**Proyecto 1**

**Inteligencia de Negocios**

1. Como problema de negocio, queremos identificar un conjunto de factores que determinen el rendimiento académico de los estudiantes de bachillerato. En particular, su desempeño en la prueba del ICFES (Saber 11). Esto con el objetivo de poder plantear medidas que estén orientadas a mejorar los resultados generales de estos estudiantes. Para el proyecto, utilizaremos como DataSet los resultados de las pruebas Saber 11 por estudiante para el periodo 2016-2 disponible en <https://www.datos.gov.co/Educaci-n/SABER-11-2016-2-RESULTADOS-SABER-11/qmn6-uaks>. La fuente asociada es el Instituto Colombiano para la Evaluación de la Educación de el Instituto Colombiano para el Fomento de la Educación Superior (ICFES).
2. Preparación del DataSet :

Los datos se prepararon de manera distinta para el árbol de decisión y el algoritmo k-medios.

En ambos casos, se **eliminaron** los siguientes campos:

* PERIODO (Number (integer))
* ESTU\_CONSECUTIVO (String)
* ESTU\_TIPO\_DOCUMENTO (String)
* ESTU\_FECHA\_NACIMIENTO (String)
* ESTU\_ETNIA (Number (integer))
* ESTU\_COD\_RESIDE\_DEPTO (Number (integer))
* ESTU\_RESIDE\_DEPTO (String)
* COLE\_COD\_ICFES (Number (integer))
* COLE\_NOMBRE\_SEDE (String)
* ESTU\_ANO\_TERMINO\_BACHILL (Number (integer))
* ESTU\_MES\_TERMINO\_BACHILL (Number (integer))
* ESTU\_RECIBE\_SALARIO (String)
* PUNT\_LECTURA\_CRITICA (Number (integer))
* PERCENTIL\_LECTURA\_CRITICA (Number (integer))
* DESEMP\_LECTURA\_CRITICA (Number (integer))
* PUNT\_MATEMATICAS (Number (integer))
* PERCENTIL\_MATEMATICAS (Number (integer))
* DESEMP\_MATEMATICAS (Number (integer))
* PUNT\_C\_NATURALES (Number (integer))
* PERCENTIL\_C\_NATURALES (Number (integer))
* DESEMP\_C\_NATURALES (Number (integer))
* PUNT\_SOCIALES\_CIUDADANAS (Number (integer))
* PERCENTIL\_SOCIALES\_CIUDADANAS (Number (integer))
* DESEMP\_SOCIALES\_CIUDADANAS (Number (integer))
* PUNT\_INGLES (Number (integer))
* PERCENTIL\_INGLES (Number (integer))
* DESEMP\_INGLES (String)
* PUNT\_GLOBAL (Number (integer))
* PERCENTIL\_GLOBAL (Number (integer))

La decisión de eliminar los campos fué basada en el perfilamiento de los datos obtenida usando pandas\_profiling en python. Algunos campos fueron eliminados porque en su mayoría eran valores nulos (como estu\_etnia o estu\_fecha\_nacimiento); otros se eliminaron porque eran maneras alternativas de expresar la variable objetivo (como percentil\_c\_naturales o percentil\_global); otros campos se eliminaron por ser IDs en bases de datos externas (como cole\_cod\_dane\_institucion o cole\_cod\_icfes) ; otros campos se eliminaron por tener únicamente un valor repetido (como periodo). Tanto el jupyter notebook utilizado para el perfilamiento (ICFES.ipynb) como los resultados del perfilamiento (profiling.html) se pueden encontrar en la misma carpeta que este archivo. Para un mayor detalle de las razones por las que se decidió eliminar los campos, favor *referirse al apéndice.*

Se discretizaron los valores de punt\_global y el resultado fue agregado como una columna llamada ***PUNT\_GLOBAL\_binned.*** Para hacer esto, se definieron intervalos sobre el campo punt\_global (en intervalos de 100) para poder agrupar estudiantes con puntajes similares. Esto es importante pues los campos de esta variable (que es a la que está orientado el análisis) toman valores de 0 a 500. Así pues, valores entre 0-100 corresponden a 1 en el campo *PUNT\_GLOBAL\_binned,* valores entre 101-200 a 2, etc.

Cabe notar que al introducir el campo *PUNT\_GLOBAL\_binned,* se eliminó la variable *PUNT\_GLOBAL.*

Para el algoritmo de k-means, se borraron adicionalmente los siguientes campos pues en su mayoría son valores continuos que nos son relevantes. Además, fue necesario manejar los valores nulos, pues la distancia no puede ser calculada si no tiene suficientes argumentos, y el algoritmo decide basándose en la distancia :

* ESTU\_PAIS\_RESIDE (Number integer))
* ESTU\_LIMITA\_MOTRIZ (String)
* ESTU\_LIMITA\_INVIDENTE (String)
* ESTU\_LIMITA\_CONDICIONESPECIAL (String)
* ESTU\_LIMITA\_SORDO (String)
* ESTU\_LIMITA\_SDOWN (String)
* ESTU\_LIMITA\_AUTISMO (String)
* ESTU\_AREA\_RESIDE (String)
* COLE\_COD\_MCPIO\_UBICACION (Number (integer))
* ESTU\_RESIDE\_MCPIO (String)
* ESTU\_COD\_RESIDE\_MCPIO (Number (integer))
* ESTU\_ZONA\_RESIDE (Number (integer))
* COLE\_COD\_DANE\_INSTITUCION (Number (double))
* ESTU\_COD\_MCPIO\_PRESENTACION (Number (integer))
* ESTU\_MCPIO\_PRESENTACION (String)
* ESTU\_DEPTO\_PRESENTACION (String)
* ESTU\_COD\_DEPTO\_PRESENTACION (Number (integer))

Entonces, se reemplazaron los valores nulos por los valores más frecuentes de la columna. Dado que en general eran pocos los valores nulos por columna, se justifica hacer esta substitución.

**3. Análisis de calidad de datos**

En cuanto a la calidad de los conjuntos de datos, afirmamos que es buena por las siguientes razones:

* 1. La información es emitida por entidades gubernamentales y oficiales, lo cual nos da cierto nivel de confianza en los datos
  2. El conjunto de datos contiene bastante información (entradas). En particular, tenemos 81 columnas y 606K filas.
  3. Los periodos de tiempo de los DataSet son recientes (alrededor de año y medio), por lo que las conclusiones encontradas tienen una alta probabilidad de ser válidas en la actualidad.
  4. Hay columnas donde la gran mayoría de los datos son nulos. Esto dificulta el preprocesamiento de los mismos. Aun así, dada la gran cantidad de información del conjunto de datos, este sigue siendo útil.

En otras palabras, después del procesamiento de datos tenemos los datos en un formato estándar y entendible, sin columnas donde la mayoría de las entradas sean valores nulos (de hecho, nuestro flujo de k-means en Knime no admite valores nulos). Como quien emite los datos es una entidad gubernamental encargada de la calificación de los exámenes y recolección de los datos, tenemos que los datos son consistentes (que puede verificarse antes del preprocesamiento con las columnas redundantes) e íntegros. Además, la naturaleza de los datos hacen que estos sean exactos. Finalmente, después del preprocesamiento, en particular, de la eliminación de columnas redundantes, el conjunto de datos usado al aplicar las técnicas de minería de datos no tiene una cantidad importante de información duplicada. Por todo lo anterior, concluimos que los datos usados por los algoritmos tiene una alta calidad y en consecuencia sus resultados tienen una alta probabilidad de ser válidos.

4. Descripción de cómo el requerimiento de negocio es resuelto con el o los requerimientos de minería de datos propuestos, para lo cual debe diligenciar la tabla que se presenta a continuación.

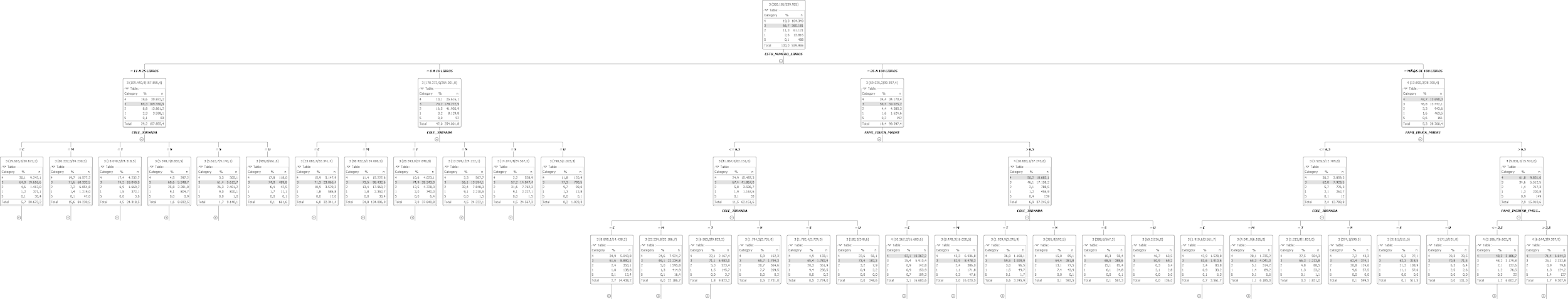
|  |  |
| --- | --- |
| **Oportunidad/problema Negocio** | Se quiere analizar el rendimiento de los estudiantes de bachillerato en el examen Saber 11. En particular se desea saber, dadas las condiciones de un estudiante, cuál será su posible resultado. También se desea saber qué estudiantes tienen características similares. Todo esto con el fin de adoptar medidas que permitan mejorar el rendimiento promedio de los estudiantes en estas pruebas. |
| **Descripción del requerimiento desde el punto de vista de minería de datos** | Se desea llevar a cabo una tarea de predicción, pues se quiere saber el posible rendimiento de un estudiante dadas determinadas condiciones socio-económicas y academicas. Por otro lado, se necesita llevar a cabo una tarea de clasificación que permita agrupar estudiantes e identificar tendencias en los resultados globales del ICFES en cada grupo. |

**Detalles de la actividad de minería de datos**

|  |  |  |
| --- | --- | --- |
| **Tarea** | **Técnica** | **Algoritmo y parámetros utilizados** |
| Predicción | Árboles de Decisión | Podado con MDL, índice Gini, Variable objetivo: punt\_global\_binned |
| Clasificación | Clustering | K-Medios con K=10 |

5. Resultado de 2 modelos analíticos

(i) Árboles de decisión

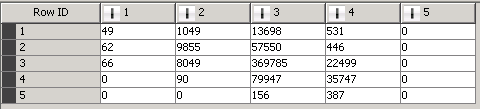


*Imagen 1: Arbol de decision*

Esta es una parte del árbol generado por Knime. El segundo nivel del árbol (hijos de la raíz) corresponde a cantidad de libros leídos. El tercer nivel al tipo de jornada y nivel de educación de la madre y el tercer nivel también al tipo de jornada y a los ingresos de la familia.

Es importante notar que se limitó a mostrar una sección del árbol pues no es práctico ni estético expandir todas las ramas. Además, por debajo del tercer nivel, no se encuentra un aumento significativo en el poder de predicción.

Matriz de confusión



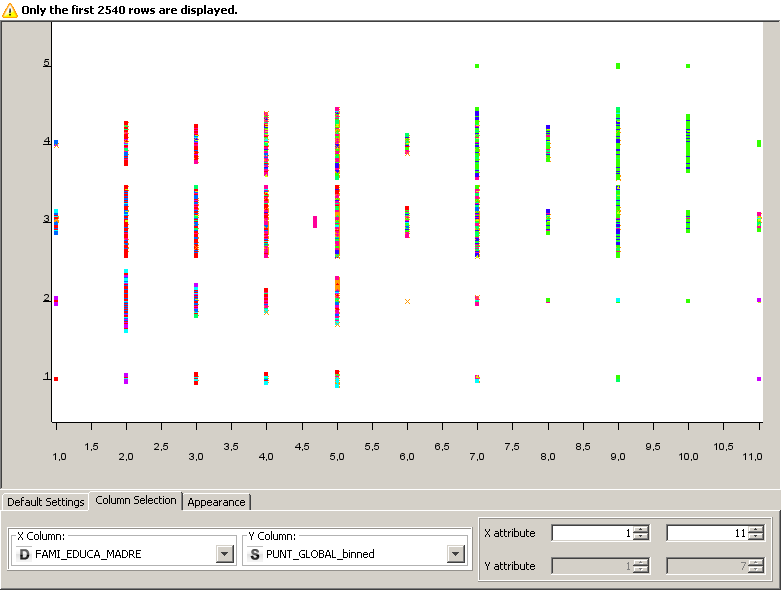
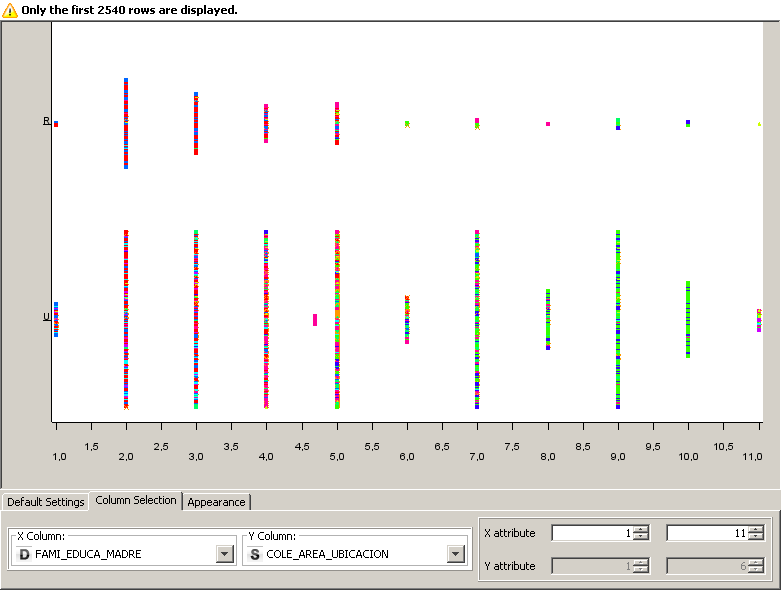
*Imagen 2: Matriz de Confusión*

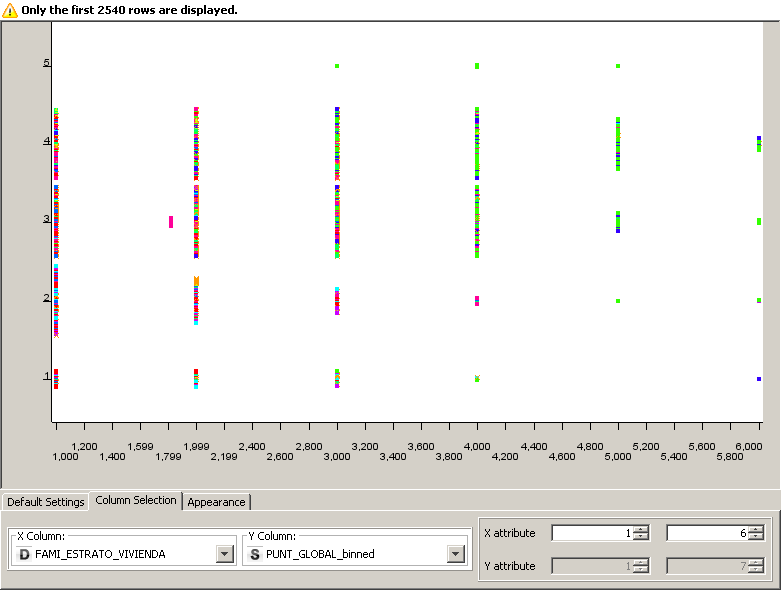
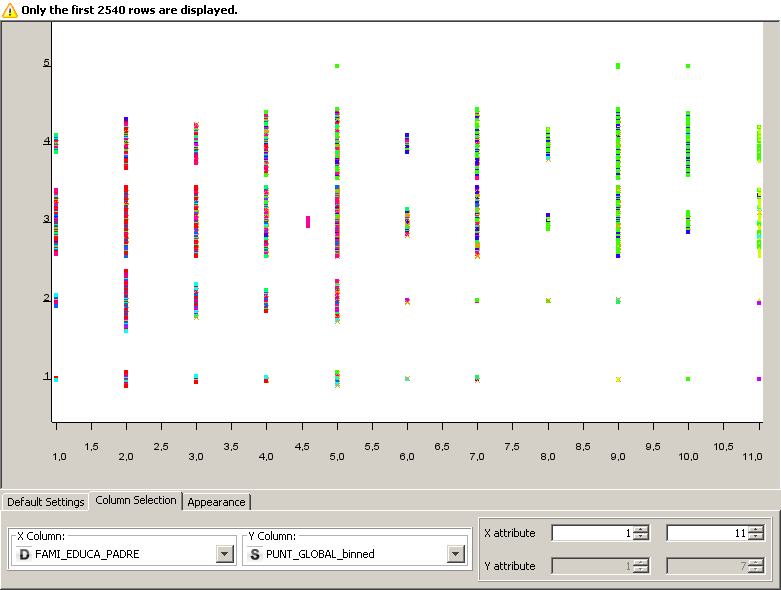
La razón para usar árboles de decisión es que es un algoritmo estándar para tareas de este tipo y además en nuestra experiencia (el laboratorio) resultó ser acertado y útil.

Se decidió podar el árbol para prevenir overfitting que invalidará los resultados encontrados, el algoritmo utilizado para podarlo fue MDL, pues es con el que trabajamos en el laboratorio y es el que Knime permite usar. La variable objetivo del arbol es punt\_global\_binned, pues la tarea de predicción consiste en tratar de determinar, basado en todas las características de un estudiante, cómo será su desempeño en la prueba, y el campo con el que tiene más sentido medir el desempeño es con el puntaje global (en intervalos).

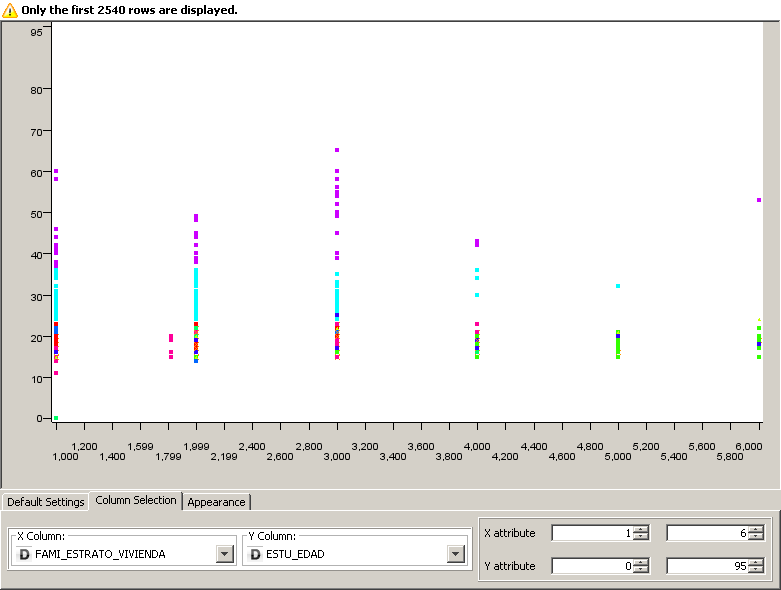
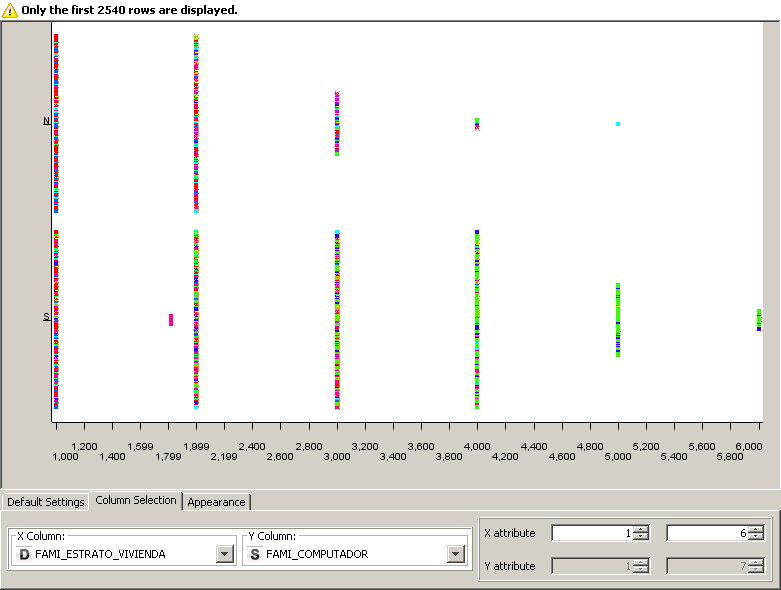
Este modelo es predictivo, entonces se puede identificar una fase de entrenamiento y una fase de prueba. El manejo que se le dió a los datos fue particionarlos de manera aleatoria con el flujo de Knime.

(ii) K-Means

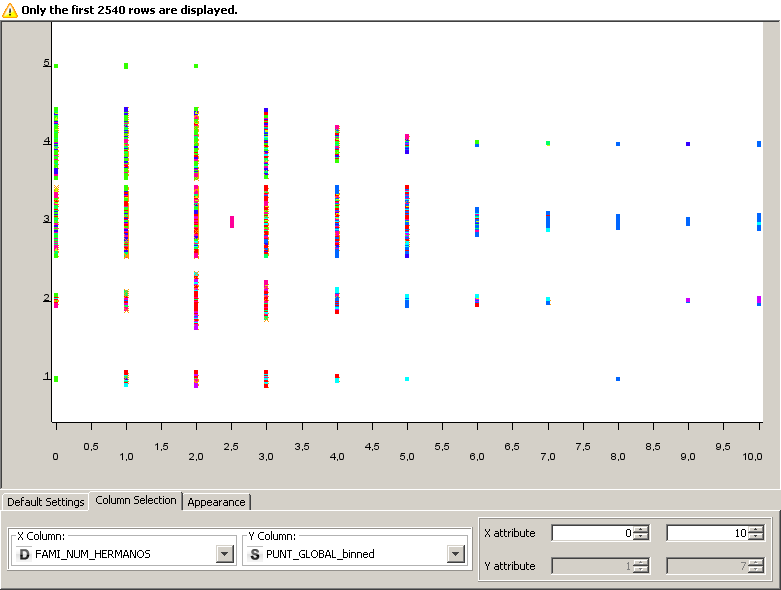
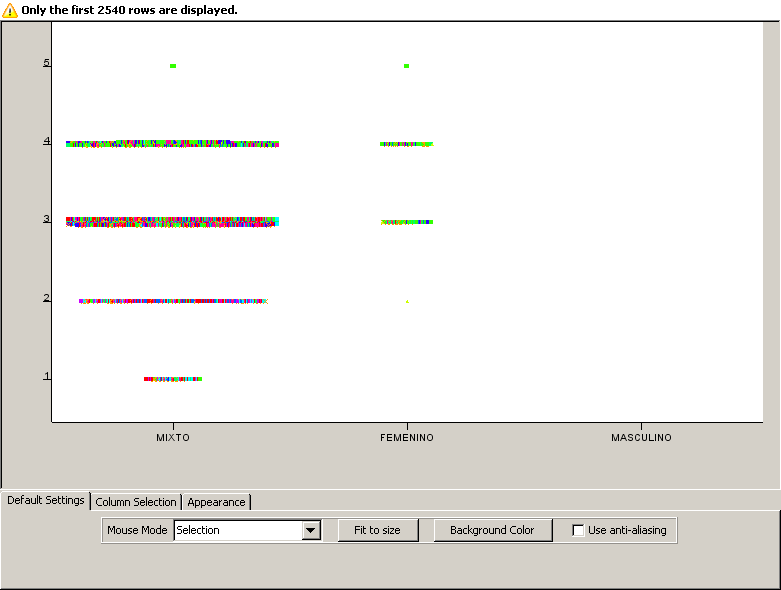


*Imagen 3 Imagen 4*

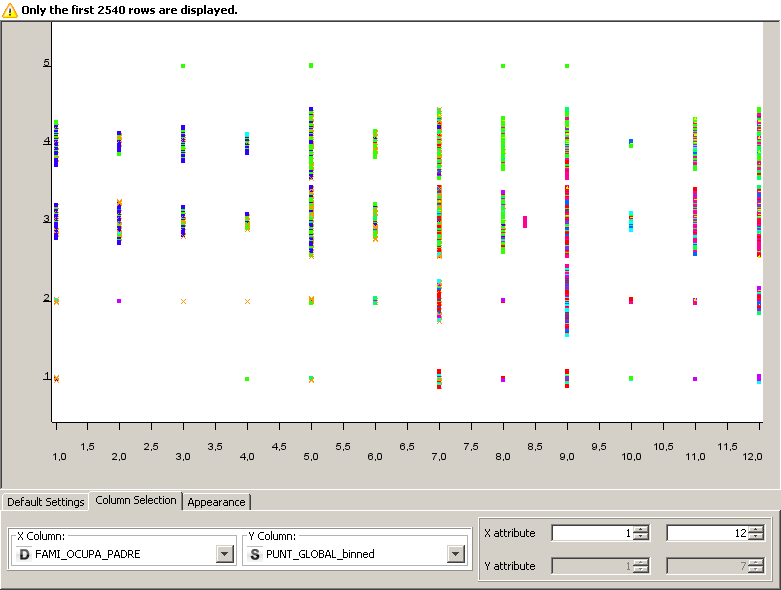
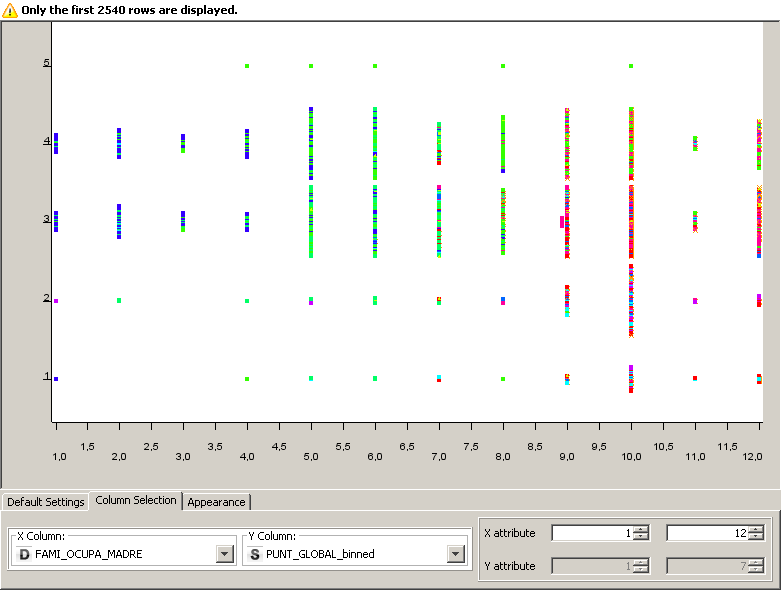
*Imagen 5 Imagen 6*



*Imagen 7 Imagen 8*



*Imagen 9 Imagen 10*



*Imagen 11 Imagen 12*

Estas imágenes corresponden a los clusters (10) resultado de la ejecución del algoritmo mostrando diferente información (para cada cluster solo se pueden representar 2 dimensiones). Se escogió k=10 después de varias iteraciones con distintos valores porque, a nuestro parecer, los resultados que mostraba con este parámetro eran más visibles. Se escogió K-Means como algoritmo de clustering porque es de rapida ejecucion y sus resultados son fáciles de interpretar. Además, es un algoritmo estándar de esta técnica y es de los que hemos visto hasta ahora en clase.

El análisis de los resultados obtenidos, tanto para el árbol como para K-Means, se explica en detalle en el punto (6)

6. Análisis de resultados

(i) Arbol de decision

El árbol de decisión clasificó correctamente el 70% de los datos. Dado el soporte (cantidad de valores que puede adoptar una variable) de la variable objetivo, consideramos que un valor del 70% correcto es un buen valor que indica que la tarea de predicción fue llevada a cabo con éxito. Una cosa que vale la pena mencionar es que la mayoría de los casos clasificados incorrectamente fueron clasificados en valores cercanos al valor real. Por ejemplo, si un estudiante cuyo puntaje global estaba en el intervalo 4 y fue clasificado incorrectamente es más probable que haya sido clasificado en el intervalo 3 que en el 1(pues el 3 está más cercano al 4 que el 1). Lo cual agrega una alguna noción de correctitud en las clasificaciones incorrectas, confirmando que la tarea cumplió con su objetivo.

Analizando los resultados puntuales de la ejecución, notamos que la variable que más determina el resultado del examen es la cantidad de libros leídos. Después de esta variable, encontramos que otro factor relevante para el resultado es el tipo de jornada de la institución del estudiante (unica, completa, mixta, nocturna) y el nivel de educación de la madre. También podemos ver que, para el caso en el que los estudiantes tienen más de 100 libros, otro de los factores relevantes es el nivel de ingresos de su familia. Lo que el árbol nos indica entonces, es que lo más relevante a la hora de predecir un posible resultado es la cantidad de libros que tiene el estudiante. Por otro lado, podemos ver que las variables relevantes en esta tarea de predicción reflejan, de algún modo u otro, las condiciones socio-económicas de un estudiante. Por ejemplo, el nivel de ingresos, o el nivel educativo de la madre (que en bastantes casos refleja bondad económica en la familia en la medida en que en muchos casos la educación debe ser pagada con recursos propios) y también, aunque en menor medida, la cantidad de libros. Al ver los valores del árbol vemos que, mientras las características reflejan mejores situaciones económicas, mejores son los resultados de los estudiantes.

(ii) K-means

Bajo un k=10, el algoritmo logró identificar varios segmentos. De todos los segmentos encontrados, hubo 4 particularmente interesantes:

* Los verdes:
  + Individuos de estrato alto, altos ingresos, muchas comodidades y alto puntaje
* Los rojos:
  + Individuo de estrato bajo, pocos ingresos, pocas comodidades y bajo puntaje
* Azules claros:
  + Individuos entre 20 y 35 años de estratos bajos y pocos recursos
* Fucsia:
  + Individuos con más de 35 años de estratos bajos y pocos recursos

De las gráficas más importantes es la de estrato vs puntaje (imágen 6). Aquí se evidencia el contraste entre los puntos verdes y los puntos rojos. Los puntos verdes están a la derecha de la gráfica (con estrato alto) y en promedio más arriba (mejor puntaje) que los puntos rojos. Es decir, se empieza a evidenciar que aquellos que tienen más recursos, en general, tienen un mejor puntaje en el ICFES. Esto está respaldado por las imágenes 4, 5 y 7 donde se vé en general una mayor educación conlleva un mayor estrato (los puntos verdes siguen estando a la derecha (gráficas 4 y 5)) y a ser más probable que exista un computador en la casa.

Por otro lado, la imágen 10 revela que entre ḿas hermanos se tenga, va a ser más probable tener una menor puntuación global en el ICFES. Sin embargo, considerando que entre más hijos se tenga, es más difícil garantizar una educación de calidad a todos por el costo que esto implica, se puede argumentar que en últimas la disminución en el puntaje se ve causada por una escasez de recursos.

Es interesante ver que los dos grupos principales descubiertos por el algoritmo fueron los de altos recursos con buenos resultados y bajos recursos con bajos resultados. Esto refleja la inequidad socioeconómica que padece lamentablemente nuestro país.

7. Estrategias que la organización debe plantear con base en los resultados obtenido.

La conclusión que se pudo sacar después de analizar los modelos analiticos es que la condición socio-económica de un estudiante tiende a determinar su rendimiento en las pruebas de estado en el siguiente sentido: como vimos en el análisis, unas mejores condiciones económicas favorecen el puntaje global. Desde el punto de vista de la organización, el problema de las condiciones es difícil de resolver. Aun así, gracias al proyecto de minería de datos desarrollado, la organización sabe sobre qué segmento de la población debe orientar sus esfuerzos para lograr un mejor desempeño general (que como se vió en el análisis es la población desfavorecida económicamente).

Lamentablemente, no hay muchas alternativas viables para mejorar en nivel educativo de las madres de los estudiantes ni los ingresos de sus familias. Sin embargo, se pueden llevar a cabo campañas que fomenten la lectura de libros (que resultó ser una de las variables que más determinaban el resultado) junto con estrategias que fomenten la inscripción a jornadas completas (C) y mixtas (M) que son las que registran mejores puntajes. Además, campañas de métodos anticonceptivos llevarían a una disminución en la cantidad de hijos por familia; esto permitiría enfocar los recursos económicos en menos hijos y por consiguiente, aumentar el puntaje de el examen ICFES.