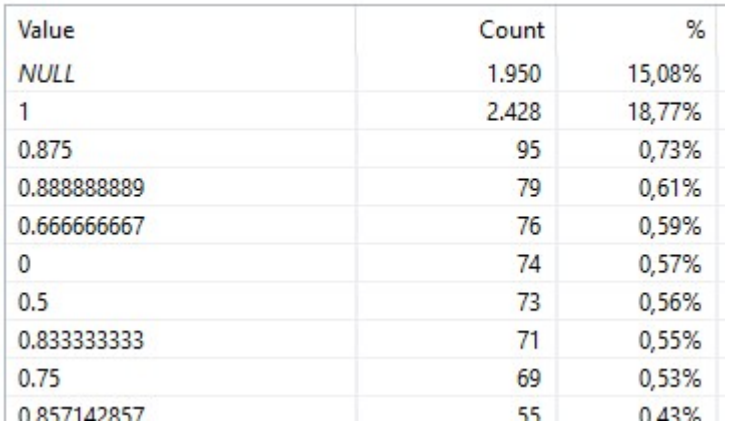
Variable	Resultado
Nivel educativo de la madre	Value / Count / % Profesional 3.712 28,70% Bachillerato completo 2.884 22,30% Técnico 1.279 9,89% Tecnólogo 1.141 8,82% Especialista 1.125 8,70% Primaria Incompleta 734 5,67% Bachillerato incompleto 716 5,54% Primaria completa 643 4,97% N/A 393 3,04% Magister 240 1,86% Doctor 68 0,53%
Cuánto suman los ingresos mensuales de su familia?	Value / Count / % Entre 2 y 5 salarios mínimos 3.766 29,11% No sabe 2.971 22,97% Entre 1 y 2 salarios mínimos 2.322 17,95% Entre 5 y 10 salarios mínimos 1.890 14,61% Más de 10 salarios mínimos 1.028 7,95% Entre 0 y 1 salario mínimo 958 7,41%
Programa Académico	Value / Count / % *NULL* 1.949 15,07% Arquitectura-Med 1.144 8,84% Comunicación Social-Period-... 984 7,61% Derecho-Med 679 5,25% Medicina-Med 606 4,68% Psicología (D)-Med 592 4,58% Enfermería-Med 553 4,28% Diseño Industrial-Med 510 3,94% Ing Aeronáutica-Med 492 3,80% Admón de Empresas-Med 472 3,65%

Variable	Resultado
Nro Semestres Programa	
Estado Académico	
Edad al iniciar el programa	

Value	Count	%
NULL	1.949	15,07%
10	7.558	58,43%
8	1.708	13,20%
9	1.054	8,15%
13	606	4,68%
6	56	0,43%
11	4	0,03%

Value	Count	%
NULL	1.949	15,07%
Estado académico normal	9.372	72,45%
Advertencia académica	1.095	8,47%
Primer reingreso	235	1,82%
Suspensión académica	122	0,94%
Segundo reingreso	111	0,86%
Segunda suspensión académi...	39	0,30%
Retiro definitivo por régimen	12	0,09%

Value	Count	%
NULL	1.949	15,07%
17	4.213	32,57%
18	2.249	17,39%
16	1.492	11,53%
19	938	7,25%
20	559	4,32%
21	328	2,54%
22	256	1,98%
23	182	1,41%

Variable	Resultado
Ubicación Semestral	
Categoría Colegio	
Porcentaje Aprobados / Intentados	100 most common values: 

Value	Count	%
NULL	1.949	15,07%
01	2.811	21,73%
02	1.355	10,48%
03	1.100	8,50%
04	1.019	7,88%
06	977	7,55%
05	942	7,28%
07	818	6,32%
08	748	5,78%
09	627	4,85%
10	460	3,56%

Value	Count	%
NULL	2.318	17,92%
MUY SUPERIOR	4.941	38,20%
SUPERIOR	2.617	20,23%
MEDIO	1.260	9,74%
ALTO	890	6,88%
BAJO	481	3,72%
INFERIOR	407	3,15%
MUY INFERIOR	21	0,16%

Value	Count	%
NULL	1.950	15,08%
1	2.428	18,77%
0.875	95	0,73%
0.888888889	79	0,61%
0.666666667	76	0,59%
0	74	0,57%
0.5	73	0,56%
0.833333333	71	0,55%
0.75	69	0,53%
0.857142857	55	0,43%

<table>
<tr><th>Variable</th><th>Resultado</th></tr>
<tr><td>Promedio Notas</td><td>100 most common values:<table><tr><th>Value</th><th>Count</th><th>%</th></tr><tr><td>NULL</td><td>1.994</td><td>15,42%</td></tr><tr><td>4.49</td><td>7</td><td>0,05%</td></tr><tr><td>3.54</td><td>6</td><td>0,05%</td></tr><tr><td>3.76</td><td>6</td><td>0,05%</td></tr><tr><td>3.85</td><td>6</td><td>0,05%</td></tr><tr><td>3.89</td><td>6</td><td>0,05%</td></tr><tr><td>3.94</td><td>6</td><td>0,05%</td></tr><tr><td>3.95</td><td>6</td><td>0,05%</td></tr><tr><td>4.24</td><td>6</td><td>0,05%</td></tr><tr><td>4.31</td><td>6</td><td>0,05%</td></tr></table></td></tr>
<tr><td>Rezagado</td><td><table><tr><th>Value</th><th>Count</th><th>%</th></tr><tr><td>NULL</td><td>1.949</td><td>15,07%</td></tr><tr><td>NO</td><td>4.280</td><td>33,09%</td></tr><tr><td>SI</td><td>3.463</td><td>26,77%</td></tr><tr><td>No se puede determinar</td><td>3.243</td><td>25,07%</td></tr></table></td></tr>
<tr><td>Desertor (Variable objetivo)</td><td><table><tr><th>Value</th><th>Count</th><th>%</th></tr><tr><td>NULL</td><td>1.949</td><td>15,07%</td></tr><tr><td>NO</td><td>10.519</td><td>81,32%</td></tr><tr><td>SI</td><td>467</td><td>3,61%</td></tr></table></td></tr>
</table>

11.5 Limpieza de datos

11.5.1 Registros duplicados

Utilizando DQ Analyzer se buscan registros duplicados por el ID del estudiante, encontrándose que 2 registros deben ser eliminados al presentar esta condición. Después esta operación se cuenta con 12933 registros.

11.5.2 Datos atípicos

En ninguna variable se encontraron datos atípicos:

- En variables numéricas se analizaron los valores máximo y mínimo según las reglas del negocio.
- En variables categóricas se analizaron las categorías existentes según las reglas del negocio.

Para establecer los rangos y valores normales de las variables se tuvo el acompañamiento de la unidad de Permanencia de la Universidad.

11.5.3 Datos ausentes

Estas son las columnas que, siendo previamente analizadas, presentan registros vacíos. Se decide eliminar los registros.

Tabla 19. Datos ausentes

Variable	% Nulos
País de nacimiento	0.01%
Departamento o estado de nacimiento	0.01%
Municipio o ciudad de nacimiento	0.02%
Vive con alguno de ellos (hermanos)	0.01%
Programa Académico	15.07%
Créditos Programa	15.07%
Nro. Semestres Programa	15.07%
Estado Académico	15.07%
Edad al iniciar el programa	15.07%
Ubicación Semestral	15.07%
Género Población Colegio	17.92%
Calendario Colegio	17.92%
Naturaleza Colegio	17.92%
Categoría Colegio	17.92%
Porcentaje aprobados vs Intentados	15.08%
Promedio notas	15.42%
Rezagado	15.07%
Desertor	15.07%

Después de eliminar 2362 registros con datos ausentes, se tiene un set de datos con 10571 registros, 202 variables predictoras y la variable objetivo. No se realizó imputación de los 2362 registros dado que los datos ausentes correspondían a variables académicas como el programa académico y otras variables categóricas.

11.6 Análisis de correlaciones

En esta sección se analizan: las variables que están altamente correlacionadas entre sí, y las correlaciones de las variables predictoras con la variable objetivo.

11.6.1 Correlación entre variables

Usando Python, se calcula la matriz de correlaciones. Como se dijo anteriormente, se buscan correlaciones altas y como criterio se establece que sean correlaciones en valor absoluto desde 0.7. Una pequeña fracción de la misma se muestra en la ilustración número 7.

Ilustración 8. Matriz de correlaciones

Mínima	0.0214	0.0756	0.1182	0.2109	0.0921	0.1412	0.2050	0.4083	0.3172	0.0700	0.0628	0.6337	0.6337	0.3682	0.5295	0.4694	0.3940
Máxima	0.5387	0.5613	0.5613	0.8348	0.3413	0.1777	0.4174	0.2729	0.1941	0.2127	0.0840	0.3190	0.3817	0.2882	0.3817	0.3615	0.1908
	País de r	Depa	Muni	Edad	Estac	Tipo	Al mo	Estra	Tipo	Tene	Zona	Estac	Con	Nive	Ocup	Pron	Estac
País de nacimiento	**1**	0.5387	0.3111	0.0172	-0.0104	0.0138	0.0021	0.0259	-0.0107	-0.0017	0.0006	-0.0119	0.0171	0.0228	0.0160	0.0036	0.0085
Departamento o estado de nacimiento	0.5387	**1**	0.5613	-0.0187	0.0055	0.0673	0.3315	0.0195	0.0404	0.0967	-0.0044	-0.0155	0.0969	0.0040	-0.0028	0.0184	0.0102
Municipio o ciudad de nacimiento	0.3111	0.5613	**1**	0.0407	0.0508	0.1152	0.3155	-0.1149	0.0867	0.0916	0.0134	0.0272	0.0774	0.0826	0.0063	0.0305	0.0179
Edad al momento de la encuesta	0.0172	-0.0187	0.0407	**1**	0.3345	-0.0136	-0.1037	0.0448	-0.0635	-0.0009	0.0043	0.1196	-0.0393	0.0094	-0.0190	-0.0623	0.1053
Estado civil	-0.0104	0.0055	0.0508	0.3345	**1**	0.0158	0.0250	-0.0715	0.0253	0.0440	0.0396	0.0629	0.0227	0.0317	-0.0165	-0.0109	0.0532
Tipo de vinculación al Sistema General de S	0.0138	0.0673	0.1152	-0.0136	0.0158	**1**	0.1465	-0.1412	0.1064	0.0614	0.0219	0.0747	0.0203	0.0239	-0.0263	0.0141	0.0466
Al momento de iniciar sus estudios en la unir	0.0021	0.3315	0.3155	-0.1037	0.0250	0.1465	**1**	-0.0805	0.1236	0.2127	-0.0069	-0.0011	0.2038	0.0204	-0.0061	0.0305	-0.0272
Estrato	0.0259	0.0195	-0.1149	0.0448	-0.0715	-0.1412	-0.0805	**1**	-0.3172	-0.0350	-0.0628	-0.0981	-0.0060	-0.1176	-0.0149	0.0404	-0.0238
Tipo de residencia	-0.0107	0.0404	0.0867	-0.0635	0.0253	0.1064	0.1236	-0.3172	**1**	0.1616	0.0840	0.0362	-0.0073	0.0706	0.0053	0.0220	0.0018
Tenencia de la residencia	-0.0017	0.0967	0.0916	-0.0009	0.0440	0.0614	0.2127	-0.0350	0.1616	**1**	0.0258	0.0035	0.0717	0.0159	0.0059	0.0111	0.0020
Zona de la residencia	0.0006	-0.0044	0.0134	0.0043	0.0396	0.0219	-0.0069	-0.0628	0.0840	0.0258	**1**	-0.0003	-0.0056	0.0095	0.0068	0.0246	0.0253
Estado actual del padre	-0.0119	-0.0155	0.0272	0.1196	0.0629	0.0747	-0.0011	-0.0981	0.0362	0.0035	-0.0003	**1**	-0.6337	-0.3682	-0.5295	-0.4694	0.1368
Convive con el padre-	0.0171	0.0969	0.0774	-0.0393	0.0227	0.0203	0.2038	-0.0060	-0.0073	0.0717	-0.0056	-0.6337	**1**	0.2820	0.3817	0.3483	-0.0758
Nivel educativo del padre	0.0228	0.0040	0.0826	0.0094	0.0317	0.0239	0.0204	-0.1176	0.0706	0.0159	0.0095	-0.3682	0.2820	**1**	0.2461	0.2882	-0.0365
Ocupación del padre	0.0160	-0.0028	0.0063	-0.0190	-0.0165	-0.0263	-0.0061	-0.0149	0.0053	0.0059	0.0068	-0.5295	0.3817	0.2461	**1**	0.3464	-0.0500
Promedio de ingresos del padre	0.0036	0.0184	0.0305	-0.0623	-0.0109	0.0141	0.0305	0.0404	0.0220	0.0111	0.0246	-0.4694	0.3483	0.2882	0.3464	**1**	-0.0523
Estado actual de la madre	0.0085	0.0102	0.0179	0.1053	0.0532	0.0466	-0.0272	-0.0238	0.0018	0.0020	0.0253	0.1368	-0.0758	-0.0365	-0.0500	-0.0523	**1**
Convive con la madre-	0.0064	0.1600	0.1804	0.0666	0.1387	0.0717	0.4174	-0.0645	0.0726	0.1312	-0.0196	0.0042	0.2949	0.0251	-0.0197	0.0171	-0.3940
Nivel educativo de la madre	-0.0064	-0.0111	0.0743	0.0685	0.0638	0.0512	0.0240	-0.1769	0.0872	0.0245	0.0067	0.0257	-0.0039	0.2249	0.0071	0.0390	-0.1951
Ocupación de la madre	0.0037	0.0083	0.0219	0.0789	0.0150	0.0410	0.0347	0.0167	0.0067	-0.0295	0.0079	-0.0077	-0.0055	0.0119	-0.0257	0.0055	-0.2587
Promedio de ingresos de la madre	-0.0167	-0.0068	-0.0400	-0.0560	-0.0364	-0.0682	0.0031	0.0829	-0.0521	-0.0124	-0.0077	-0.0094	0.0309	-0.0330	-0.0220	0.1348	-0.2720
Quién es su acudiente-	0.0002	0.0824	0.1507	0.2005	0.2490	0.0784	0.1941	-0.1012	0.0710	0.1041	0.0312	0.0800	0.0530	0.0371	-0.0138	-0.0028	0.1348
Cuántos hermanos tiene-	-0.0085	0.0853	0.1574	0.1703	0.1172	0.0921	0.1553	-0.0899	0.0921	0.0559	0.0187	-0.0185	0.0641	0.0937	0.0163	0.0534	0.0154
Cuál es el lugar que ocupa entre sus hermar	-0.0074	0.0557	0.0999	0.0955	0.0589	0.0430	0.0982	-0.0216	0.0436	0.0108	0.0101	-0.0378	0.0039	0.0717	0.0401	0.0333	-0.0005
Vive con alguno de ellos? (Hermanos)	0.0064	0.1247	0.1378	0.1494	0.1333	0.0581	0.2440	-0.0346	0.0879	0.0953	-0.0001	-0.0490	0.1249	0.0490	0.0201	0.0421	-0.0019
Vive con otros miembros de la familia-	-0.0044	-0.0306	0.0039	-0.0476	-0.0240	0.0242	0.0165	-0.1341	0.0482	0.0066	-0.0398	0.1184	0.0279	-0.0372	-0.0529	-0.0517	0.0342
Convive con otras personas que no sean de	0.0180	0.1753	0.1675	-0.0290	0.1032	0.0843	0.3558	0.0049	0.1824	0.2010	0.0037	0.0317	0.1344	-0.0084	-0.0211	0.0248	0.0253
Actualmente tiene personas a cargo-	-0.0150	-0.0113	0.0131	0.2787	0.2095	-0.0219	-0.0019	-0.0630	-0.0057	0.0423	0.0271	0.0654	0.0087	-0.0011	-0.0212	-0.0195	0.0427

Siendo así, y después de analizar todas las correlaciones en la matriz, se puede decir que las siguientes variables tienen alta correlación numérica con un valor mayor a 0.7 (en valor absoluto).

Tabla 20. Variables correlacionadas

Variable	Correlacionada con	Variable a eliminar
Edad al iniciar el programa	Edad al momento de la encuesta	Edad al iniciar el programa
Cuántos hermanos tiene	Cuál es el lugar que ocupa entre sus hermanos	No se elimina ninguna ya que no son redundantes
Créditos programa	Número de semestres programa	Créditos programa ya que tiene menos correlación con la variable deserción
Promedio notas	Porcentaje aprobados / intentados	No se elimina ninguna ya que no son redundantes

Después de eliminar estas dos variables se cuenta con 200 variables predictoras y los mismos 10571 registros.

11.6.2 Correlación con la variable objetivo

Se busca la correlación con la variable DESERTOR para ver cuáles tienen mayor relación. Se utiliza Weka, obteniendo un ranking como el siguiente:

Ilustración 9. Correlación con la variable objetivo

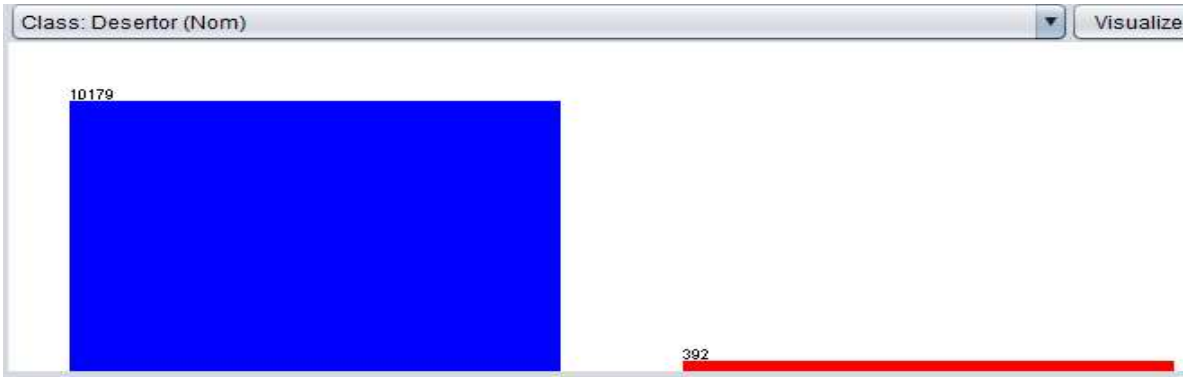
```
Attribute Evaluator (supervised, Class (nominal): 201 Desertor):
        Correlation Ranking Filter
Ranked attributes:
 0.2433354    198 Porcentaje Aprobados / Intentados
 0.202217     199 Promedio Notas
 0.1422846    192 Estado Académico
 0.0726402    191 Nro Semestres Programa
 0.0686922    200 Rezagado
 0.0620024    123 Considera usted que requiere de otros apoyos como: [Crédito]
 0.0619798    193 Ubicación Semestral
 0.0534261      7 Al momento de iniciar sus estudios en la universidad se trasladó de Ciudad o Municipio-
 0.0523159    128 Durante su educación media presentó alguna de las siguientes situaciones- [Dificultades académicas]
 0.0510476    136 Motivo por el cual eligió la carrera [Interés propio]
 0.0510283    138 Motivo por el cual eligió la carrera [Presión familiar]
 0.0499235    129 Durante su educación media presentó alguna de las siguientes situaciones- [Ninguno]
 0.0452771    102 Actualmente trabaja-
 0.0425702     23  Cuántos hermanos tiene-
 0.0418186    178 Qué tipo de actividad realiza con mayor frecuencia en su tiempo libre- [Compartir con los amigos]
 0.0409187      4 Edad al momento de la encuesta
 0.0402145     73 Posee Necesidades Educativas Especiales (NEE) como: [Déficit de atención con Hiperactividad (TDAH)]
 0.0399734    109 Forma de pago de la matrícula [Beneficiario de beca]
 0.0398602    127 Durante su educación media presentó alguna de las siguientes situaciones- [Pérdida de años]
 0.0387518      6 Tipo de vinculación al Sistema General de Seguridad Social en Salud
 0.0369324    175 Qué tipo de actividad realiza con mayor frecuencia en su tiempo libre- [Dormir]
```

Se nombran las cinco variables que más influyen en la deserción según este set de datos y este método aplicado, ellas son: Porcentaje aprobados/intentados, promedio de notas, estado académico, número de semestres del programa y rezagado.

11.7 Balanceo de datos

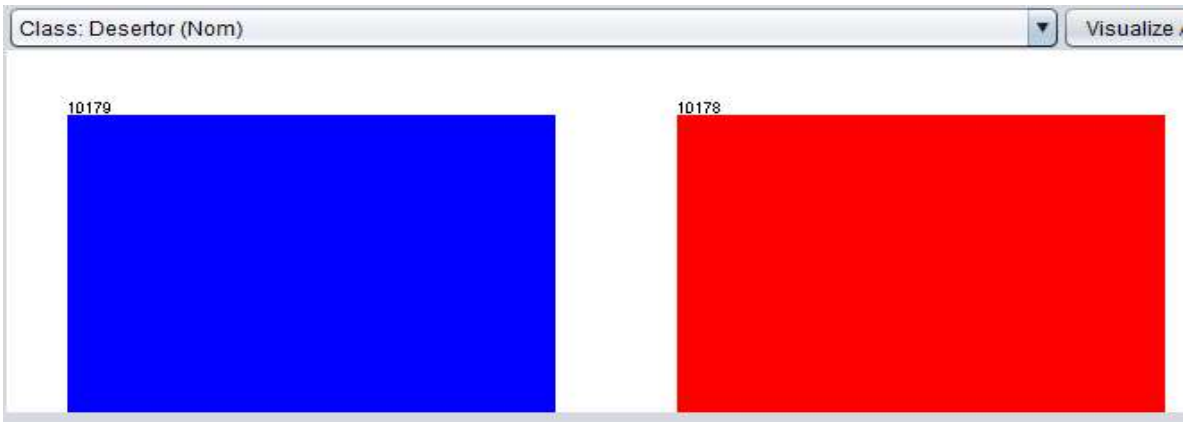
El histograma de frecuencias de la variable DESERTOR nos muestra que hay un desbalanceo en la variable objetivo.

Ilustración 10. Variable objetivo desbalanceada



Con ayuda de Weka, se aplica el filtro SMOTE para balancear los registros, obteniendo lo siguiente.

Ilustración 11. Variable objetivo balanceada



12. MODELADO

Para realizar los siguientes análisis se tienen 2 conjuntos de datos conv200 variables predictoras, el primer conjunto son datos NO BALANCEADOS, y el segundo conjunto son datos BALANCEADOS. A continuación, se detalla la creación de los modelos analíticos, se indica aquellos que usaron los datos balanceados:

- Selección de factores que más tiene relación con la deserción por medio de un sistema de votación de 3 métodos: correlaciones, ganancia y árbol de decisión (datos balanceados).
- Perfiles de estudiantes que desertan por medio de un análisis de clustering.
- Perfiles de estudianes que no desertan por medio de un análisis de clustering
- Reglas de asociación para buscar la co-ocurrencia de eventos en los datos
- Predicción de deserción usando todas las variables (datos balanceados)
- Predicción de deserción usando sólo las variables seleccionadas en el primer experimento (datos balanceados).

12.1 MODELO 1: Selección de factores

Objetivo: Seleccionar las variables más relevantes para la deserción.

Métodos: ganancia del atributo, análisis de correlaciones y árbol de decisión.

12.1.1 Ganancia del atributo

Este método (InfoGainAttributeEval) lo ofrece Weka para clasificar los atributos de forma individual basado en la ganancia de la información respecto a la variable

objetivo. Se aplica al set de datos sin balancear y se obtienen las primeras 30 variables que, según este método, más influyen en la deserción.

Tabla 21. Resultado método ganancia del atributo

Variable	Ganancia
Porcentaje Aprobados / Intentados	0.02946534
Promedio Notas	0.02486397
Municipio o ciudad de nacimiento	0.02249378
Estado Académico	0.02114532
Programa Académico	0.00789112
Ubicación Semestral	0.00776753
Rezagado	0.00776630
Departamento o estado de nacimiento	0.00477538
Edad al momento de la encuesta	0.00376837
Nro Semestres Programa	0.00350152
Categoría Colegio	0.00283616
Quién es su acudiente-	0.00281852
Considera usted que requiere de otros apoyos como: [Crédito]	0.00233352
Al momento de iniciar sus estudios en la universidad se trasladó de Ciudad o Municipio-	0.00187305
Durante su educación media presentó alguna de las siguientes situaciones- [Ninguno]	0.00166499
Durante su educación media presentó alguna de las siguientes situaciones- [Dificultades académicas]	0.00165443
Motivo por el cual eligió la carrera [Interés propio]	0.00152156
Forma de pago de la matrícula [Beneficiario de beca]	0.00136921
Cuántos hermanos tiene-	0.00131796
Actualmente trabaja-	0.00130360
Qué tipo de actividad realiza con mayor frecuencia en su tiempo libre- [Compartir con los amigos]	0.00128028
Motivo por el cual eligió la carrera [Presión familiar]	0.00121395

Variable	Ganancia
País de nacimiento	0.00116376
Ocupación del padre	0.00114528
Estado civil	0.00110757
Promedio de ingresos del padre	0.00108540
Tipo de vinculación al Sistema General de Seguridad Social en Salud	0.00106301
Cómo se siente con los métodos de estudio que utiliza usualmente:	0.00103710
Qué tipo de actividad realiza con mayor frecuencia en su tiempo libre-[Dormir]	0.00101438
Nivel educativo de la madre	0.00100472

12.1.2 Análisis de correlaciones

Tomando el análisis de correlaciones con la variable objetivo ya realizado en Weka, aplicado al set de datos sin balancear, se toman las primeras 30 variables.

Tabla 22. Resultado método análisis de correlaciones

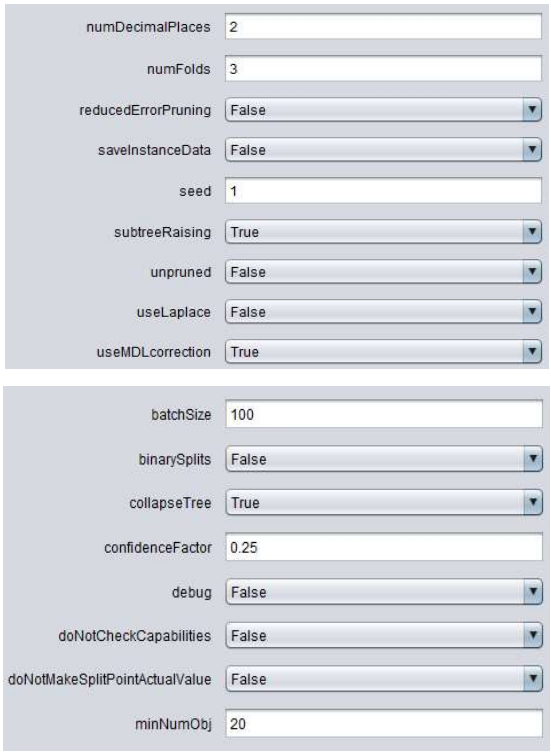
Variable	Correlación
Porcentaje Aprobados / Intentados	0.243335
Promedio Notas	0.202217
Estado Académico	0.142285
Nro Semestres Programa	0.072640
Rezagado	0.068692
Considera usted que requiere de otros apoyos como: [Crédito]	0.062002
Ubicación Semestral	0.061980
Al momento de iniciar sus estudios en la universidad se trasladó de Ciudad o Municipio-	0.053426
Durante su educación media presentó alguna de las siguientes situaciones- [Dificultades académicas]	0.052316
Motivo por el cual eligió la carrera [Interés propio]	0.051048

Variable	Correlación
Motivo por el cual eligió la carrera [Presión familiar]	0.051028
Durante su educación media presentó alguna de las siguientes situaciones- [Ninguno]	0.049924
Actualmente trabaja-	0.045277
Cuántos hermanos tiene-	0.042570
Qué tipo de actividad realiza con mayor frecuencia en su tiempo libre- [Compartir con los amigos]	0.041819
Edad al momento de la encuesta	0.040919
Posee Necesidades Educativas Especiales (NEE) como: [Déficit de atención con Hiperactividad (TDAH)]	0.040215
Forma de pago de la matrícula [Beneficiario de beca]	0.039973
Durante su educación media presentó alguna de las siguientes situaciones- [Pérdida de años]	0.039860
Tipo de vinculación al Sistema General de Seguridad Social en Salud	0.038752
Qué tipo de actividad realiza con mayor frecuencia en su tiempo libre- [Dormir]	0.036932
Actualmente tiene personas a cargo-	0.036422
Departamento o estado de nacimiento	0.035986
Forma de pago de la matrícula [De contado (Efectivo)]	0.033882
Motivo por el cual eligió la carrera [Presión social]	0.032660
Forma de pago de la matrícula [Fondo EPM]	0.031266
Padece alguna de las siguientes enfermedades: [Síndrome de hiperactividad]	0.029055
Naturaleza Colegio	0.028917
Categoría Colegio	0.028814
Posee Necesidades Educativas Especiales (NEE) como: [Ninguna]	0.028496

12.1.3 Árbol de decisión

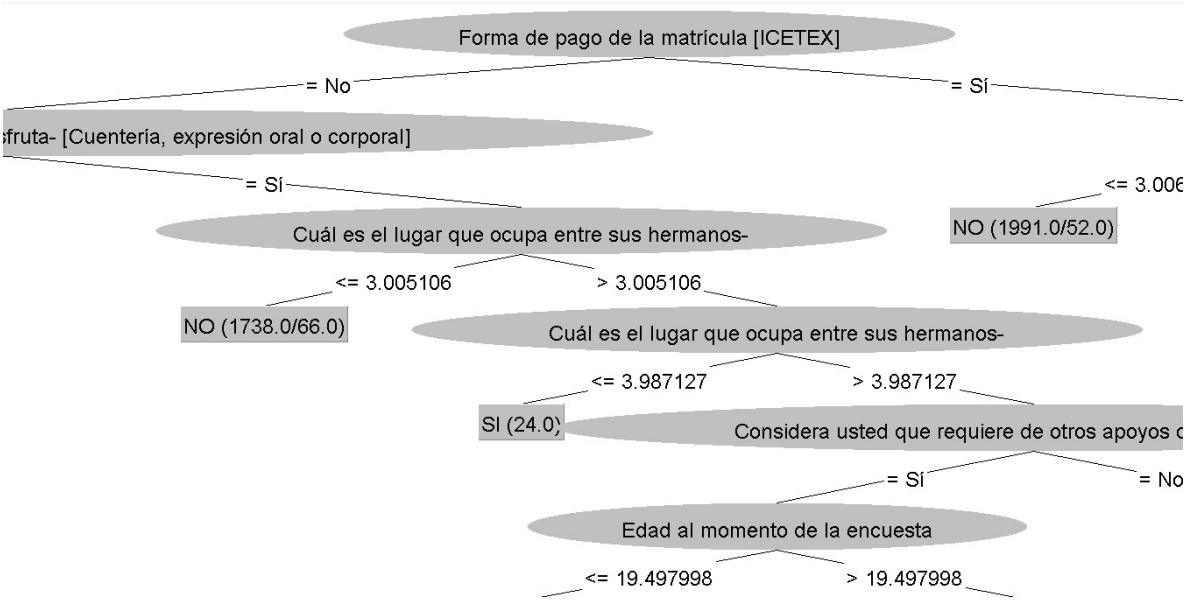
Tomando el set de datos balanceados, se aplica el método de árbol de decisión J48 con los siguientes parámetros:

Ilustración 12. Parámetros árbol de decisión para selección de factores



Arrojando el árbol de clasificación de la ilustración 13.

Ilustración 13. Árbol de decisión para selección de factores



Este árbol arroja las siguientes medidas de calidad con validación cruzada, indicando una exactitud de instancias correctamente clasificadas del 96.4%

Ilustración 14. Medidas de calidad del árbol de decisión para selección de factores

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       19629               96.4238 %
Incorrectly Classified Instances       728                3.5762 %
Kappa statistic                          0.9285
Mean absolute error                      0.0625
Root mean squared error                  0.1829
Relative absolute error                 12.5041 %
Root relative squared error             36.5847 %
Total Number of Instances            20357

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,969    0,041    0,960      0,969    0,964      0,929    0,975     0,956     NO
                 0,959    0,031    0,969      0,959    0,964      0,929    0,975     0,978     SI
Weighted Avg.    0,964    0,036    0,964      0,964    0,964      0,929    0,975     0,967
```

En la matriz de confusión de la tabla 23 se puede observar que sólo se obtienen 728 errores en la predicción y 19.629 aciertos.

Tabla 23. Matriz de confusión del árbol de decisión para selección de factores

		Predicción	
		Clasificados como NO	Clasificados como SI
Real	NO	**9864**	315
	SI	413	**9765**

Después de analizar el árbol resultante, las 30 variables más influyentes en la deserción, según este método, son aquellas variables que se encuentran en la parte superior del árbol. En la tabla 24 se listan las variables en el mismo orden de aparición en el árbol.

Tabla 24. Resultado método árbol de decisión

Variable
Forma de pago de la matrícula [ICETEX]
Qué tipo de espectáculos o actividades disfruta- [Cuentería, expresión oral o corporal]
Cuántos hermanos tiene-
Ha recibido alguno o todos de los siguientes apoyos: [Ninguno]
Cuál es el lugar que ocupa entre sus hermanos
Promedio Notas
Qué tipo de actividad realiza con mayor frecuencia en su tiempo libre- [Jugar video Juegos]
Considera usted que requiere de otros apoyos como: [Ninguno]
Pertenece a alguna etnia o resguardo indígena
Padece alguna de las siguientes enfermedades: [Migraña]
Cuáles de los siguientes métodos de estudio utiliza usualmente- [Subrayar]
Usted considera que requiere ayuda financiera para el pago de la matrícula

Variable
Forma de pago de la matrícula [Entidad Bancaria]
Ocupación del padre
Considera que tiene dificultad en alguno de los siguientes aspectos- [Las actividades artísticas y manuales]
Quién toma las decisiones en su familia- [Usted]
Cuáles de los siguientes métodos de estudio utiliza usualmente- [Repasar después de cada clase]
Quién toma las decisiones en su familia- [Mamá]
Participa en grupos o realiza actividades relacionadas con: [Sociales]
Porcentaje Aprobados / Intentados
Padece alguna de las siguientes enfermedades: [Trastorno de ansiedad / Depresión / Trastorno Bipolar]
Cuáles de los siguientes métodos de estudio utiliza usualmente- [Memorización mecánica (sin comprender lo que estudio)]
Posee Necesidades Educativas Especiales (NEE) como: [Baja Visión diagnosticada]
Tipo de vinculación al Sistema General de Seguridad Social en Salud = Régimen contributivo (EPS)
Actualmente en su familia presentan alguna de las siguientes situaciones- [Separación de los padres]
Zona de la residencia
Edad al momento de la encuesta
Considera usted que requiere de otros apoyos como: [Beca]
En su barrio se presentan algunas de estas situaciones o fenómenos- [Hurto]
Vive con otros miembros de la familia

12.1.4 Resultado selección de factores

Se realiza una tabla de frecuencia de acuerdo a la ocurrencia en cada de los tres métodos aplicados anteriormente.

Tabla 25. Selección de factores - frecuencia por método

Variable	Ganancia	Correlación	Árbol	Frecuencia
Cuántos hermanos tiene	X	X	X	3
Edad al momento de la encuesta	X	X	X	3
Porcentaje Aprobados / Intentados	X	X	X	3
Promedio Notas	X	X	X	3
Tipo de vinculación al Sistema General de Seguridad Social en Salud	X	X	X	3
Actualmente trabaja-	X	X		2
Al momento de iniciar sus estudios en la universidad se trasladó de Ciudad o Municipio-	X	X		2
Categoría Colegio	X	X		2
Considera usted que requiere de otros apoyos como: [Crédito]	X	X		2
Departamento o estado de nacimiento	X	X		2
Durante su educación media presentó alguna de las siguientes situaciones- [Dificultades académicas]	X	X		2
Durante su educación media presentó alguna de las siguientes situaciones- [Ninguno]	X	X		2
Estado Académico	X	X		2
Forma de pago de la matrícula [Beneficiario de beca]	X	X		2
Motivo por el cual eligió la carrera [Interés propio]	X	X		2
Motivo por el cual eligió la carrera [Presión familiar]	X	X		2
Nro Semestres Programa	X	X		2
Ocupación del padre	X		X	2
Qué tipo de actividad realiza con mayor frecuencia en su tiempo libre- [Compartir con los amigos]	X	X		2
Qué tipo de actividad realiza con mayor frecuencia en su tiempo libre- [Dormir]	X	X		2
Rezagado	X	X		2
Ubicación Semestral	X	X		2
Actualmente en su familia presentan alguna de las siguientes situaciones- [Separación de los padres]			X	1
Actualmente tiene personas a cargo-		X		1
Cómo se siente con los métodos de estudio que utiliza usualmente:	X			1
Considera que tiene dificultad en alguno de los siguientes aspectos-			X	1

Variable	Ganancia	Correlación	Árbol	Frecuencia
[Las actividades artísticas y manuales]				
Considera usted que requiere de otros apoyos como: [Beca]			X	1
Considera usted que requiere de otros apoyos como: [Ninguno]			X	1
Cuál es el lugar que ocupa entre sus hermanos			X	1
Cuáles de los siguientes métodos de estudio utiliza usualmente- [Memorización mecánica (sin comprender lo que estudio)]			X	1
Cuáles de los siguientes métodos de estudio utiliza usualmente- [Repasar después de cada clase]			X	1
Cuáles de los siguientes métodos de estudio utiliza usualmente- [Subrayar]			X	1
Durante su educación media presentó alguna de las siguientes situaciones- [Pérdida de años]		X		1
En su barrio se presentan algunas de estas situaciones o fenómenos- [Hurto]			X	1
Estado civil	X			1
Forma de pago de la matrícula [De contado (Efectivo)]		X		1
Forma de pago de la matrícula [Entidad Bancaria]			X	1
Forma de pago de la matrícula [Fondo EPM]		X		1
Forma de pago de la matrícula [ICETEX]			X	1
Ha recibido alguno o todos de los siguientes apoyos: [Ninguno]			X	1
Motivo por el cual eligió la carrera [Presión social]		X		1
Municipio o ciudad de nacimiento	X			1
Naturaleza Colegio		X		1
Nivel educativo de la madre	X			1
Padece alguna de las siguientes enfermedades: [Migraña]			X	1
Padece alguna de las siguientes enfermedades: [Síndrome de hiperactividad]		X		1
Padece alguna de las siguientes enfermedades: [Trastorno de ansiedad / Depresión / Trastorno Bipolar]			X	1

Variable	Ganancia	Correlación	Árbol	Frecuencia
País de nacimiento	X			1
Participa en grupos o realiza actividades relacionadas con: [Sociales]			X	1
Pertenece a alguna etnia o resguardo indígena			X	1
Posee Necesidades Educativas Especiales (NEE) como: [Baja Visión diagnosticada]			X	1
Posee Necesidades Educativas Especiales (NEE) como: [Déficit de atención con Hiperactividad (TDAH)]		X		1
Posee Necesidades Educativas Especiales (NEE) como: [Ninguna]		X		1
Programa Académico	X			1
Promedio de ingresos del padre	X			1
Qué tipo de actividad realiza con mayor frecuencia en su tiempo libre- [Jugar video Juegos]			X	1
Qué tipo de espectáculos o actividades disfruta- [Cuentería, expresión oral o corporal]			X	1
Quién es su acudiente-	X			1
Quién toma las decisiones en su familia- [Mamá]			X	1
Quién toma las decisiones en su familia- [Usted]			X	1
Usted considera que requiere ayuda financiera para el pago de la matrícula			X	1
Vive con otros miembros de la familia			X	1
Zona de la residencia			X	1

Se toman las variables que aparezcan en dos de los tres métodos aplicados. Siendo así, se eligen 22 variables que se convertirán en los factores seleccionados, ellos son:

Tabla 26. Factores seleccionados

Factor
Cuántos hermanos tiene
Edad al momento de la encuesta
Porcentaje Aprobados / Intentados

Factor
Promedio Notas
Tipo de vinculación al Sistema General de Seguridad Social en Salud
Actualmente trabaja-
Al momento de iniciar sus estudios en la universidad se trasladó de Ciudad o Municipio-
Categoría Colegio
Considera usted que requiere de otros apoyos como: [Crédito]
Departamento o estado de nacimiento
Durante su educación media presentó alguna de las siguientes situaciones- [Dificultades académicas]
Durante su educación media presentó alguna de las siguientes situaciones- [Ninguno]
Estado Académico
Forma de pago de la matrícula [Beneficiario de beca]
Motivo por el cual eligió la carrera [Interés propio]
Motivo por el cual eligió la carrera [Presión familiar]
Nro Semestres Programa
Ocupación del padre
Qué tipo de actividad realiza con mayor frecuencia en su tiempo libre- [Compartir con los amigos]
Qué tipo de actividad realiza con mayor frecuencia en su tiempo libre- [Dormir]
Rezagado
Ubicación Semestral

12.2. MODELO 2: Clúster de tipos de estudiantes que desertan.

Objetivo: Encontrar las características básicas de los tipos de estudiantes que desertan.

Método: K-means.

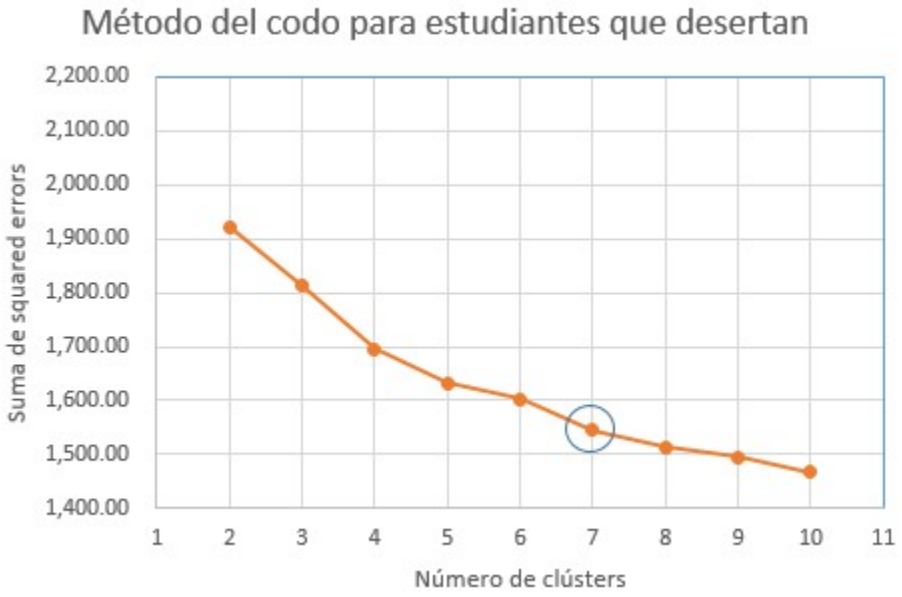
Primero que todo, se utiliza el método del codo para realizar evaluación del número de clústeres óptimo. Se aplica el método K-means con diferentes números de clústeres evaluando en cada uno de ellos la suma de squared errors, lo cual nos indica el nivel de cohesión entre ellos. El resultado se observa en la siguiente tabla.

Tabla 27. Resultado K-means por clúster para estudiantes que desertan

Número de clústeres	Suma de squared errors
2	1922.73
3	1812.38
4	1696.12
5	1631.9
6	1601.75
7	1544.93
8	1512.83
9	1494.24
10	1466.27

Graficando estos resultados se observa un punto de inflexión con 7 clústeres.

Ilustración 15. Método del codo - estudiantes que desertan



Tomando un total de 7 clusters, se aplica el método K-means y por medio de los centroides del resultado se realiza el perfilamiento de los estudiantes. Los perfiles encontrados son:

Clúster 1: Incomprendidos

Estudiantes con un promedio de edad de 19 años, sin necesidad de trasladarse de residencia para ir a estudiar, cuyo padre trabaja como empleado, entre uno y dos hermanos, sin necesidad de crédito para pagar la matrícula, sin dificultades académicas en su educación media, no les gusta dormir en su tiempo libre, no les gusta compartir con los amigos, provenientes de muy buen colegio en términos académicos, en los dos primeros semestres de su carrera y con un muy bajo rendimiento académico en la universidad.

Clúster 2: Dormilones

Estudiantes con un promedio de edad de 18 años, sin necesidad de trasladarse de residencia para ir a estudiar, cuyo padre trabaja como independiente, entre uno y

dos hermanos, sin necesidad de crédito para pagar la matrícula, sin dificultades académicas en su educación media, les gusta dormir en su tiempo libre, les gusta compartir con los amigos, provenientes de muy buen colegio en términos académicos, en los dos primeros semestres de su carrera y con buen rendimiento académico en la universidad.

Clúster 3: Independientes

Estudiantes con un promedio de edad de 26 años, sin necesidad de trasladarse de residencia para ir a estudiar, cuyo padre ya ha fallecido o no viven con él, entre uno y dos hermanos, con necesidad de crédito para pagar la matrícula, sin dificultades académicas en su educación media, les gusta dormir en su tiempo libre, les gusta compartir con los amigos, provenientes de buen colegio en términos académicos, en los cinco primeros semestres de su carrera, con aceptable rendimiento académico en la universidad y rezagados en mínimo dos semestres en su carrera.

Clúster 4: Problemas económicos

Estudiantes con un promedio de edad de 22 años, sin necesidad de trasladarse de residencia para ir a estudiar, cuyo padre trabaja como independiente, entre uno y dos hermanos, con necesidad de crédito para pagar la matrícula, sin dificultades académicas en su educación media, no les gusta dormir en su tiempo libre, les gusta compartir con los amigos, provenientes de muy buen colegio en términos académicos, en los primero cuatro semestres de su carrera, con buen rendimiento académico en la universidad y rezagados en mínimo dos semestres.

Clúster 5: Relajados

Estudiantes con un promedio de edad de 23 años, sin necesidad de trasladarse de residencia para ir a estudiar, cuyo padre trabaja como independiente, con un

hermano, sin necesidad de crédito para pagar la matrícula, con dificultades académicas en su educación media, les gusta dormir en su tiempo libre, les gusta compartir con los amigos, provenientes de muy buen colegio en términos académicos, en los primeros cuatro semestres de su carrera, con bajo rendimiento académico en la universidad y rezagados en mínimo dos semestres.

Clúster 6: Pilos no tan pilos

Estudiantes con un promedio de edad de 21 años, con necesidad de trasladarse de residencia para ir a estudiar, cuyo padre trabaja como independiente, entre uno y dos hermanos, sin necesidad de crédito para pagar la matrícula, sin dificultades académicas en su educación media, no les gusta dormir en su tiempo libre, no les gusta compartir con los amigos, provenientes de un colegio con rendimiento medio en términos académicos, en los primeros dos semestres de su carrera y con bajo rendimiento académico en la universidad.

Clúster 7: BomBril

Estudiantes con un promedio de edad de 27 años, sin necesidad de trasladarse de residencia para ir a estudiar, cuyo padre trabaja como empleado, entre uno y dos hermanos, sin necesidad de crédito para pagar la matrícula, sin dificultades académicas en su educación media, no les gusta dormir en su tiempo libre, no les gusta compartir con los amigos, provenientes de un colegio con rendimiento alto en términos académicos, en los últimos semestres de su carrera, con rendimiento académico aceptable en la universidad y rezagados en mínimo dos semestres.

12.3. MODELO 3: Clúster de tipos de estudiantes que NO desertan.

Objetivo: Encontrar las características básicas de los tipos de estudiantes que NO desertan

Método: K-means

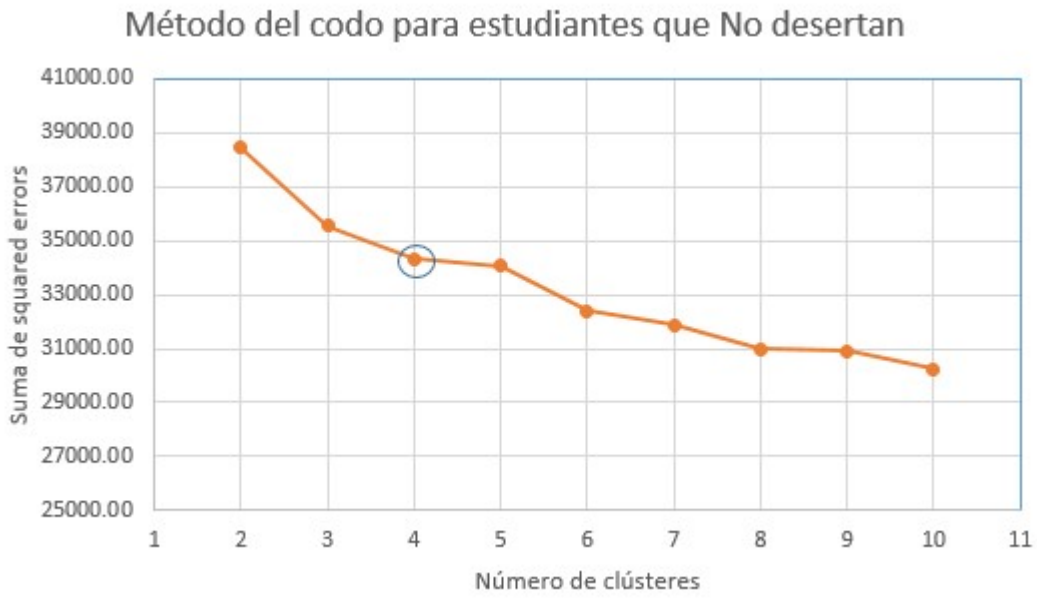
Al igual que en el método anterior, se utiliza el método del codo para encontrar la cantidad de clústeres óptimo para los estudiantes que no desertan.

Tabla 28. Resultado K-means por clúster para estudiantes que no desertan

Número de clústeres	Suma de squared errors
2	38437.19
3	35572.01
4	34331.05
5	34056.76
6	32389.42
7	31883.87
8	30999.60
9	30913.23
10	30227.33

Graficando estos resultados se observa un punto de inflexión con 4 clústeres.

Ilustración 16. Método del codo - estudiantes que no desertan



Al crear 4 clusters con el método k-means, los perfiles encontrados son:

Clúster 1: Persistentes

Estudiantes con un promedio de edad de 22 años, sin necesidad de trasladarse de residencia para ir a estudiar, cuyo padre trabaja como independiente, con un hermano, sin necesidad de crédito para pagar la matrícula, sin dificultades académicas en su educación media, no les gusta dormir en su tiempo libre, les gusta compartir con los amigos, provenientes de muy buen colegio en términos académicos, con avance de seis semestres en su carrera, con muy buen rendimiento académico en la universidad y rezagados en mínimo dos semestres.

Clúster 2: Juiciosos semestres intermedios

Estudiantes con un promedio de edad de 20 años, sin necesidad de trasladarse de residencia para ir a estudiar, cuyo padre trabaja como empleado, con un hermano, sin necesidad de crédito para pagar la matrícula, sin dificultades académicas en su educación media, no les gusta dormir en su tiempo libre, no les gusta compartir con los amigos, provenientes de muy buen colegio en términos académicos, con avance de cuatro semestres en su carrera, con muy buen rendimiento académico en la universidad y sin rezago en su avance semestral.

Clúster 3: Juiciosos primeros semestres

Estudiantes con un promedio de edad de 19 años, sin necesidad de trasladarse de residencia para ir a estudiar, cuyo padre trabaja como independiente, entre uno y dos hermanos, sin necesidad de crédito para pagar la matrícula, sin dificultades académicas en su educación media, no les gusta dormir en su tiempo libre, no les

gusta compartir con los amigos, provenientes de buen colegio en términos académicos, en sus primeros dos semestres de su carrera y con muy buen rendimiento académico en la universidad.

Clúster 4: Avanzados en la carrera

Estudiantes con un promedio de edad de 20 años, sin necesidad de trasladarse de residencia para ir a estudiar, cuyo padre trabaja como independiente, con un hermano, sin necesidad de crédito para pagar la matrícula, sin dificultades académicas en su educación media, no les gusta dormir en su tiempo libre, les gusta compartir con los amigos, provenientes de muy buen colegio en términos académicos, con avance de cinco semestres en su carrera, con muy buen rendimiento académico en la universidad y sin rezago en su avance semestral.

12.4. MODELO 4: Reglas de asociación

Objetivo: encontrar relaciones entre los datos en relación a la variable objetivo.

Método: A-priori

Se utilizan los datos no balanceados y con las 22 variables seleccionadas en el primer modelo desarrollado de selección de factores. Antes de aplicar el método, se hace necesario categorizar las variables numéricas. Para ello se observa la descripción estadística de los datos y se aplican los siguientes grupos para cada variable:

Ilustración 17. Categorización de variables para aplicar método A-priori

Variable	Categorización
Edad al momento de la encuesta	• 14-17 • 18-22 • 23-26 • 26+

Variable	Categorización
Cuántos hermanos tiene?	• sin hermanos • un hermano • dos hermanos • tres o más
Número de semestre de programa	• 6 o menos • 7-9 • 10 • 10+
Ubicación semestral	• primero • segundo • tercero • cuarto • quinto • sexto • séptimo • octavo • noveno • décimo • onceavo o más
% créditos aprobados / intentados	• 0%-50% • 51%-80% • 81%-90% • 91%+
Promedio de notas	• 0 - 2.99 • 3 - 3.30 • 3.31 - 3.70 • 3.71 - 4.30 • 4.31+

El método arroja diferentes reglas de asociación, se configuró para generar 1000 reglas diferentes y se tomaron las primeras 50 que tienen un nivel de confianza de entre 0.98 y 0.99. Todas estas reglas están asociadas con los estudiantes que no desertan. Algunas de ellas, de esas primeras 50 son:

- Motivo por el cual eligió la carrera [Interés propio]=Sí Porcentaje Aprobados / Intentados=91%+ 4909 ==> Desertor=NO 4844 conf:(0.99)

Interpretación: Los estudiantes que eligen la carrera por interés propio y que tienen un porcentaje de créditos aprobados/intentados superior al 91% normalmente no desertan.

- Motivo por el cual eligió la carrera [Interés propio]=Sí Motivo por el cual eligió la carrera [Presión familiar]=No Porcentaje Aprobados / Intentados=91%+ 4882 ==> Desertor=NO 4817 conf:(0.99)

Interpretación: Los estudiantes que eligen la carrera por interés propio y que no tiene presión familiar para esa misma elección y tienen un porcentaje de créditos aprobados/intentados superior al 91% normalmente no desertan.

- Motivo por el cual eligió la carrera [Presión familiar]=No Estado Académico=Estado académico normal Porcentaje Aprobados / Intentados=91%+ 5110 ==> Desertor=NO 5037 conf:(0.99)

Interpretación: Los estudiantes que no tuvieron presión familiar para elegir la carrera, sin novedades en su estado académico y con porcentaje de créditos aprobados/intentados superior al 91% normalmente no desertan.

- Estado Académico=Estado académico normal Porcentaje Aprobados / Intentados=91%+ 5153 ==> Desertor=NO 5078 conf:(0.99)

Interpretación: Los estudiantes sin novedades en su estado académico y con porcentaje de créditos aprobados/intentados 91% normalmente no desertan.

- Durante su educación media presentó alguna de las siguientes situaciones-[Dificultades académicas]=No Estado Académico=Estado académico normal Porcentaje Aprobados / Intentados=91%+ 4859 ==> Desertor=NO 4787 conf:(0.99)

Interpretación: Los estudiantes que no presentaron dificultades académicas en el colegio, sin novedades en su estado académico y con porcentaje de créditos aprobados/intentados superior al 91% normalmente no desertan.

- Al momento de iniciar sus estudios en la universidad se trasladó de Ciudad o Municipio-=No Actualmente trabaja-=No Considera usted que requiere de otros apoyos como: [Crédito]=No Durante su educación media presentó alguna de las siguientes situaciones- [Dificultades académicas]=No Motivo por el cual eligió la carrera [Presión familiar]=No Estado Académico=Estado académico normal 4917 ==> Desertor=NO 4837 conf:(0.98)

Interpretación: Los estudiantes que no se trasladaron de domicilio, que no necesitan de crédito para pagar su matrícula, que no presentaron dificultades académicas, sin presión familiar para elegir la carrera y sin dificultades en rendimiento académico normalmente no desertan.

- Durante su educación media presentó alguna de las siguientes situaciones-[Dificultades académicas]=No Estado Académico=Estado académico

normal Promedio Notas=3.7 - 4.299 4852 ==> Desertor=NO 4772 conf:(0.98)

Interpretación: Los estudiantes que no tuvieron dificultades académicas en el colegio, sin dificultades académicas en la universidad y con un promedio de 3.7, normalmente no desertan.

Generando un resumen de aparición de esas 50 primeras reglas y ordenándolas por frecuencia de aparición se obtiene la siguiente tabla:

Tabla 29. Resumen reglas de asociación estudiantes que no desertan

Regla	Frecuencia
Estado Académico=Estado académico normal	44
Al momento de iniciar sus estudios en la universidad se trasladó de Ciudad o Municipio-=No	31
Actualmente trabaja-=No	30
Motivo por el cual eligió la carrera [Presión familiar]=No	26
Departamento o estado de nacimiento=Antioquia	24
Motivo por el cual eligió la carrera [Interés propio]=Sí	20
Considera usted que requiere de otros apoyos como: [Crédito]=No	20
Durante su educación media presentó alguna de las siguientes situaciones- [Dificultades académicas]=No	19
Tipo de vinculación al Sistema General de Seguridad Social en Salud=Régimen contributivo (EPS)	13
Porcentaje Aprobados / Intentados=91%+	9
Promedio Notas=3.7 - 4.299	3

Estas reglas visualizan algunas características principales de los estudiantes que no presentan deserción, dado que no fue posible obtener reglas asociadas a los

estudiantes que desertan ya que son muy pocos en la muestra y para este tipo de análisis descriptivo NO se deben balancear los datos.

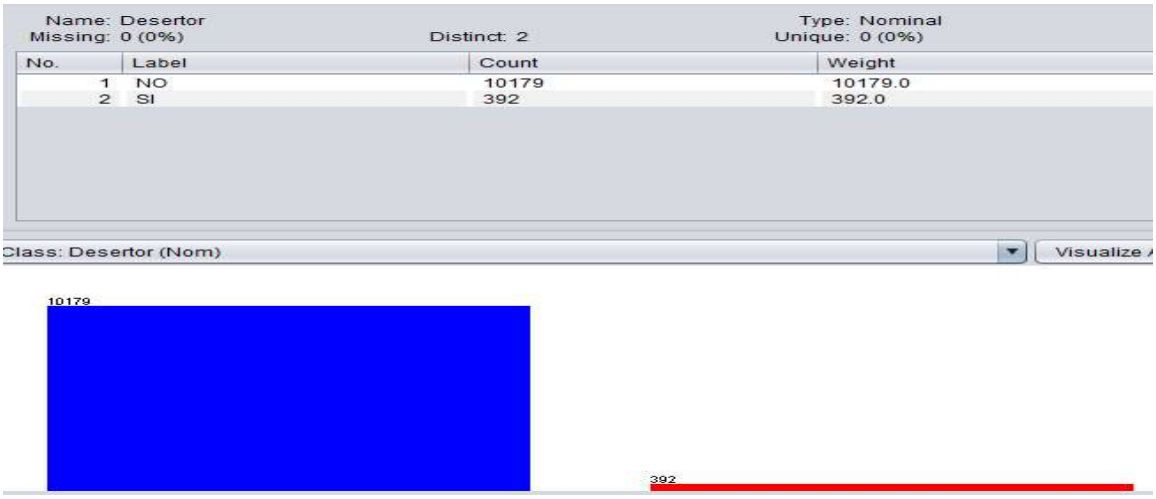
12.5. MODELO 5: Predicción de la deserción

Objetivo: aplicar diversas técnicas de predicción sobre el conjunto de datos previo a la selección de factores y con los factores seleccionados.

Métodos: árbol J48, árbol RandomForest, NaiveBayes, SMO (soporte vectorial)

Ya se dispone de un primer set de datos que contiene 200 variables y está balanceado, ahora se hace necesario crear un segundo set de datos balanceados correspondiente a los 22 factores seleccionados. Para este segundo set se cuenta con lo siguiente:

Ilustración 18. Datos sin balancear factores seleccionados



Al balancearlos con el filtro SMOTE se obtiene:

Ilustración 19. Datos balanceados factores seleccionados

Ahora ya se cuenta con dos sets de datos balanceados, el que se obtuvo después de la preparación de datos que contiene 200 variables; y este que se acaba de balancear con los 22 factores seleccionados. A continuación, se aplicarán los modelos predictivos.

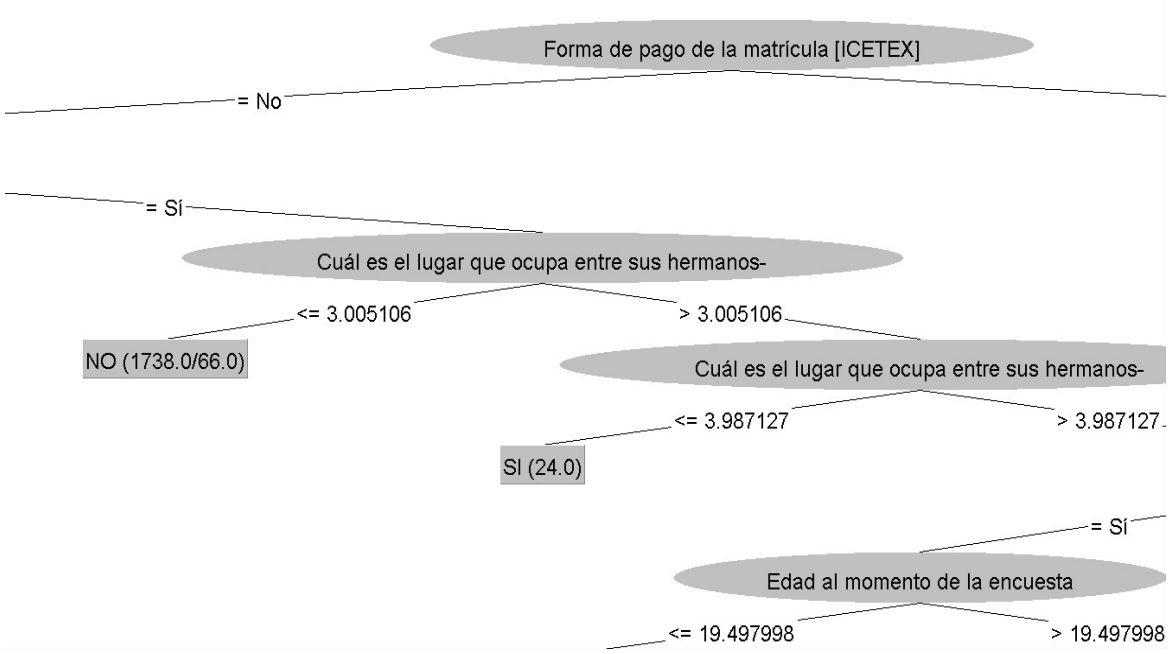
12.5.1. Predicción con todas las variables

Para realizar la predicción se aplican 4 métodos que son comparados para seleccionar el mejor, los métodos aplicados son: un árbol de decisión, un bosque de árboles, un método bayesiano y una máquina de soporte vectorial.

12.5.1.1. Árbol de decisión

Se aplica en weka el árbol de decisión J48 con las 200 variables. Al aplicar este método se obtiene un árbol extenso. Una muestra de lo que se encuentra en la copa del árbol es lo siguiente:

Ilustración 20. Árbol de decisión con todas las variables



Los resultados de evaluación en la calidad del arbol se muestran a continuación:

Ilustración 21. Medidas de calidad del árbol J48 con todas las variables

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       19629               96.4238 %
Incorrectly Classified Instances       728                3.5762 %
Kappa statistic                          0.9285
Mean absolute error                      0.0625
Root mean squared error                  0.1829
Relative absolute error                 12.5041 %
Root relative squared error             36.5847 %
Total Number of Instances            20357

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,969     0,041     0,960      0,969     0,964     0,929     0,975     0,956     NO
               0,959     0,031     0,969      0,959     0,964     0,929     0,975     0,978     SI
Weighted Avg.  0,964     0,036     0,964      0,964     0,964     0,929     0,975     0,967
```

Se obtiene la matriz de confusión:

Tabla 30. Matriz de confusión árbol J48 con todas las variables

		Predicción	
		Clasificados como NO	Clasificados como SI
Real	NO	**9864**	315
	SI	413	**9765**

12.5.1.2. Bosque de árboles

Se crea un bosque de árboles con el método RandomForest de weka. Con este método se obtienen los siguientes resultados en la calidad del modelo:

Ilustración 22. Medidas de calidad del árbol RandomForest con todas las variables

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       19966               98.0793 %
Incorrectly Classified Instances       391                1.9207 %
Kappa statistic                          0.9616
Mean absolute error                      0.0644
Root mean squared error                  0.1382
Relative absolute error                 12.8836 %
Root relative squared error             27.6438 %
Total Number of Instances            20357

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               1,000    0,038    0,963      1,000    0,981      0,962    0,993     0,989     NO
               0,962    0,000    1,000      0,962    0,980      0,962    0,993     0,995     SI
Weighted Avg.  0,981    0,019    0,982      0,981    0,981      0,962    0,993     0,992
```

Y la siguiente matriz de confusión:

Tabla 31. Matriz de confusión árbol RandomForest con todas las variables

		Predicción	
		Clasificados como NO	Clasificados como SI
Real	NO	**10179**	0
	SI	391	**9787**

12.5.1.3. Método bayesiano

Se aplica el método NaiveBayes de weka, con este método se obtienen los siguientes resultados:

Ilustración 23. Medidas de calidad NaiveBayes con todas las variables

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       19550               96.0358 %
Incorrectly Classified Instances       807                3.9642 %
Kappa statistic                          0.9207
Mean absolute error                      0.0408
Root mean squared error                  0.1888
Relative absolute error                  8.161  %
Root relative squared error             37.7658 %
Total Number of Instances            20357

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,954     0,034     0,966       0,954   0,960      0,921    0,982     0,970     NO
               0,966     0,046     0,955       0,966   0,961      0,921    0,982     0,985     SI
Weighted Avg.  0,960     0,040     0,960       0,960   0,960      0,921    0,982     0,978
```

Y la siguiente matriz de confusión:

Tabla 32. Matriz de confusión NaiveBayes con todas las variables

		Predicción	
		Clasificados como NO	Clasificados como SI
Real	NO	**9715**	464
	SI	343	**9835**

12.5.1.4. Máquina de soporte vectorial

Se aplica el método SMO de weka, con este método se obtienen los siguientes resultados:

Ilustración 24. Medidas de calidad SMO con todas las variables

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       19871               97.6126 %
Incorrectly Classified Instances       486                2.3874 %
Kappa statistic                          0.9523
Mean absolute error                      0.0239
Root mean squared error                  0.1545
Relative absolute error                  4.7748 %
Root relative squared error             30.9023 %
Total Number of Instances            20357

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,986     0,034     0,967       0,986   0,976      0,952    0,976     0,960     NO
               0,966     0,014     0,986       0,966   0,976      0,952    0,976     0,969     SI
Weighted Avg.  0,976     0,024     0,976       0,976   0,976      0,952    0,976     0,965
```

Y la siguiente matriz de confusión:

Tabla 33. Matriz de confusión SMO con todas las variables

		Predicción	
		Clasificados como NO	Clasificados como SI
Real	NO	**10035**	144
	SI	342	**9836**

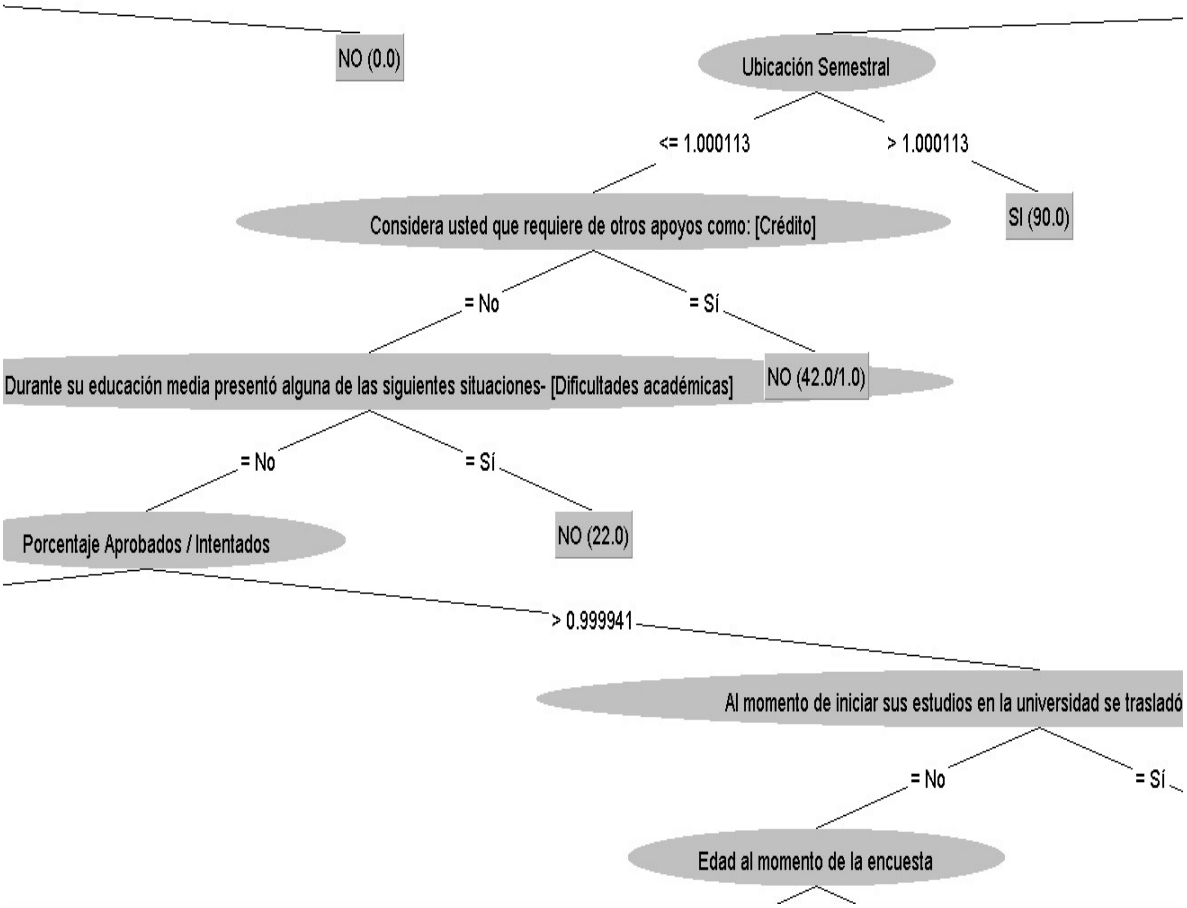
12.5.2. Predicción con los factores seleccionados

Para realizar la predicción se aplican 4 métodos que son comparados para seleccionar el mejor, los métodos aplicados a las 22 variables previamente seleccionaas son: un árbol de decisión, un bosque de árboles, un método bayesiano y una máquina de soporte vectorial.

12.5.2.1. Árbol de decisión

Al aplicar el método árbol J48 de weka sobre la sábana con los 22 factores seleccionados se obtiene un árbol extenso. Una pequeña sección del árbol se muestra a continuación:

Ilustración 25. Árbol de decisión con factores seleccionados



Los resultados de la calidad del árbol se muestran a continuación:

Ilustración 26. Medidas de calidad del árbol J48 con factores seleccionados

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       19180               94.2136 %
Incorrectly Classified Instances      1178                5.7864 %
Kappa statistic                          0.8843
Mean absolute error                      0.0805
Root mean squared error                  0.2287
Relative absolute error                 16.0937 %
Root relative squared error             45.7387 %
Total Number of Instances            20358

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,933     0,049     0,950       0,933   0,942      0,884    0,960     0,945     NO
               0,951     0,067     0,934       0,951   0,943      0,884    0,960     0,945     SI
Weighted Avg.  0,942     0,058     0,942       0,942   0,942      0,884    0,960     0,945
```

Se obtiene la matriz de confusión:

Tabla 34. Matriz de confusión árbol J48 con factores seleccionados

		Predicción	
		Clasificados como NO	Clasificados como SI
Real	NO	**9497**	682
	SI	496	**9683**

12.5.2.2. Bosque de árboles

Con el método RandomForest de weka se obtienen los siguientes resultados:

Ilustración 27. Medidas de calidad del árbol RandomForest con factores seleccionados

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       19733               96.93   %
Incorrectly Classified Instances       625                3.07   %
Kappa statistic                          0.9386
Mean absolute error                      0.0849
Root mean squared error                  0.1691
Relative absolute error                 16.9795 %
Root relative squared error             33.8189 %
Total Number of Instances            20358

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,971     0,033     0,967       0,971   0,969      0,939    0,992     0,990     NO
               0,967     0,029     0,971       0,967   0,969      0,939    0,992     0,993     SI
Weighted Avg.  0,969     0,031     0,969       0,969   0,969      0,939    0,992     0,991
```

Y la siguiente matriz de confusión:

Tabla 35. Matriz de confusión árbol RandomForest con factores seleccionados

		Predicción	
		Clasificados como NO	Clasificados como SI
Real	NO	**9888**	291
	SI	334	**9845**

12.5.2.3. Método bayesiano

Con el método NaiveBayes se obtienen los siguientes resultados:

Ilustración 28. Medidas de calidad de NaiveBayes con factores seleccionados

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       16023               78.7062 %
Incorrectly Classified Instances      4335               21.2938 %
Kappa statistic                          0.5741
Mean absolute error                      0.2341
Root mean squared error                  0.3852
Relative absolute error                 46.8131 %
Root relative squared error             77.0432 %
Total Number of Instances            20358

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,874     0,300     0,745       0,874   0,804      0,583    0,893     0,902     NO
               0,700     0,126     0,848       0,700   0,767      0,583    0,893     0,879     SI
Weighted Avg.  0,787     0,213     0,796       0,787   0,785      0,583    0,893     0,890
```

Y su respectiva matriz de confusión:

Tabla 36. Matriz de confusión NaiveBayes con factores seleccionados

		Predicción	
		Clasificados como NO	Clasificados como SI
Real	NO	**8897**	1282
	SI	3053	**7126**

12.5.2.4. Máquina de soporte vectorial

Con el método SMO de weka se obtienen los siguientes resultados:

Ilustración 29. Medidas de calidad de SMO con factores seleccionados

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       17210               84.5368 %
Incorrectly Classified Instances      3148               15.4632 %
Kappa statistic                          0.6907
Mean absolute error                      0.1546
Root mean squared error                  0.3932
Relative absolute error                 30.9264 %
Root relative squared error             78.6466 %
Total Number of Instances            20358

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,837    0,146    0,851      0,837    0,844      0,691    0,845     0,794     NO
               0,854    0,163    0,840      0,854    0,847      0,691    0,845     0,790     SI
Weighted Avg.  0,845    0,155    0,845      0,845    0,845      0,691    0,845     0,792
```

Y su respectiva matriz de confusión:

Tabla 37. Matriz de confusión SMO con factores seleccionados

		Predicción	
		Clasificados como NO	Clasificados como SI
Real	NO	**8518**	1661
	SI	1487	**8692**

13. DESPLIEGUE

13.1 Análisis de resultados de los modelos

Se revisan en conjunto los resultados entregados por los diferentes modelos aplicados en el capítulo anterior.

El modelo para la selección de factores delimita el conjunto de variables a analizar para los modelos de perfilamiento (clústeres), reglas de asociación y especialmente para la generación de los modelos predictivos. Este acotamiento de variables facilita además la puesta en marcha de todo el proceso de modelamiento en cuanto a tiempos de procesamiento.

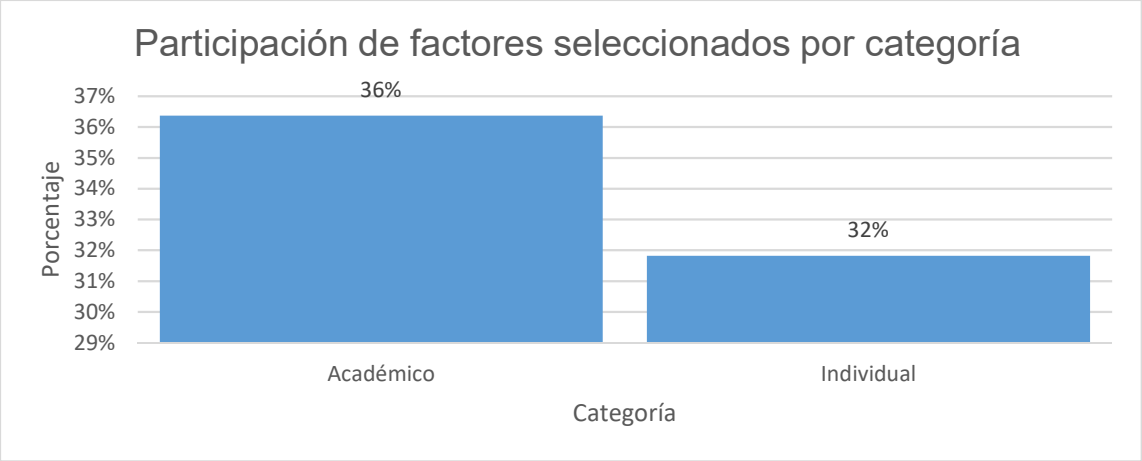
Tabla 38. Factores más relacionados con la Deserción

Factores más relacionados con la Deserción
Cuántos hermanos tiene
Edad al momento de la encuesta
Porcentaje Aprobados / Intentados
Promedio Notas
Tipo de vinculación al Sistema General de Seguridad Social en Salud
Actualmente trabaja-
Al momento de iniciar sus estudios en la universidad se trasladó de Ciudad o Municipio-
Categoría Colegio
Considera usted que requiere de otros apoyos como: [Crédito]
Departamento o estado de nacimiento
Durante su educación media presentó alguna de las siguientes situaciones- [Dificultades académicas]
Durante su educación media presentó alguna de las siguientes situaciones- [Ninguno]
Estado Académico
Forma de pago de la matrícula [Beneficiario de beca]
Motivo por el cual eligió la carrera [Interés propio]
Motivo por el cual eligió la carrera [Presión familiar]

Factores más relacionados con la Deserción
Nro Semestres Programa
Ocupación del padre
Qué tipo de actividad realiza con mayor frecuencia en su tiempo libre- [Compartir con los amigos]
Qué tipo de actividad realiza con mayor frecuencia en su tiempo libre- [Dormir]
Rezagado
Ubicación Semestral

Se puede decir además que dichos factores, en su mayoría, son coherentes con el marco teórico respecto a las variables determinantes para la deserción. Se obtienen características de tipo individual, desempeño académico, apoyos institucionales y condiciones socioeconómicas. Realizando una clasificación de cada una de ellas se obtiene la ilustración número 30, donde se muestra que por ejemplo la mayoría de los factores seleccionados corresponden al componente académico, seguido por características individuales del estudiante, condiciones socioeconómicas y por último aparece los apoyos institucionales.

Ilustración 30. Categorías de variables principales de los factores seleccionados



Aplicando estos factores en el perfilamiento de los estudiantes, tanto los que desertan como los que no, se observa un agrupamiento lógico que permite describir

algunas características básicas tanto de forma positiva como negativa en cada uno de los grupos generados.

Se encontraron 7 perfiles de estudiantes que desertan:

- **Perfil 1 - Incomprendidos**: Estudiantes con un promedio de edad de 19 años, sin necesidad de trasladarse de residencia para ir a estudiar, cuyo padre trabaja como empleado, entre uno y dos hermanos, sin necesidad de crédito para pagar la matrícula, sin dificultades académicas en su educación media, no les gusta dormir en su tiempo libre, no les gusta compartir con los amigos, provenientes de muy buen colegio en términos académicos, en los dos primeros semestres de su carrera y con un muy bajo rendimiento académico en la universidad.

- **Perfil 2 - Dormilones**: Estudiantes con un promedio de edad de 18 años, sin necesidad de trasladarse de residencia para ir a estudiar, cuyo padre trabaja como independiente, entre uno y dos hermanos, sin necesidad de crédito para pagar la matrícula, sin dificultades académicas en su educación media, les gusta dormir en su tiempo libre, les gusta compartir con los amigos, provenientes de muy buen colegio en términos académicos, en los dos primeros semestres de su carrera y con buen rendimiento académico en la universidad.
- **Perfil 3 – Independientes**: Estudiantes con un promedio de edad de 26 años, sin necesidad de trasladarse de residencia para ir a estudiar, cuyo padre ya ha fallecido o no viven con él, entre uno y dos hermanos, con necesidad de crédito para pagar la matrícula, sin dificultades académicas en su educación media, les gusta dormir en su tiempo libre, les gusta compartir con los amigos, provenientes de buen colegio en términos académicos, en los cinco primeros semestres de su carrera, con aceptable rendimiento académico en la universidad y rezagados en mínimo dos semestres en su carrera.

- **Perfil 4 - Problemas económicos**: Estudiantes con un promedio de edad de 22 años, sin necesidad de trasladarse de residencia para ir a estudiar, cuyo padre trabaja como independiente, entre uno y dos hermanos, con necesidad de crédito para pagar la matrícula, sin dificultades académicas en su educación media, no les gusta dormir en su tiempo libre, les gusta compartir con los amigos, provenientes de muy buen colegio en términos académicos, en los primero cuatro semestres de su carrera, con buen rendimiento académico en la universidad y rezagados en mínimo dos semestres.
- **Perfil 5 – Relajados**: Estudiantes con un promedio de edad de 23 años, sin necesidad de trasladarse de residencia para ir a estudiar, cuyo padre trabaja como independiente, con un hermano, sin necesidad de crédito para pagar la matrícula, con dificultades académicas en su educación media, les gusta dormir en su tiempo libre, les gusta compartir con los amigos, provenientes de muy buen colegio en términos académicos, en los primeros cuatro semestres de su carrera, con bajo rendimiento académico en la universidad y rezagados en mínimo dos semestres.
- **Perfil 6 - Pilos no tan pilos**: Estudiantes con un promedio de edad de 21 años, con necesidad de trasladarse de residencia para ir a estudiar, cuyo padre trabaja como independiente, entre uno y dos hermanos, sin necesidad de crédito para pagar la matrícula, sin dificultades académicas en su educación media, no les gusta dormir en su tiempo libre, no les gusta compartir con los amigos, provenientes de un colegio con rendimiento medio en términos académicos, en los primeros dos semestres de su carrera y con bajo rendimiento académico en la universidad.
- **Clúster 7- BomBril**: Estudiantes con un promedio de edad de 27 años, sin necesidad de trasladarse de residencia para ir a estudiar, cuyo padre trabaja como empleado, entre uno y dos hermanos, sin necesidad de crédito para pagar la matrícula, sin dificultades académicas en su educación media, no les gusta dormir en su tiempo libre, no les gusta compartir con los amigos, provenientes de un colegio con rendimiento alto en términos académicos, en

los últimos semestres de su carrera, con rendimiento académico aceptable en la universidad y rezagados en mínimo dos semestres.

De los estudiante que NO desertan, se encontraron 4 perfiles:

- **Perfil 1 - Persistentes**: Estudiantes con un promedio de edad de 22 años, sin necesidad de trasladarse de residencia para ir a estudiar, cuyo padre trabaja como independiente, con un hermano, sin necesidad de crédito para pagar la matrícula, sin dificultades académicas en su educación media, no les gusta dormir en su tiempo libre, les gusta compartir con los amigos, provenientes de muy buen colegio en términos académicos, con avance de seis semestres en su carrera, con muy buen rendimiento académico en la universidad y rezagados en mínimo dos semestres.

- **Perfil 2 - Juiciosos semestres intermedios**: Estudiantes con un promedio de edad de 20 años, sin necesidad de trasladarse de residencia para ir a estudiar, cuyo padre trabaja como empleado, con un hermano, sin necesidad de crédito para pagar la matrícula, sin dificultades académicas en su educación media, no les gusta dormir en su tiempo libre, no les gusta compartir con los amigos, provenientes de muy buen colegio en términos académicos, con avance de cuatro semestres en su carrera, con muy buen rendimiento académico en la universidad y sin rezago en su avance semestral.
- **Perfil 3 - Juiciosos primeros semestres:** Estudiantes con un promedio de edad de 19 años, sin necesidad de trasladarse de residencia para ir a estudiar, cuyo padre trabaja como independiente, entre uno y dos hermanos, sin necesidad de crédito para pagar la matrícula, sin dificultades académicas en su educación media, no les gusta dormir en su tiempo libre, no les gusta compartir con los amigos, provenientes de buen colegio en términos

académicos, en sus primeros dos semestres de su carrera y con muy buen rendimiento académico en la universidad.

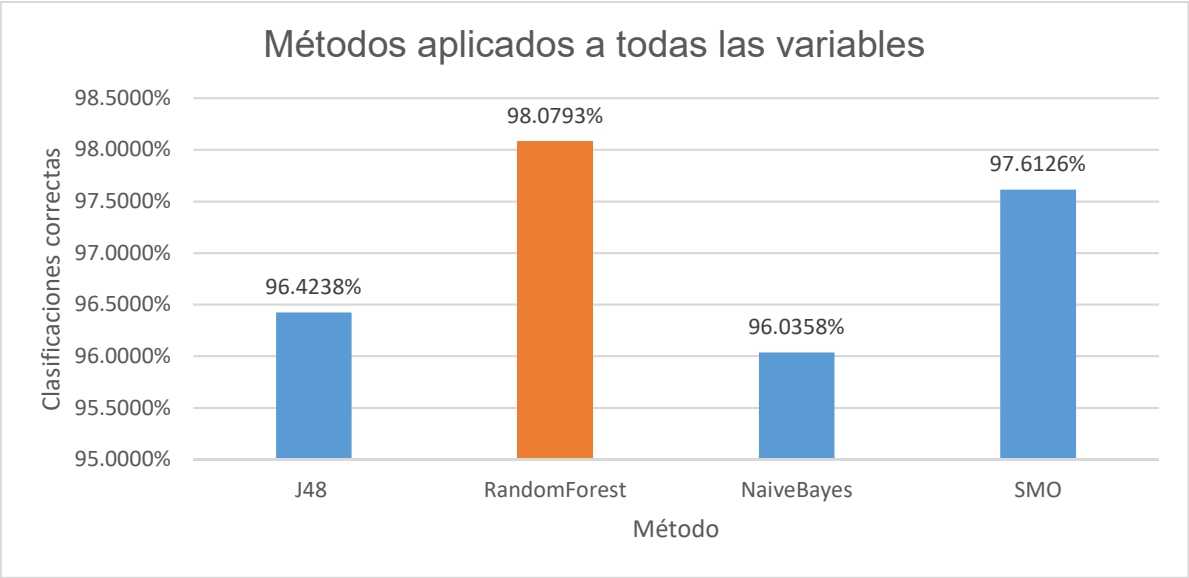
- **Perfil 4 - Avanzados en la carrera:** Estudiantes con un promedio de edad de 20 años, sin necesidad de trasladarse de residencia para ir a estudiar, cuyo padre trabaja como independiente, con un hermano, sin necesidad de crédito para pagar la matrícula, sin dificultades académicas en su educación media, no les gusta dormir en su tiempo libre, les gusta compartir con los amigos, provenientes de muy buen colegio en términos académicos, con avance de cinco semestres en su carrera, con muy buen rendimiento académico en la universidad y sin rezago en su avance semestral.

En relación a las reglas de asociación, se observan algunas recurrencias en los estudiantes no desertores:

- Los estudiantes que eligen la carrera por interés propio y que tienen un porcentaje de créditos aprobados/intentados superior al 91% normalmente no desertan.
- Los estudiantes que no presentaron dificultades académicas en el colegio, sin novedades en su estado académico y con porcentaje de créditos aprobados/intentados superior al 91% normalmente no desertan.
- Los estudiantes que no tuvieron dificultades académicas en el colegio, sin dificultades académicas en la universidad y con un promedio de 3.7, normalmente no desertan.

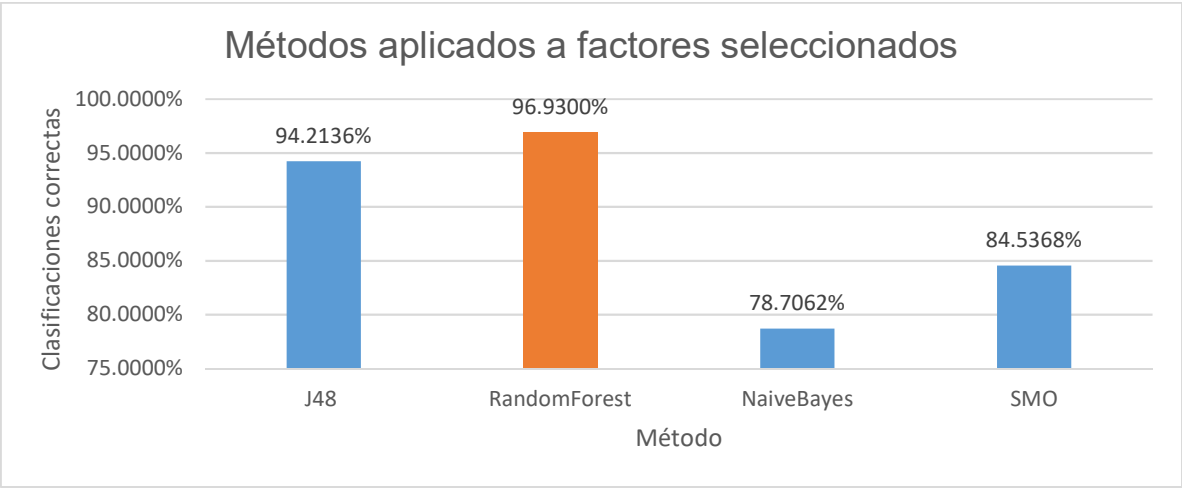
Ahora evaluando los resultados de los modelos aplicados para predicción, se puede inicialmente evidenciar que, de los cuatro modelos aplicados con todas las 200 variables, el mejor desempeño lo obtiene el modelo el árbol de decisión RandomForest al utilizar todas las variables. Esto de cierta forma coincide con el estado del arte, donde normalmente se referencian los árboles de decisión con buenos resultados en cuanto a modelos predictivos para deserción.

Ilustración 31. Porcentaje de precisión para los modelos predictivos con todas las variables



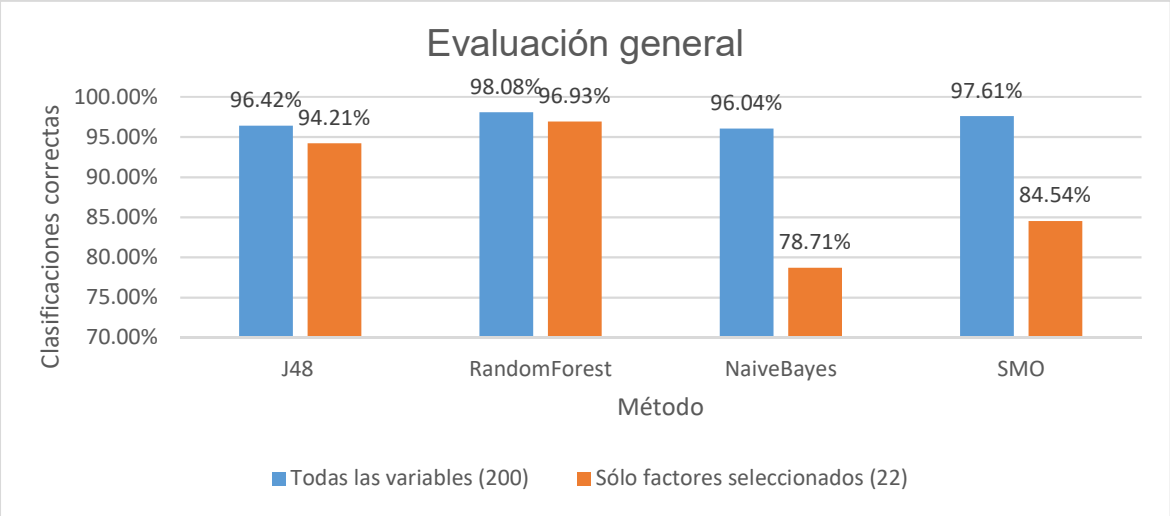
Lo mismo ocurre al reducir el número de variables a los 22 factores seleccionados, nuevamente el método RandomForest obtiene el mejor desempeño.

Ilustración 32. Porcentaje de precisión para los modelos predictivos con factores seleccionados



Ahora evaluando el impacto entre realizar la predicción con todas las variables y sólo con los factores seleccionados, se observa que, aunque los desempeños caen un poco, no es de forma significativa. Además, el trabajar con todas las variables hace que, dependiendo del método a aplicar, el tiempo se aumente considerablemente al momento del despliegue.

Ilustración 33. Resumen general modelos predictivos



13.2 Recomendaciones

Teniendo en cuenta la considerable cantidad de preguntas que se aplicaron en la encuesta de perfil integral, y que NO todas resultaron relevantes para prevenir la deserción, se recomienda a la Unidad de Permanencia de la Universidad capturar desde primer semestre las 22 variables que quedaron seleccionadas como las más relacionadas con la deserción y a su vez, actualizar semestre a semestre dicha información.

Teniendo presente los 7 perfiles encontrados para los estudiantes desertores, se recomienda crear políticas de apoyo desde la Unidad de Permanencia según las características de cada tipo de estudiante.

Con las 22 variables seleccionadas, además de los patrones de conducta encontrados con los estudiantes NO desertores en las reglas de asociación, se recomienda a la Unidad de Permanencia filtrar aquellos estudiantes en los cuales se deben enfocar, es decir en aquellos que SI tienen riesgo de deserción.

Finalmente, sobre el modelo predictivo de deserción se sugiere implantar el modelo creado con las 22 variables obtenidas en la selección de factores con el método RandomForest (bosque de árboles).

14. CONCLUSIONES

- En este trabajo se realizó un estudio de la deserción en estudiantes de pregrado de la Universidad Pontificia Bolivariana. Se analizaron los datos capturados en la encuesta de perfil integral aplicada por la Universidad a los estudiantes durante los años 2015 y 2016. Adicionalmente, se analizaron los datos académicos de los estudiantes y del colegio de procedencia.

- En el estudio de desarrollaron 6 modelos analíticos: 1) Selección de factores que más tiene relación con la deserción por medio de un sistema de votación de 3 métodos: correlaciones, ganancia y árbol de decisión (datos balanceados). 2) Perfiles de estudiantes que desertan por medio de un análisis de clustering. 3) Perfiles de estudianes que no desertan por medio de un análisis de clustering. 4) Reglas de asociación para buscar la co-ocurrencia de eventos en los datos. 5) Predicción de deserción usando todas las variables. 6) Predicción de deserción usando sólo las variables seleccionadas en el primer experimento.

- La encuesta de perfil integral aplicada por la Bienestar Universitario tenía inicialmente 277 columnas, de las cuales eran susceptibles de analizar como variables 200, pero tan sólo unas cuantas influían directamente en la probabilidad de deserción. Esto indica que para futuras caracterizaciones se debe ser más preciso en la selección de preguntas y los factores seleccionados por este caso de estudio podrían servir como base para ayudar a delimitar dichos aspectos a tener en cuenta.

- No todos los métodos para realizar clasificación, en este caso para predecir la deserción, tienen comportamientos homogéneos en cuanto a resultados. Por eso se hace necesario aplicar varios métodos y elegir el que mejor se

ajuste. El árbol de decisión RandomForest presentó el mejor desempeño entre los 4 métodos aplicados para este caso de estudio.

15. TRABAJOS FUTUROS

Las cifras oficiales de deserción por cohorte calculadas por el MEN muestran cifras muy altas respecto a la UPB. Sin embargo, las obtenidas en este trabajo; que aunque no se miden exactamente igual y tampoco trabajó con deserción por cohorte, muestran unas cifras muy diferentes. Sería interesante analizar el por qué de estas diferencias y de ser el caso realizar una segunda fase de experimentos con los ajustes necesarios.

Realizar encuestas con los estudiantes que han desertado y mejorar con dichos datos la efectividad del modelo predictivo.

REFERENCIAS

Astin, A. W. (1975). *Preventing students from dropping out*. Retrieved from https://programs.honolulu.hawaii.edu/intranet/sites/programs.honolulu.hawaii.edu.intranet/files/predicting-freshman-dropout.pdf

Baepler, P., & Murdoch, C. J. (2010). International Journal for the Scholarship of Teaching and Learning Academic Analytics and Data Mining in Higher Education Academic Analytics and Data Mining in Higher Education. *International Journal for the Scholarship of Teaching and Learning*, *4*(2). https://doi.org/10.20429/ijsotl.2010.040217

Bayer, J., Bydžovs, H., Eryk, J., Aš Obšívač, T., & Popelínsk, L. (2012). *Predicting drop-out from social behaviour of students*. Retrieved from https://is.muni.cz/th/uw0s1/papers/EDM_2012.pdf

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (1999). *Step-by-step data mining guide*. Retrieved from https://www.the-modeling-agency.com/crisp-dm.pdf

Congreso de Colombia. (2006). Ley 1098 de 2006. *Ley 1098 de 2006*, *2006*(46), 1–118. Retrieved from http://www.ins.gov.co:81/normatividad/Leyes/LEY 1098 DE 2006.pdf

Congreso de Colombia. (2008). *Ley 1266 de 2008*. Retrieved from http://wp.presidencia.gov.co/sitios/normativa/leyes/Documents/Juridica/Ley 1266 de 31 de diciembre 2008.pdf

Estudios Económicos de la OCDE Colombia www.oecd.org/eco/surveys/economic-survey-colombia.htm. (2017). Retrieved from http://www.oecd.org/eco/surveys/Colombia-2017-OECD-economic-survey-overview-spanish.pdf

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). *Knowledge Discovery in Databases: An Overview*. Retrieved from https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1011/929

González Fiegehen, L. E. (2006). *Repitencia y deserción universitaria en América Latina*. (April), 13. Retrieved from https://www.unila.edu.br/sites/default/files/files/Fiegehen, Luis Eduardo González - Repitencia y deserción universitaria en América Latina.pdf

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining. Concepts and Techniques, 3rd Edition*. Retrieved from http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

Heckman, J. J. (2008). *SCHOOLS, SKILLS, AND SYNAPSES*. https://doi.org/10.1111/j.1465-7295.2008.00163.x

IBM. (2012). Manual de CRISP-DM de IBM SPSS Modeler. Retrieved April 15, 2018, from ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/es/CRISP-DM.pdf

Jisc. (2013). CASE STUDY A: Traffic lights and interventions: Signals at Purdue University. *Learning Analytics in Higher Education*. Retrieved from https://analytics.jiscinvolve.org/wp/files/2016/04/CASE-STUDY-A-Purdue-University.pdf

Kotsiantis, S., Patriarcheas, K., & Xenos, M. (2010). A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education. *Knowledge-Based Systems*, *23*, 529–535. https://doi.org/10.1016/j.knosys.2010.03.010

Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2003). Preventing student dropout in distance learning using machine learning techniques. *Lecture Notes in Artificial Intelligence Subseries of Lecture Notes in Computer Science*, *2774 PART*, 267–274. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-8344235939&partnerID=40&md5=3bb04edd5cc41a409eee9d1fc302abef

Levitz, R. S., Noel, L., & Richter, B. J. (1999). Strategic Moves for Retention Success. *New Directions for Higher Education*, *1999*(108), 31–49.

https://doi.org/10.1002/he.10803

Márquez Vera, C., Romero Morales, C., & Ventura Soto, S. (2012). Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos. *IEEE-RITA*, *7*(3). Retrieved from http://rita.det.uvigo.es/201208/uploads/IEEE-RITA.2012.V7.N3.A1.pdf

MEN, M. de E. N. (2009). *Deserción estudiantil en la educación superior colombiana. Metodología de seguimiento, diagnóstico y elementos para su prevención*. Retrieved from https://www.mineducacion.gov.co/sistemasdeinformacion/1735/articles-254702_libro_desercion.pdf

MEN, M. de E. N. (2014). *MANUAL DE PREGUNTAS FRECUENTES*. Retrieved from https://www.mineducacion.gov.co/sistemasdeinformacion/1735/articles-254704_archivo_pdf_manual_preg_frecuentes.pdf

MEN, M. de E. N. (2016). *Estadisticas de Educacion Superior*.

Ocaranza, O., & Quiroz, M. (2006). *Deserción Estudiantil en el Pregrado en la Pontificia Universidad Católica de Valparaíso, Chile*. Retrieved from http://www.universidadfutura.org/wp-content/uploads/2012/05/Repitencia-y-Deserción-Universitaria-en-América-Latina1.pdf

Oñate Bowen, A. A. (2016). *Análisis de la Deserción y Permanencia Académica en la Educación Superior Aplicando Minería De Datos* (Universidad Nacional de Colomiba). Retrieved from http://bdigital.unal.edu.co/53635/1/alvaroagustinoñatebowen.2016.pdf

Pal, S. (2012). Mining Educational Data Using Classification to Decrease Dropout Rate of Students. *INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY SCIENCES AND ENGINEERING*, *3*(5). Retrieved from www.ijmse.org

Reason, R. D. (2009). Student Variables that Predict Retention: Recent Research and New Developments. *NASPA Journal*, *46*(3). Retrieved from https://pdfs.semanticscholar.org/c110/06dcdcf410747105e75a93b7845d0f34e0ec.pdf

Riquelme, J. C., Ruiz, R., & Gilbert, K. (2006). ARTÍCULO Minería de Datos: Conceptos y Tendencias. *Inteligencia Artificial, Revista Iberoamericana de*

Inteligencia Artificial. No, *29*, 11–18. Retrieved from https://idus.us.es/xmlui/bitstream/handle/11441/43290/Minería de datos.pdf?sequence=1&isAllowed=y

Timarán Pereira, R. (2013). *Detección de Patrones de Bajo Rendimiento Académico y Deserción Estudiantil con Técnicas de Minería de Datos*. Retrieved from http://www.iiis.org/cds2008/cd2009cSc/CISCI2009/PapersPdf/C692YV.pdf

Tinto, V. (1975). Dropout from Higher Education: A Theoretical Synthesis of Recent Research. *Review of Educational Research Winter*, *5*(1), 8–9. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.874.5361&rep=rep1&type=pdf

Tinto, V. (1989). *DEFINIR LA DESERCION: UNA CUESTION DE PERSPECTIVA*. Retrieved from https://www.researchgate.net/publication/252868573

Torres, F. S., & Zúñiga, J. M. (2012). *La Deserción en la Educación Superior en Colombia durante la Primera Década del Siglo XXI: ¿Por qué ha aumentado tanto?* Retrieved from www.cadena.com.co

UPB. (2018). Misión, Visión y Valores. Retrieved May 31, 2018, from https://www.upb.edu.co/es/identidad-principios-historia/mision-vision-valores

Vélez Martínez, N. (2016). *EFECTIVIDAD DE LAS ESTRATEGIAS DE RETENCIÓN ESTUDIANTIL EN LA LICENCIATURA INGLÉS-ESPAÑOL MODALIDAD DISTANCIA DE LA UNIVERSIDAD PONTIFICIA BOLIVARIANA DE MEDELLÍN, ENTRE LOS AÑOS 2013 Y 2015*. Retrieved from https://repository.upb.edu.co/bitstream/handle/20.500.11912/3619/EFECTIVIDAD DE LAS ESTRATEGIAS DE RETENCIÓN ESTUDIANTIL.pdf?sequence=1

Wirth, R., & Hipp, J. (2002). *CRISP-DM: Towards a Standard Process Model for Data Mining*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.198.5133&rep=rep1&type=pdf

Yang, D., Sinha, T., Adamson, D., & Rose, C. P. (2017). *“Turn on, Tune in, Drop out” : Anticipating student dropouts in Massive Open Online Courses*. Retrieved from https://www.cs.cmu.edu/~diyiy/docs/nips13.pdf