ESTUDIO SOBRE LA DESERCIÓN ESCOLAR UTILIZANDO MINERÍA DE DATOS EN LA MODALIDAD DE ENSEÑANZA VIRTUAL EN LA INSTITUCIÓN UNIVERSITARIA POLITÉCNICO GRANCOLOMBIANO

Jhonny Cano, Julian Olarte, Henry Solarte

Resumen
En la actualidad la Institución Universitaria Politécnico Grancolombiano, oferta programas académicos a través de modalidad virtual. En ésta modalidad de enseñanza hay una elevada tasa de deserción estudiantil; Se propone usar minería de datos como herramienta que puede contribuir a determinar las causas de deserción con el fin de adoptar mecanismos que disminuyan este fenómeno. Para lograr este objetivo se cuenta con diversos orígenes de datos suministrados por el claustro universitario, los cuales se depuran y preparan para su posterior utilización en una vista minable y luego se procesan mediante un aplicativo llamado RapidMiner. Los hallazgos encontrados arrojan causales comunes que llevan a la deserción estudiantil en modalidad virtual y brindan conocimiento básico y fundamental con miras a instituir procesos que eviten la deserción.

Keywords
Modalidad Virtual, Deserción, Vista Minable, Minería de datos

[1] *Universidad Politécnico Grancolombiano, Maestría en Ingeniería de Sistemas, Bogotá, Colombia*
*__Correos de los autores__: jhonnycano@hotmail.com, jolarter@poligran.edu.co, henrysolarte@hotmail.com

Índice

Introducción

La Universidad Politécnico Grancolombiano tiene como misión ofrecer educación con altos estándares de calidad a la sociedad colombiana, y de este modo contribuir a su cualificación y crecimiento.

Por este motivo, y con miras de llegar a la mayor parte del territorio de nuestro país, la Universidad ha incursionado en el mercado educacional a través de la modalidad virtual.

No obstante lo anterior, dentro de esta modalidad académica existe un alto número de deserciones, lo cual se traduce en una disminución de conocimiento y personal capacitado en el país.[1]

En este contexto, la Ingeniería de Sistemas puede suministrar herramientas valiosas que permitan a esta comunidad académica conocer las causales que originan el fenómeno de la deserción, con el fin de adoptar las medidas necesarias para disminuir esta problemática.

La minería de datos entonces se vislumbra como una

herramienta material y efectiva que nos permite lograr este cometido por medio de un proceso claramente definido, cuyo resultado final contribuirá a entender el fenómeno y promover acciones para su erradicación.

En dicho proceso existen diversas tareas que nos permiten realizar la transición de los datos en información y éste en conocimiento.

Entre las primeras tareas se encuentra la comprensión tanto del negocio, como de los datos con los que se cuenta para realizar la tarea de minería, esto con el fin de plantear los objetivos a los que se quiere llegar por medio del proyecto de minería de datos, y cómo se traduce este objetivo de minería en un objetivo de negocio.

Luego de entender los datos disponibles, se debe proceder a una fase de preparación de los mismos, con el fin de que se ajusten a las entradas requeridas por los algoritmos de minería identificados como factibles de aplicar en las fases iniciales del proyecto.

Posteriormente, en la fase de modelado, se procesan y calibran los parámetros de los modelos para extraer los mejores resultados a partir de los algoritmos elegidos.

Existen otras fases posteriores, que permiten evaluar el resultado del proyecto de minería y generalizar su aplicación dentro de la organización.

1. Trabajos Relacionados

Respecto al tema objeto de investigación, se han adelantado varios estudios, entre los cuales destacan:

La investigación efectuada por Márquez, Romero y Ventura en la Unidad Académica Preparatoria de la Universidad Autónoma de Zacatecas (UAPUAZ) de México, investigación que acude a la minería de datos para analizar el fenómeno de la deserción universitaria. Se ha mostrado la utilidad de las técnicas de selección de características al utilizar un conjunto reducido de 15 atributos de entre los 77 disponibles inicialmente, se han seleccionado 10 algoritmos de clasificación disponibles por la herramienta de minería de datos Weka, obteniendo un modelo de salida comprensible para el usuario, y se obtienen reglas de clasificación del tipo “Si – Entonces” o árboles de decisión.

A su vez, la Universidad de la Sabana por conducto de Maria Claudia Moreno y Stephanie Mendez, acudió a la minería de datos con el fin de realizar un estudio de deserción universitaria en la facultad de ingeniería. En esta investigación se aplicó la metodología *Rough set*, que busca la identificación de las variables criticas influyentes.

En la Universidad de Nariño, los investigadores Ricardo Timarán Pereira, Andrés Calderón Romero, Javier Jiménez Toledo, adelantaron una investigación sobre el descubrimiento de perfiles de deserción estudiantil con técnicas de minería de datos. Dentro de la misma se recurrió a las técnicas de clasificación y clustering sobre los datos de los estudiantes utilizando el aplicativo Weka.

2. Justificación

El Ministerio de Educación Nacional, define la deserción como una situación a la que se enfrenta un estudiante cuando aspira y no logra concluir su proyecto educativo, considerándose como desertor a aquel individuo, que siendo estudiante de una institución de educación superior, no presenta actividad académica durante dos semestres académicos consecutivos, lo cual equivale a un año de inactividad académica.[2]

La educación Superior en Colombia presenta altas tasas de deserción estudiantil, especialmente en los primeros semestres académicos, lo cual conlleva a efectos de tipo financiero, académico y social, tanto para las Instituciones de Educación Superior (IES) como para el estudiante, la región, y el país.

La Institución Universitaria Politécnico Grancolombiano, no es ajena a esta situación, toda vez que en la modalidad de educación virtual ofertada por este claustro universitario, se presenta en gran volumen deserción estudiantil, generando disminución de cobertura y de la materialización del quehacer institucional.

En este contexto es de vital importancia analizar las causas que originan esta problemática, esto, con la finalidad última de lograr disminución de estas tasas y así contribuir a la cualificación y progresión del estado, ya que al incrementar la población profesional de una nación, se crea desarrollo progreso y avances en las diferentes esferas que la conforman, lo que genera altos niveles de crecimiento y tecnificación.

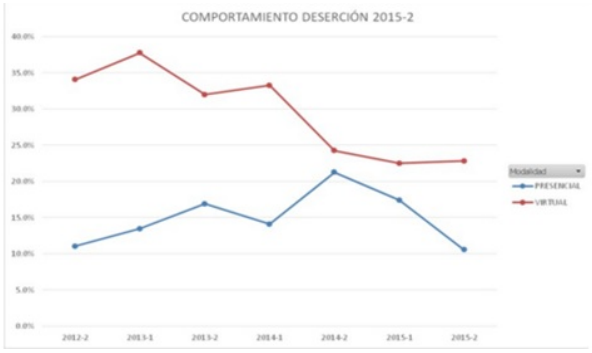
Ahora bien, la Ingeniería de Sistemas, a través de áreas como la minería de datos, permite realizar investigaciones que arrojan información sobre las causas que originan deserción, lo cual se traduce en un insumo vital que permite adoptar estrategias, medidas y acciones, que inciden de forma directa y efectiva en la prevención de la deserción estudiantil.

Esto a su vez, permite mitigar este fenómeno derivando en un proceso académico sostenible y reduciendo afectaciones que van desde la órbita económica hasta la sociocultural.

3. Preparación de datos

El Politécnico Grancolombiano recibe cada semestre un número significativo de estudiantes en sus programas académicos en modalidad presencial y virtual. La tasa de permanencia de los estudiantes y su continuidad en esta institución, se calcula sobre las matrículas recibidas en cada semestre, versus las matrículas recibidas en el semestre anterior. A su vez, el complemento del indicador permanencia es el indicador de deserción, que presenta los siguientes datos históricos para

los últimos cuatro años (siete periodos):



Se espera que con este proyecto de minería de datos se identifiquen atributos, reglas o agrupaciones naturales de los estudiantes que toman la decisión de retirarse de un programa académico y crear programas, campañas y en general decisiones que desestimulen la deserción.

Por tanto, se requiere realizar una caracterización de los estudiantes para comprender el fenómeno de deserción y posteriormente se pretende crear un modelo que permita identificar los estudiantes que están en riesgo de desertar de un programa académico. Para delimitar el problema, se estudiarán únicamente los estudiantes de la modalidad virtual y su fenómeno de deserción durante los dos primeros semestres.
Gracias a la información suministrada por el Politécnico Grancolombiano, se procesaron seis tablas principales de las cuales se derivó el trabajo de investigación.

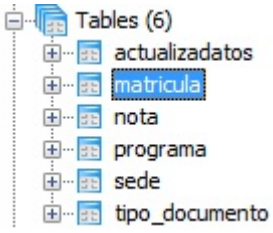


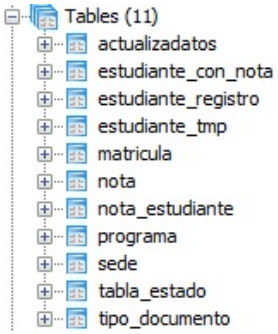
Tabla **ACTUALIZADATOS**, tiene información sobre los estudiantes que actualizaron su información personal en el año 2015. Cuenta con un total de 99.549 registros y 27 columnas.
Tabla **NOTA** posee información de las notas de los estudiantes desde el año 2008. Esta tabla contiene un total de 867.011 registros y 32 columnas.
Tabla **MATRICULA**, base de datos de las transacciones de matrícula, renovaciones, cambios de jornada, entre otros desde el 2003.
Información secundaria empleada en la investigación fue accesada de las tablas: Programa, sede, TipoDocumento.
Tabla **PROGRAMA**, contiene información de los programas de la institución universitaria Politécnico Grancolombiano, con un total de 99 registros y un total de 7 columnas.
Tabla **SEDE** con un total de 149 registros y un total de 5 columnas.
Tabla **TIPO_DOCUMENTO** con un total de 12 registros y un total de 2 columnas.

El análisis de calidad de datos se realiza una vez migrados los archivos de Excel a la base de datos de Postgres; previo a este análisis, se realizó reformateo de los nombres de las columnas para facilitar su trabajo con las mismas. A continuación se procede a verificar la consistencia de los campos, se tratará de realizar diferentes procedimientos en aquellos campos que no estén totalmente diligenciados, en caso de que no se pueda lograr su consistencia se procederá a eliminarlos.

De forma posterior, se procedió a almacenar la base de datos en un servidor de Amazon y un proveedor de Cloud llamado Heroku. Se precisa que Heroku es una plataforma de servicio de computación en la Nube que soporta distintos lenguajes de programación (php, Ruby, Python, etc.) y bases de datos (Postgres, MySql, etc.)

El objetivo de colocar la base de datos en la nube es permitir que las personas del grupo puedan trabajar con datos en tiempo real, y así dividir el trabajo en 3 Partes:
1. Tabla Estudiantes, en donde cada estudiante fuera único, determinando ciertos atributos para esta tabla.
2. Tabla HistoricoMatriculas para conocer el comportamiento de los estudiantes respecto a las matrículas.
3. Tabla de Nota para conocer el comportamiento de las notas de los estudiantes y su incidencia en la deserción.

Es así como se pasó de 6 tablas inicialmente a un total de 11 tablas, cifra última con la cual se desarrolló el trabajo.



3.1 Extracción de datos para estudiantes

Una de las primeras tareas detectadas en la depuración de los datos consiste en adaptar los datos de estudiantes, de modo que quede una tabla donde cada registro represente un único estudiante. Para llegar a este resultado se realizaron los siguientes pasos:

- Se extrajeron datos de estudiante de las diferentes tablas relacionadas (actualizadatos, matricula)
- Se cruzaron de modo que se pudiese acceder a una llave primaria con significado para el negocio (codigo_estudiante)

- Se agruparon los datos para evitar la repetición de estudiantes.
- Una vez encontrada la tabla de estudiante, se verificó que no hayan estudiantes repetidos.

Al realizar la agrupación de datos, se detectó que el estudiante se repetía, con los mismos datos, pero con diferentes jornadas o sedes, esto ocurre porque un estudiante puede encontrarse inscrito en programas como inglés virtual, alianza con el SENA, etc., adicional a los programas convencionales.

Para poder individualizar esta información, se realizó el proceso similar a la numeración 1-n con miras a extraer información de filas a columnas para los atributos considerados como más importantes para la vista minable.

Así pues, los campos resultantes derivaron en:

• Jornada_virtual
• Sedesena_virtual
• Sede_ingles_virtual
• Sede_pregrado_virtual
• Sede_postgrado_virtual

3.2 Extracción de datos para notas

La tabla nota discutida en apartados anteriores, contiene un registro por cada nota definitiva de un estudiante en una materia determinada, relacionando además atributos como la carrera, la nota, el programa, entre otros; hubiese sido interesante contar con la información del profesor que dictó la materia, pero dadas las restricciones de tiempo, no fue posible acceder a dicha información.

Dado que para la vista minable la decisión fue que cada registro representase a un estudiante, esta información de notas debió ser manejada en agrupaciones para poder asociarla a un estudiante y que la misma fuera accedida en un solo registro; por tanto, se tomó la decisión de extraer la información de notas de los tres primeros periodos de cada estudiante, y para cada periodo.

Para extraer esta información, se siguió el procedimiento descrito a continuación:

En primer término, se individualizaron los registros de la tabla nota (tabla origen de las notas) por medio de una consulta GROUP BY; los datos de agrupación se crearon, dejando vacíos para calcular en pasos posteriores.

Para el segundo paso, se tomó el identificador de los primeros tres periodos para cada estudiante (campos s1_id, s2_id, s3_id); en caso que no lo estuviera, también era importante marcar dicho identificador como NULL;

4. Clustering

Una vez culminada la fase de extracción de datos para estudiantes, extracción de datos para notas y extracción de datos para matriculas, se creó una tabla que contiene la vista minable, esta tabla se llamó vista_minable_clustering y cuenta con un total de 38 campos que son indispensables para el estudio en relación.

4.1 Síntesis de los ejercicios de minería de datos

Una vez obtenida la vista minable a partir de los datos a analizar, se procede a hacer los siguientes ejercicios:
• Separación de desertores según el periodo en el que desertaron.
• Generación de orígenes de datos para los diferentes ejercicios
• Preparación de datos para Clustering
• Conclusiones preliminares

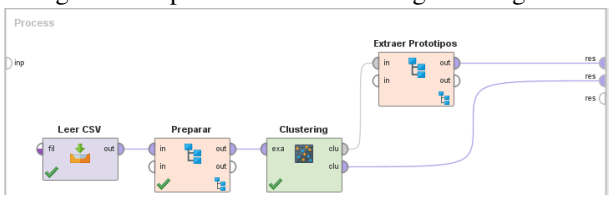
En la vista minable del anterior ejercicio, se obtuvo una cantidad de 4390 registros con clase "desertor", los cuales en algunas columnas tienen valores nulos, dependiendo del semestre en el que abandonaron sus estudios. Para poder manejar esta situación, se realizó una separación de desertores de acuerdo al último periodo en el cual tuvieran notas registradas. De este modo, se filtró la vista minable del ejercicio inicial en tres consultas:

• Desertores de primer semestre: Estudiantes marcados como "desertor"que no registraron notas en segundo semestre (1157).
• Desertores de segundo semestre: Estudiantes marcados como "desertor"que registraron notas en segundo semestre, pero

no en tercer semestre (670).
• Desertores de tercer semestre: Estudiantes marcados como "desertor"que registraron notas en tercer semestre (2563).

De cada consulta, solamente se extrajeron los datos relevantes, de forma tal que las columnas que tenían valores NULL quedaron excluidas de cada análisis.

Para cada grupo de desertores, se realizó el ejercicio de clustering como se puede observar en la siguiente figura:



Como se puede apreciar, la primera tarea consiste en la lectura de datos desde un archivo csv extraído desde Postgres, para luego realizar las preparaciones de datos que se explicarán en secciones siguientes, y luego someter los datos transformados al algoritmo de clustering y como medida final, el almacenamiento del resultado detallado y de prototipos detectados en archivos de Excel; en las siguientes secciones se realiza una revisión del proceso de preparación realizado.

4.2 Clustering de desertores en primer semestre

Una vez extraídos los datos de desertores de primer semestre, se realizó una preparación de datos con el objetivo de exponer los datos a un algoritmo de clustering (k-means) que nos permitiera detectar patrones de similaridad o equivalencia en aquellos estudiantes.

Las siguientes fueron las tareas realizadas durante la preparación de datos:

• Exclusión de columnas (Materia)

Debido a que hay muchas materias en esta vista, y teniendo en cuenta que la misma información se encuentra de forma agrupada en la variable "área", se decide excluir la columna de materia para facilitar los ejercicios sub-siguientes; esta decisión se sustenta realizando el primer ejercicio de clustering (desertores en primer periodo) sin encontrar diferencias sustanciales en los clustering resultantes.

• Materia y área (Nominal to Numerical)

Para convertir la variable "área.en valores compatibles con el algoritmo de clustering, se realizó una numerización 1-n; al ver que hay relación entre las variables de materia y área, se decidió realizar el ejercicio de clustering dos veces, la primera vez usando la variable "materiaz la segunda vez, prescindiendo de ella para ver si se encontraban grupos más definidos de esta manera; como se explicó anteriormente, al hacer el ejercicio completo, no se encontraron grandes diferencias en los clusters encontrados, por lo que se decide proseguir con los demás ejercicios prescindiendo de la variable materia.

• Edad (Discretize y Nominal to Numerical)

Para no usar la edad como un argumento entero, se discretizó este valor en diferentes categorías (0-18, 19-21, 22-24, 25-27, 28-30, 31-33, 33+).

Luego, estos valores fueron transformados en columnas (numerización) para poder someterlos al algoritmo de clustering.

• Valores numéricos (Normalize)

Las variables restantes fueron normalizadas en rango 0-1: materias cursadas y perdidas, nota mínima y máxima, promedio.

Para este primer ejercicio de clustering, se encontraron varios clusters que inicialmente no nos brindaron información evidente; sin embargo, después de una cuidadosa observación de los prototipos generados por la herramienta, encontramos las tendencias que resaltamos en los clústers a continuación que juzgamos como más interesantes:

• Cluster 0:

- Estudiantes que no perdieron ninguna materia en primer semestre.
- Promedio general todas las materias: (4.21).
- Promedio general nota mínima: (3.72).
- 70 % de estudiantes entre 25 y 33 años, 90 % entre 22 y 33 años, 0 % estudiantes mayores de 33 años.
- Materias no vistas en el primer periodo: (física, humanidades, idiomas, electivas).
- 50 % de notas mínimas en materias de educación, administración, finanzas, economía y contabilidad.
- 53 % de notas máximas en materias de administración, contabilidad y psicología.
- Resultados no concluyentes en materias de matemáticas (nota mínima: 17.3 %, nota máxima: 17.9 %).

• Cluster 4:

- Estudiantes que pudieron perder alguna materia en primer semestre pero en promedio tuvieron buen rendimiento.
- Promedio general todas las materias: (4.25).
- Promedio general nota mínima: (3.68).
- 100 % estudiantes mayores de 33 años.
- Materias no vistas en el primer periodo: (biología, programación, ingeniería industrial, idiomas, electivas).

- 50 % de notas mínimas en materias de matemáticas, educación y administración.
- 52 % de notas máximas en materias de Administración de sistemas, psicología y finanzas.

s1_nota_min_area = Administración	0,2923
s1_nota_min_area = Finanzas	0,1346
s1_nota_min_area = Educación	0,1154
s1_nota_min_area = Ingenieria Industrial	0,0923
s1_nota_min_area = Matemáticas	0,0654
s1_nota_min_area = Biología	0,0538
s1_nota_min_area = Economia	0,0462
s1_nota_min_area = Electiva en Sistemas	0,0423
s1_nota_min_area = Ingenieria de Sistemas	0,0385
s1_nota_min_area = Pasajes y Servicio a bordo	0,0308
s1_nota_min_area = Administración de Sisitemas	0,0269
s1_nota_min_area = Psicología	0,0154
s1_nota_min_area = Contabilidad	0,0077
s1_nota_min_area = Mercadeo y Publicidad	0,0077
s1_nota_min_area = Pogramaciòn y Desarrollo	0,0077
s1_nota_min_area = Ingles	0,0077
s1_nota_min_area = Seguros	0,0038
s1_nota_min_area = Humanidades	0,0038
s1_nota_min_area = Electiva	0,0038
s1_nota_min_area = Diplomados	0,0038
s1_nota_min_area = Politica y Laboral	-
s1_nota_min_area = Administración Pública	-
s1_nota_min_area = Periodismo	-
s1_nota_min_area = fisica	-
s1_nota_min_area = Ecologia y Medio Ambiente	-
s1_nota_min_area = Banca	-
s1_nota_min_area = Idiomas	-
s1_nota_min_area = Sistemas de Base	-
s1_nota_max_area = Administración de Sisitemas	0,3423
s1_nota_max_area = Ingenieria de Sistemas	0,0923
s1_nota_max_area = Idiomas	0,0846
s1_nota_max_area = Psicología	0,0808
s1_nota_max_area = Economia	0,0808
s1_nota_max_area = Finanzas	0,0808
s1_nota_max_area = Educación	0,0654
s1_nota_max_area = Administración	0,0385
s1_nota_max_area = Matemáticas	0,0308
s1_nota_max_area = Opción Grado	0,0308
s1_nota_max_area = Politica y Laboral	0,0192
s1_nota_max_area = Pasajes y Servicio a bordo	0,0115
s1_nota_max_area = Diplomados	0,0115
s1_nota_max_area = Electiva en Sistemas	0,0077
s1_nota_max_area = Biología	0,0077
s1_nota_max_area = fisica	0,0038
s1_nota_max_area = Humanidades	0,0038
s1_nota_max_area = Ecologia y Medio Ambiente	0,0038
s1_nota_max_area = Sistemas de Base	0,0038
s1_nota_max_area = Contabilidad	-
s1_nota_max_area = Ingenieria Industrial	-
s1_nota_max_area = Periodismo	-
s1_nota_max_area = Mercadeo y Publicidad	-
s1_nota_max_area = Metodologia del Estudio	-
s1_nota_max_area = Administración Pública	-
s1_nota_max_area = Ingles	-
s1_nota_max_area = Seguros	-

- **Cluster 1:**

- Estudiantes que perdieron el 87 % de las materias que inscribieron.
- Promedio general todas las materias: (0.51).
- Promedio general nota mínima: (0.11).
- Variable de edad no conclusiva, 20 % en cada categoría.
- 50 % de notas mínimas en materias de administración, finanzas y educación.
- 52 % de notas máximas en materias de Administración de sistemas, ingeniería de sistemas e idiomas
- Materias no vistas en el primer periodo: (Política laboral, administración, contabilidad).

Cluster desertores S1.xls

4.3 Clustering de desertores en segundo semestre

Para los desertores en segundo semestre se realizaron las mismas tareas de preparación de datos, incluyendo las nuevas columnas correspondientes a información de segundo semestre.

En este segundo ejercicio de clustering, se realizó la misma tarea de observación encontrando lo que sigue:

- **Cluster 1:**

  - No perdieron materias en primer semestre.
  - En segundo semestre perdieron entre 3 y 4 materias.
  - Promedio general primer semestre: 3.8.
  - Promedio general segundo semestre: 2.4.
  - Promedio general nota mínima primer semestre: 3.2.
  - Promedio general nota máxima primer semestre: 4.4.
  - Promedio general nota mínima segundo semestre: 1.1.
  - Promedio general nota máxima segundo semestre: 3.7.
  - Tendencias de edad: 47.8 % total mayores de 31, y 64 % total mayores de 28.
  - 67 % de notas mínimas primer semestre en materias de matemáticas, educación y contabilidad.
  - 60 % de notas máximas primer semestre en materias de administración de sistemas y matemáticas.
  - 63 % de notas mínimas segundo semestre en materias de matemáticas, ingeniería industrial, administración y contabilidad.
  - 61 % de notas máximas segundo semestre en materias de Política, contabilidad, administración y humanidades.

- **Cluster 2:**

  - No perdieron materias en primer semestre.
  - No perdieron materias en segundo semestre.
  - Promedio general primer semestre: 3.9.
  - Promedio general segundo semestre: 3.9.
  - Cantidad de materias inscritas primer semestre: 10.2 (Homologaciones de otras instituciones).
  - Promedio general nota mínima primer semestre: 3.3.
  - Promedio general nota máxima primer semestre: 4.5.
  - Promedio general nota mínima segundo semestre: 3.4.
  - Promedio general nota máxima segundo semestre: 4.5.
  - 60 % de notas mínimas segundo semestre en materias de contabilidad, economía e idiomas.
  - 63 % de notas máximas segundo semestre en materias de contabilidad, matemáticas, ecología e idiomas.

- **Cluster 0:**

  - Perdieron 70 % de materias en primer semestre.
  - No perdieron materias en segundo semestre.
  - Inscribieron 8 materias en primer semestre (homologaciones).
  - Idiomas como nota mínima en el 17 % de los casos, tanto en primer como segundo semestre.

- **Cluster 4:**

  - Perdieron todas (o casi todas en su gran mayoría) las materias en primer semestre.
  - Perdieron todas las materias en segundo semestre.
  - Nota mínima en primer semestre: 0.9.
  - Nota máxima en primer semestre: 3.1.
  - Nota mínima en primer semestre: 0.15.
  - Nota máxima en primer semestre: 1.1.

4.4 Clustering de desertores en tercer semestre

Para los marcados como desertores en tercer semestre se identificaron los siguientes grupos especiales:

- **Cluster 0:**

  - Perdieron en promedio 2 a 3 materias en primer semestre
  - Perdieron en promedio 4 a 5 materias en segundo semestre
  - Perdieron en promedio 2 materias en tercer semestre
  - Promedio primer semestre: 3.1
  - Promedio segundo semestre: 2.2
  - Promedio tercer semestre: 2.8
  - Cantidad de estudiantes repartidos proporcionalmente en las categorías de edad
  - 28 % de notas mínimas primer semestre en materias de matemáticas
  - 35 % de notas mínimas segundo semestre en materias de matemáticas
  - 11 % de notas mínimas segundo semestre en materias de inglés
  - 36 % de notas mínimas tercer semestre en materias de matemáticas

- **Cluster 1:**

  - En promedio perdieron 0 o 1 materias en cada semestre de los tres semestres cursados
  - 100 % Mayores de 33 años

- **Cluster 4:**

  - Perdieron en promedio 3 materias en primer semestre
  - Perdieron en promedio 5 materias en segundo semestre
  - Perdieron en promedio 5 materias en tercer semestre
  - 51 % de notas mínimas segundo semestre en materias de matemáticas y economía
  - 30 % de notas mínimas segundo semestre en materias de economía

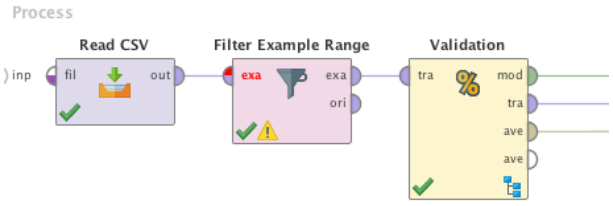
5. Predicción usando árboles de decisión

Una vez analizado el conjunto de datos se puede observar diferencias importantes en los clusters obtenidos. Mediante la técnica de árboles de decisión se establece un mecanismo para identificar posibles candidatos a desertar. Teniendo en cuenta los datos se crean tres experimientos diferentes, cada uno con grupos de datos diferentes dependiendo de los estudiantes que desertan luego del primer semestre, los estudiantes que desertan luego del segundo semestre y los que desertan luego del tercero.

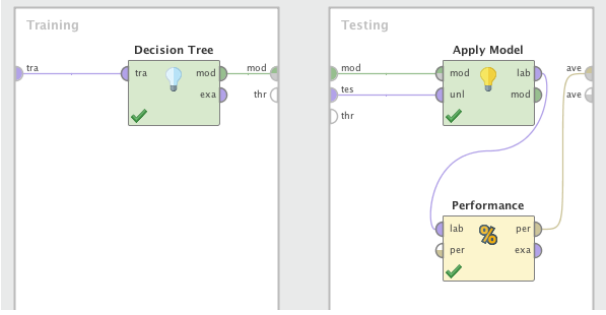
Para preparar éste conjunto de datos se escogen únicamente aquellos estudiantes que desertaron luego del tercer semestre y se prepara el conjunto de datos con cerca de 2500 desertores, un número similar de graduados y alrededor de 10000 estudiantes activos. Se realiza un proceso de preparación de datos similar al modelo de clustering.

5.1 Estudiantes que desertan en el tercer semestre

Para realizar el proceso de generación del árbol de decisión se usó el siguiente diagrama de procesos sobre la herramienta RapidMiner:



Y específicamente el bloque de validación contendrá los siguientes elementos:



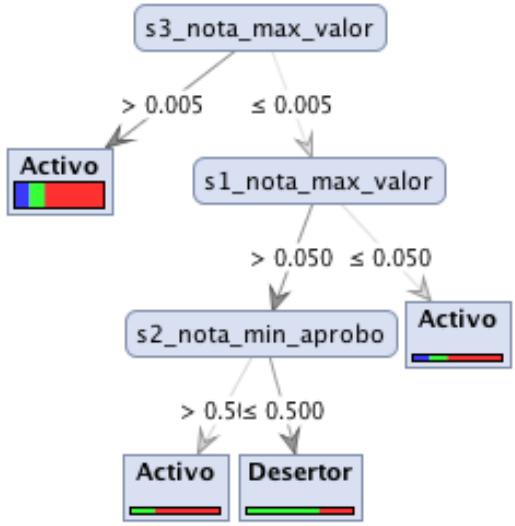
Los parámetros del bloque de validación (tipo split) se encuentran en la siguiente tabla.

En este proceso se usan 75 porciento de datos para entrenar el árbol de decisión y 25 porciento para realizar pruebas.

El resultado final del árbol de decisión se presenta en la siguiente figura.

Parámetro	Valor
Split	relative
Split Ratio	0.75
Sampling type	stratified sampling
criterion	gain ratio
maximal depth	20
confidence	0.25
minimal gain	0.001

Cuadro 1. Parámetros del bloque de validación.



Debido a la uniformidad de los datos no se logró tener un árbol sino hasta el minimal gain de 0.001. Esto indica que hay muy pocas variables que realmente establecen relación de causa y aún así no hay buenos datos de decisión con la vista minable usada.

Los resultados de clasificación para los datos de prueba con el árbol se extienden a continuación.

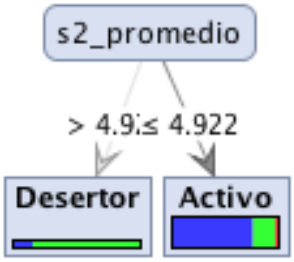
```
s3_nota_max_valor > 0.005: Activo
{Graduado=2558, Desertor=2476, Activo=9459}
s3_nota_max_valor <= 0.005
|   s1_nota_max_valor > 0.050
|   |   s2_nota_min_aprobo > 0.500: Activo
        {Graduado=0, Desertor=6, Activo=15}
|   |   s2_nota_min_aprobo <= 0.500: Desertor
        {Graduado=1, Desertor=79, Activo=36}
|   s1_nota_max_valor <= 0.050:
    Activo {Graduado=2, Desertor=2, Activo=6}
```

Se puede observar que, a pesar de los rangos de los valores, que las variables que generan un mayor valor de confianza son s3_nota_max_valor, s1_nota_max_valor y s2_nota_min_aprobo. Éstos valores tienen una influencia importante en la deserción

pero no lo suficientemente significativa como para ser tenida en cuenta debido a los valores de notas bajas para la comparación y uso de la regla.

5.2 Estudiantes que desertan en el segundo semestre

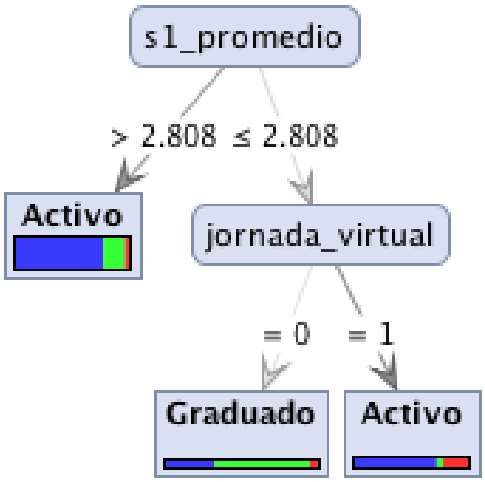
El proceso fue el mismo para el conjunto de datos y el resultado fue el siguiente.



Para éste caso la variable encontrada para la clasificación era s2_promedio pero de nuevo los valores límite no pueden ser usados en la práctica. Las combinación de las variables no permite hacer un ejercicio de clasificación.

5.3 Estudiantes que desertan en el primer semestre

El proceso fue el mismo para el conjunto de datos y el resultado fue el siguiente con algunas diferencias:



No hay datos para clasificar a los desertores o la información no es útil para clasificar. Se debe considerar que los datos no son suficientes para el ejercicio de predicción.

6. Conclusiones

Es prudente en este punto informar al lector que los datos obtenidos del ejercicio de clustering, tanto el detalle como los prototipos detectados, se dejan en el repositorio público de github en la siguiente dirección:

https://github.com/jhonnycano/poli-201601-mineria/tree/master/data/

Con base en los resultados encontrados en el ejercicio de clustering expuesto anteriormente, las conclusiones a las que podemos llegar son las siguientes:

- Los estudiantes con edades mayores tienden a desertar, independiente de los resultados académicos que obtengan.
- Los estudiantes que llegan de otras instituciones y pierden varias materias, tienden a desertar más seguido.
- Las materias de las áreas de economía y matemáticas estimulan un alto grado de deserción en el tercer semestre.
- Los estudiantes mayores de 28 años que pierden más de 3 materias en segundo semestre, tienen una mayor tendencia a la deserción.

Para los datos encontrados en el ejercicio de predicción se debe tener en cuenta que el conjunto de datos basado en las notas de los tres semestres iniciales en donde se presenta deserción no son suficientes para establecer causas del fenómeno. Se podría presentar la hipótesis que la actividad académica no es causa directa de la deserción pero deberá ser comprobada con otros ejercicios.

6.1 Mejoras y trabajos futuros

Hubiera sido mucho más útil contar con información demográfica de los estudiantes con el fin de detectar patrones de deserción relacionados con información de género, estrato socioeconómico, condiciones de vivienda, ingresos, etc.

Por otra parte, se encontró que se podrían realizar otro tipo de detecciones de grupos naturales, pero teniendo en cuenta otras variables que no se descubrieron al inicio del proyecto, tales como homologación de materias cuando los estudiantes vienen de otras instituciones, o tendencias de cumplimiento de pago de matrículas.

Se espera repetir el ejercicio con una mayor cantidad de variables. Tal como se recomienda en [3][4]

Referencias

[1] Rodrigo Ocampo Mejía and Jorge Raúl Ossa Botero. Exclusión social y sus relaciones con la deserción universitaria maestros y maestras pensando en clave de diversidad. 2014.

[2] Carlos Márquez Vera, Cristóbal Romero Morales, and Sebastián Ventura Soto. Predicción del fracaso escolar mediante técnicas de minería de datos. *Revista Iberoamericana de Tecnologías del/da Aprendizaje/Aprendizagem*, 109, 2012.

[3] Stephanie Méndez Morales, Maria Claudia Moreno Santiago, Mauricio Restrepo López, Héctor Andrés López Ospina, et al. *Uso de la Metodología Rough Sets para la identificación de variables críticas influyentes en una base de datos*. PhD thesis, 2012.

[4] Sergio Valero Orea, Alejandro Salvador Vargas, and Marcela García Alonso. Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. *Ene*, 779(73):33, 2005.