[[1]](#footnote-0)

SaberPro: Predicción del puntaje por correlación y árboles de decisión

**William A. Rodríguez Mora  
Yasmin Sarbaoui**

Resumen

**Por medio de los datos suministrados públicamente por el icfes, específicamente en las pruebas de educación superior: SaberPro, se pretende identificar las relaciones entre distintas variables y el puntaje final obtenido en el área de matemáticas, es este abordaje un problema de predicción. Con esto, estimar el puntaje que puede obtener un estudiante dadas las variables con mayor correlación. Para esto nos servimos de un método clave en la clasificación supervisada: Árboles de decisión.**

***Palabras Clave*— Examen de evaluación de la educación superior, icfes, predicción puntaje, Saberpro.**

# INTRODUCCIÓN

El Instituto Colombiano para el Fomento de la Educación Superior brinda públicamente datos para fomentar la investigación alrededor de los exámenes de evaluación de la calidad en la educación colombiana. Hemos abordado los datos disponibles de la prueba Saber Pro, anteriormente denominada ECAES. Esta es una prueba realizada por el Estado por medio del ministerio de educación y del instituto ICFES. Está dirigida a evaluar oficialmente las competencias adquiridas por los estudiantes próximos a graduarse de programas de pregrado de las distintas instituciones de Educación Superior[[2]](#footnote-1).

El proceso de descubrimiento de conocimiento atravesará varias etapas, entre la primera de ellas, el pre-procesamiento, donde eliminaremos variables irrelevantes para nuestro objetivo puntual, limpieza de datos e integración de los datos, con el debido manejo de los casos con datos faltantes, que argumentamos más adelante..

Entre los atributos disponibles hay 87 variables distintas, de las cuales eliminamos 16 por tratarse de atributos irrelevantes para el análisis que no contienen información útil para la minería de datos ya que es información de tipo identificador y para esto ya hemos conservado la llave principal: Consecutivo del estudiante (ESTU\_CONSECUTIVO).

# Pre-procesamiento de los datos

Nuestro conjunto de datos contiene 87 variables y 324.850 registros. Tanto Nominales, ordinales, intervalo como variables continuas y Discretas.

En nuestro ejercicio el pre-procesamiento tuvo lugar en dos ocasiones, la primera ocasión sirvió para limpiar los elementos más intuitivos que era evidente para continuar con los algoritmos de asociación y agrupación, así como acercar el conjunto de datos al problema definido al inicio del proyecto.

Así nuestro conjunto de datos, luego del procesamiento preciso y de algunos ensayos con los algoritmos de asociación y agrupación, pasó a reducir nuevamente el conjunto de datos, las razones serán expuestas más adelante, haciendo énfasis a qué procedimiento pertenece a qué fase de preprocesamiento.

## Limpieza

No se encontró información duplicada, cada registro tiene un identificador único que refleja el identificador de cada inscrito al examen SaberPro.

Para agilizar la búsqueda de conocimiento en los datos que poseemos, hemos de descartar miles de datos inútiles, que no nos aportan información relevante y con quienes desperdiciamos recursos computacionales, como es el caso de los siguientes 16 atributos eliminados:

ESTU\_EXAM\_CODMPIO\_PRESENTA,

ESTU\_EXAM\_MPIO\_PRESENTACION y

ESTU\_EXAM\_DEPT\_PRESENTACION

Que corresponden a la ubicación de la presentación del examen, una variable que consideramos irrelevante en la búsqueda de correlaciones con el puntaje final.

Por otro lado, si bien creemos que la fecha de nacimiento es importante, dejaremos de esta únicamente el año, dado que no consideramos relevante las diferencias por días y tampoco por meses. Siendo así también eliminamos 2 variables:

ESTU\_NACIMIENTO\_DIA y ESTU\_NACIMIENTO\_MES.

De la residencia del estudiante dejaremos sólo el código del municipio y no su nombre. Así eliminamos el atributo: ESTU\_RESIDE\_MPIO.

Así mismo con el nombre de la institución donde estudia, se dejará sólo el código.

Se elimina el código ICFES de la institución dado que es uno por cada institución y ya tenemos un identificador para esta, estos códigos a eliminar no representan información relacionada al puntaje sino a procesos administrativos.

El consecutivo SNIES se dejará y se eliminará el nombre del programa académico representado por el código SNIES.

Dado que la mayoría de los siguientes códigos no tienen información y son otros identificadores para el mismo colegio donde estudió el estudiante, se procederá a eliminarlos: PLAN\_CODIGODANESEDE, PLAN\_CODIGODANEINSTITUCION, PLAN\_NOMBRECOMUN, PLAN\_MUNICIPIOSDR, DIPO\_NOMBREMUNICIPIO y DIPO\_NOMBREDEPARTAMENTO.

La variable semestre de presentación del examen no arroja información útil, se puede despreciar porque se tiene año: ESTU\_EXAM\_SEMESTRE\_PRESTAC.

Segunda fase

Esta segunda fase fue póstuma y a partir de un conocimiento más profundo del modelo de datos, lo que permitió conocer mejor la relación entre las variables y el sesgo que estaban sufriendo por proporciones incorrectas que existían, como la cantidad de variables de un misma área permeando la información y volcándola hacia atributos económicos y sociales, excluyendo de esta forma la proporción en que infieren otras áreas también importantes para nuestro objetivo, sea el caso del área educativa, donde sólo 3 variables medían atributos de esta área, desproporcionada área respecto al área económica, donde más de 10 variables alrededor de aspectos económicos menguaban la participación de otras áreas.

## Reducción

Se hizo un muestreo sobre todo el conjunto de datos para reducir costos computacionales y acelerar los algoritmos que se apliquen y demás exploraciones en búsqueda de patrones que nos genere conocimiento.

Es un muestreo aleatorio simple y sin reemplazo, se hizo por medio de una distribución uniforme. Se particionó todo el conjunto en 33 partes divididas aleatoriamente por la función random que ofrece MySQL en su versión 5.5.49. La metodología fue la siguiente:

1. Crear una nueva columna.
2. Asignarle un valor aleatorio entre 0 y 33.
3. Función usada: CEIL(RAND()\*33).
4. Se selecciona algún grupo de entre los 33 creados.
5. Se estimó que este grupo debía contener cerca de 10 mil registros (9844 estimados).

Hemos seleccionado el primer grupo y este contiene 9771 registros, muy cercano al valor esperado y que comprueba la distribución uniforme utilizada.

A continuación medimos la entropía de la muestra y la comparamos con la entropía total del conjunto para encontrar la ganancia de información y las implicaciones que nos evidencié ésta respecto a cuánta información nuestro conjunto es capaz todavía de aportar.

Si las medidas son cercanas, el conjunto de datos reducido aproxima el original conjunto de datos

Esto lo hacemos midiendo la similaridad entre los distintos datos del conjunto:

Fórmula 1. *Entropía del conjunto S.*

Donde es la similaridad entre el elemento *i* y *j*.

Los resultados obtenidos fueron los siguientes:

Entropía del nuevo conjunto es un valor cercano a 134 y el del conjunto anterior cercano a 210. Por tanto el valor nuevo de entropía es aceptable para nuestro ejercicio.

En tanto que en la segunda fase del pre-procesamiento no se ha hecho reducción del conjunto de datos, más que algunas eliminaciones de registros con valor 0 en puntaje y otras variables nulas que consideramos relevantes y que no podrían ser calculadas sin añadir sesgo.

## Agregación

Dado que muy pocos estudiantes sufren de discapacidades (0,22% del conjunto total y 0,17% en la muestra) se discretizan las 5 variables de discapacidad: ESTU\_DISC\_SORDO\_CEGUERA, ESTU\_DISC\_INVIDENTE, ESTU\_DISC\_SORDO\_CON\_INTERPRETE, ESTU\_DISC\_SORDO\_SIN\_INTERPRETE, ESTU\_DISC\_MOTRIZ.

En una sola variable que denota la presencia de discapacidad.

UPDATE temp SET DISCAPACIDAD = ESTU\_DISC\_INVIDENTE + ESTU\_DISC\_SORDO\_CON\_INTERPRETE + ESTU\_DISC\_SORDO\_SIN\_INTERPRETE + ESTU\_DISC\_MOTRIZ

Ahora con 5 variables reducidas a 1 sola hemos quedado con 66 variables.

La variable ESTU\_RESIDE\_CODMPIO identifica un municipio dentro de un departamento, se discretiza también este valor, agrupándolos por el nombre del departamento. Por facilidad en la interpretación dejaremos el nombre del departamento en vez de su código, quedando así con 65 variables.

Otras variables eliminadas:

ESTU\_COLEGIOTERMINO\_COD

ESTU\_TITULO\_BTO

GRRE\_CODIGOICFES

GRRE\_NOMBRE

CAPR\_PUNTAJE\_COMP

CAPR\_PUNTAJE\_LECT

CAPR\_PUNTAJE\_RAZO

CAPR\_PUNTAJE\_INGL

CAPR\_PUNTAJE\_COMU

Quedando después de la eliminación de variables redundantes o irrelevantes con un total de 56 variables.

Existen 2 registros con valores atípicos y carencia de información. Uno de ellos no tiene nada y el otro posee sólo los puntajes del examen. Dado que la información adicional al puntaje será la que nos ayudará a realizar la predicción, un registro sin estos datos no nos es útil. Procedemos a eliminar ambos registros despreciando el efecto que pueda tener dado el tamaño del conjunto, para quedar con un total de 9769 registros.

Hay atributos que ya están discretizados previamente en el dataset. Este es el caso de ESTU\_PJE\_CREDITOS que es el porcentaje de créditos necesarios para optar al título que ha cursado y aprobado hasta el momento. Sus valores posibles son:

0: No sigue sistema de créditos

1: Menos del 75%

2: Entre el 76% y el 80%

3: Entre el 81% y el 90%

4: El 90% o más

Y INST\_VLR\_MATRICULA\_ANT que es el valor anual de la matrícula del año anterior:

0: No pagó matrícula

1: Menos de 500 mil

2: Entre 500 mil y menos de 1 millón

3: Entre 1 millón y 3 millones

4: Entre 3 millones y 5 millones

5: Más de 5 millones

Como grado de educación y ocupación actual de los padres discretizada en más de 20 grupos. Por hablar de algunas de las variables discretizadas.

Segunda fase

En la segunda fase del pre-procesamiento, la agregación fue el proceso al cual acudimos con más frecuencia.

## Entre los datos productos de agregados, por mencionar, tenemos:

## COD\_EDUCA\_PADRES, originalmente era un atributo para madre y padre.

## ESTU\_TOMOCURSO engloba la suma o agregación de variables como horas por semana del curso, pagó instructores privados entre otros.

## OCUP\_PROGENIT también es una variable que engloba los atributos de cada padre.

## ECONOMIA es quizás la variable que engloba la mayor cantidad de atributos. Concentra el valor de material de los pisos, pertenencias como: Nevera, celular, televisor, internet entre otros servicios.

## INST\_SECTOR fue binarizado en público o privado.

* 1. Discretización

Entre las variables discretizadas están:

* INST\_COD\_INSTITUCION: Esta variable fue discretizada en 6bins, constituidos por los 5 primeros institutos o centros educativos, ordenados de mayor a menor por número de estudiantes, añadiendo un 6to bin que son los demás.
* ESTU\_HORAS\_TRABAJO discretizado en 0, 1-19,20 a 1, 2 y 3 respectivamente.
* Finalmente: *clase* fue la variable discretizada del puntaje (*puntajeMates*) que obtuvo cada persona por el área de matemáticas.

Este atributo nuevo que hará las veces de la clase a predecir, está dividida en 3 bins, divididos por frecuencia. Así, del dominio de posibles valores de *puntajeMates* se discretizan los bins así: .  
Y tiene las siguientes medias y varianzas.

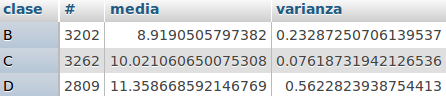


Imagen 1. Resultado de MySQL, AVG y VAR.  
  
Finalmente hemos obtenido las siguientes variables:

1. ESTU\_CONSECUTI
2. ESTU\_GENERO
3. ESTU\_NACIMIENTO\_ANNO
4. ESTU\_ESTADO\_CIVIL
5. DISCAPACIDAD
6. ESTU\_RESIDE\_DEPT
7. INST\_COD\_INSTITUCION
8. INST\_CARACTERACADEMICO
9. INST\_SECTOR
10. ESTU\_PJE\_CREDITOS
11. INST\_VLR\_MATRICULA\_ANT
12. ESTU\_TOMOCURSO
13. ESTU\_EXAM\_ANNO\_PRESENTACION
14. ESTU\_HOGAR\_ACTUAL
15. FAMI\_NUM\_PERS\_GRUP\_FAM
16. ESTU\_SN\_CABEZA\_FMLIA
17. FAMI\_NUM\_PERS\_CARGO
18. COD\_EDUCA\_PADRES
19. OCUP\_PROGENIT
20. ESTU\_ESTRATO
21. ECONOMIA
22. FAMI\_ING\_FMLIAR\_MENSUAL
23. ESTU\_TRABAJA
24. ESTU\_HORAS\_TRABAJO
25. clase

# Aplicación algoritmos de asociación

Dado que la aplicación de algoritmos de asociación mejoró notoriamente después del pre-procesamiento, omitiremos la asociación de la primera fase que no produjo reglas de calidad, dejando manifiesto a continuación de la gran diferencia entre las mejores reglas de cada fase de aplicación de los algoritmos de asociación:

Mejor regla de asociación encontrada antes del preprocesamiento fase 2: ESTU\_PJE\_CREDITOS=1 ECON\_SN\_SERVICIO\_TV=1 ECON\_SN\_NEVERA=1 5360 ==> ECON\_SN\_CELULAR=1 5082 <conf:(0.95)

Mejor regla de asociación posterior al preprocesamiento fase2: ESTU\_NACIMIENTO\_ANNO=mayor, INST\_CARACTERACADEMICO=UNIVERSIDAD (2215) ==> ESTU\_ESTADO\_CIVIL=A <conf:(0.97)>

La segunda regla mostrada nos lleva a asumir que si el estudiante tiene una fecha de nacimiento mayor a la mitad de los estudiantes y estudia en un centro educativo de carácter universidad, entonces es muy probable que su estado civil sea soltero.

## Reglas de asociación creadas

Las más destacadas han sido estas:

ESTU\_NACIMIENTO\_ANNO=mayor,FAMI\_ING\_FMLIAR\_MENSUAL=alto (2293) ==> ESTU\_ESTADO\_CIVIL=A | <conf:(0.97)>

ESTU\_SN\_CABEZA\_FMLIA(1751)==>ESTU\_TRABAJAzconf:(0.86)>

INST\_CARACTERACADEMICO=UNIVERSIDAD,INST\_SECTOR=PRIV (2172) ==> FAMI\_ING\_FMLIAR\_MENSUAL=alto <conf:(0.76)>

ESTU\_NACIMIENTO\_ANNO=mayor, INST\_SECTOR=PRIV, ESTU\_EXAM\_ANNO\_PRESENTACION=mayor (2000) ==> ESTU\_ESTADO\_CIVIL=A | <conf:(0.97)>

INST\_CARACTERACADEMICO=UNIVERSITARIA,FAMI\_ING\_FMLIAR\_MENSUAL=alto 1405 ==> ESTU\_TRABAJA <conf:(0.71)>

ESTU\_EXAM\_ANNO\_PRESENTACION=mayor,FAMI\_ING\_FMLIAR\_MENSUAL=alto ==> ESTU\_ESTADO\_CIVIL=A | <conf:(0.96)>

ESTU\_NACIMIENTO\_ANNO=mayor,ESTU\_EXAM\_ANNO\_PRESENTACION=mayor (4278) ==> ESTU\_ESTADO\_CIVIL=A | <conf:(0.96)>

ESTU\_NACIMIENTO\_ANNO=mayor (4628) ==> ESTU\_ESTADO\_CIVIL=A | <conf:(0.96)>

INST\_CARACTERACADEMICO=UNIVERSIDAD,ESTU\_EXAM\_ANNO\_PRESENTACION=mayor==>ESTU\_ESTADO\_CIVIL=A conf:0.95

Con las cuales se nota muy bien la mejora respecto a las anteriores.

Interpretando algunas asociaciones se podría afirmar que:

Si el estudiante es más joven que la otra mitad del conjunto y además el ingreso familiar es alto, muy seguramente su estado civil es soltero.

También se podría afirmar que si es cabeza de familia, entonces el estudiante labora. También, con una confianza del 0.76 si el carácter académico del centro educativo es *Universidad* y la financiación es privada, entonces es muy probable que el ingreso familiar sea alto. De la misma forma se le adjunta una proposición similar: Si estudia en Universidad y el ingreso familiar es alto, es probable que el estudiante trabaje. Más adelante se establecen reglas repetitivas y con confianza de 0.66 de si el ingreso familiar es alto, entonces es probable que esté en una privada y al revés.

Para forzar la aparición de reglas positivas (relaciones entre campos con 1) de presencia del atributo en el inscrito a la prueba, se modificó el comportamiento de 0 en el algoritmo, tomándolo como ausencia de valor, así no se crean tantas reglas con ausencia de los atributos. Sin embargo algunas reglas también fueron analizadas antes de asignarle a 0 el valor de ausencia.

Con esta modificación se evidencia la mejora de reglas respecto a las anteriores, donde entre las menos útiles teníamos: Si un inscrito carece de servicio de tv, lavadora, computador entonces no tendrá tampoco nevera. También si carece de computador, muy probablemente carece de acceso a internet.

# Aplicación algoritmos de agrupación

El primer obstáculo con el que nos encontramos en la fase 1 del preprocesamiento, fue la igualdad de relevancia entre columnas del conjunto, pues en el empleo de los algoritmos de agrupación presentes en los programas más reconocidos de minería de datos cada dimensión del conjunto es valorada en la misma proporción, suponiendo que el valor entre las similaridades en las *n* dimensiones del conjunto tienen una misma incidencia. Este problema se soluciona en la segunda fase porque dejamos una misma cantidad de variables por área.

Con el algoritmo EM (Expectation Maximization) obtuvimos, entre los atributos agrupados más significativamente:

Attribute | 0 1 2 3 4

| (0.14) (0.14) (0.2) (0.17) (0.35)

==================================================================

ESTU\_GENERO

F 766.0821 241.2137 941.8806 1004.7327 1713.0908

M 402.3815 884.6177 713.1808 359.6514 1141.1686  
ESTU\_ESTADO\_CIVIL

B 203.5964 233.9344 78.1233 301.2636 33.0822

E 172.1195 197.3486 130.6725 212.5064 34.353

A 765.9437 667.159 1437.1334 810.4328 2787.3311

D 25.5496 27.6076 10.4562 38.9045 1.4821

C 4.2544 2.7817 1.6761 4.2768 1.0111

INST\_SECTOR

PUBL 1042.6321 429.9046 1585.811 84.1129 949.5395

PRIV 125.8315 695.9268 69.2504 1280.2712 1904.72  
FAMI\_ING\_FMLIAR\_MENSUAL

bajo 641.8826 479.6442 1323.3315 670.9189 960.2228

alto 526.581 646.1872 331.7299 693.4652 1894.0366

Y ahora la clase  
clase

B 554.0753 175.3822 645.6957 784.1026 677.7441

C 312.4154 562.1102 630.6077 413.9 949.9667

D 302.9728 389.339 379.758 167.3816 1227.5486

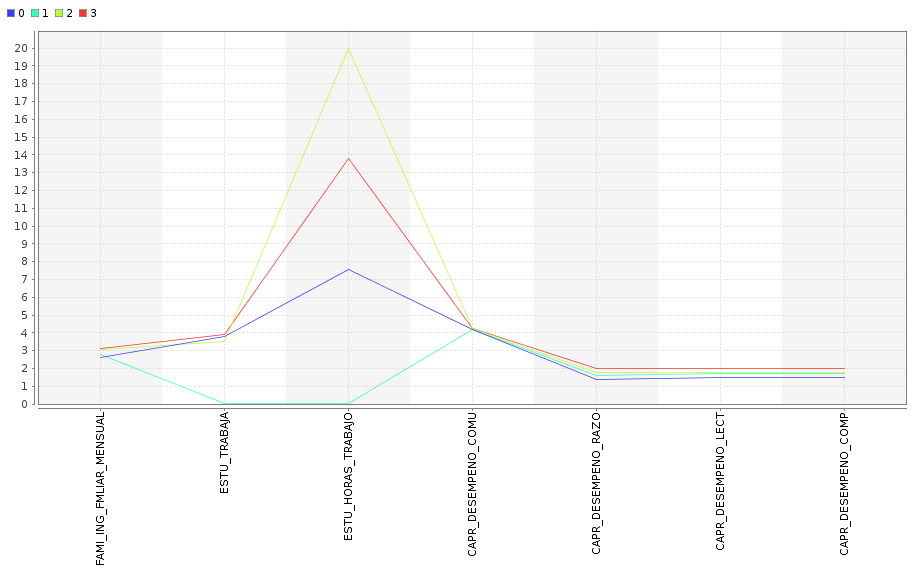
Así nos damos cuenta que variables como: INST\_COD\_INSTITUCION, ESTU\_RESIDE\_DEPT, ESTU\_HOGAR\_ACTUAL, FAMI\_NUM\_PERS\_GRUP\_FAM y ESTU\_PJE\_CREDITOS no presentan una distribución inclinada hacia un grupo, sino que están casi uniformemente distribuidas, por lo que no las enunciamos. Las eliminaremos para el paso de clasificación.

Por otro lado en el atributo INST\_CARACTERACADEMICO, el valor TECNICA PROFESIONAL se agrupa en el mismo cluster que agrupa gran parte de INSTITUCION UNIVERSITARIA, pero esta última tiene el resto de su participación conjunta a UNIVERSIDAD, por último INSTITUCION TECNOLOGICA está sola en un grupo. Esto no nos parece que revele información, por tanto vamos a suprimir esta variable para el paso de clasificación.

Las únicas variables que fueron mostradas con sus respectivos valores en cada grupo son las que más fielmente marcan la separación de los clusters, eliminando posible ruido.

Según nuestra gráfica de codos es con 4 grupos donde más obtenemos una proporción satisfactoria entre la suma del error cuadrado y el número de grupos.

Con este número de grupos obtenemos la dimensión estudiantes horas de trabajo como la que mayor discriminaba la selección de un grupo u otro.

  
Gráfico 2. Creación de clusters algoritmo EM anterior a la fase 2.

Ahora vamos a analizar la distribución de la variable *clase* luego de eliminar los atributos antes mencionados, que creemos no aportan información relevante para discriminar.

clase

B 520.4794 1195.8252 1118.6955

C 777.5351 792.5468 1296.9181

D 851.759 428.7339 1184.507

Se ha reducido el número de *clusters* a 3 y se ha logrado reducir la entropía.

Los clusters están asignados de la siguiente forma: Cluster

Attribute 0 1 2

(0.26) (0.3) (0.44)

Se escogió el 12% de datos para evaluar, estos datos quedaron así distribuidos:

Clustered Instances

0 213 ( 19%)

1 304 ( 27%)

2 596 ( 54%)

# Aplicación algoritmos de clasificación

Dadas todas las mejoras anteriormente mencionadas, se produjeron los siguientes resultados con los algoritmos de clasificación.

El algoritmo J48 fue quien mejor ofreció resultados con validación cruzada. Dadas 3 posibles clases, clasificó correctamente 46% de los datos y obtuvo un 0.63 de área bajo la curva ROC.

El número de hojas del árbol producido por el algoritmo fue 187 y el tamaño total del árbol fue de 331.

Es un árbol bastante grande en comparación al reducido número de atributos que se ha dejado para el algoritmo, dados los análisis en las etapas ya comentadas.

En total manejamos 12 atributos (ESTU\_GENERO, ESTU\_NACIMIENTO\_ANNO,ESTU\_ESTADO\_CIVIL,DISCAPACIDAD,INST\_SECTOR,INST\_VLR\_MATRICULA\_ANT,ESTU\_TOMOCURSO,ESTU\_EXAM\_ANNO\_PRESENTACION,ESTU\_SN\_CABEZA\_FMLIA,FAMI\_ING\_FMLIAR\_MENSUAL,ESTU\_TRABAJA,clase) Donde analizamos de nuevo la entropía entre el último conjunto utilizado para agrupación y este, dado que redujimos las dimensiones.

Según la fórmula 1. y los datos obtenidos en la anterior búsqueda, tenemos cerca de 134 de valor de entropía del anterior conjunto, con la reducción hecha el valor que tenemos ahora de entropía es aprox. 111. Lo que nos parece aceptable dado que mejora notoriamente el desempeño de los algoritmos de clasificación, pasando de un área bajo la curva ROC de 0.44 a 0.63.

La validación del modelo de árbol de clasificación se hizo con *cross-validation* with folds 10, así este nos arrojó 4249 instancias correctamente clasificadas, es decir: 45.8311% del conjunto total.

Entre otras medidas entregadas tenemos:

* Kappa statistic 0.188
* Mean absolute error 0.4114
* Root mean squared error 0.4616
* Relative absolute error 92.7691 %
* Root relative squared error 98.033 %

Las medidas TP Rate, FP Rate, Precision, Recall, F-Measure y ROC Area se muestran en la siguiente tabla:

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0,500 0,280 0,485 0,500 0,492 0,218 0,658 0,471 B

0,345 0,283 0,398 0,345 0,369 0,064 0,540 0,380 C

0,543 0,250 0,485 0,543 0,512 0,284 0,695 0,464 D

0,458 0,272 0,454 0,458 0,455 0,184 0,628 0,437 PROMEDIO

La matriz de confusión que obtuvimos:

=== Confusion Matrix ===

a b c <-- classified as

1600 979 622 | a = B

1142 1125 995 | b = C

Y finalmente el árbol que se produjo lo presentamos a trozos dado su tamaño.

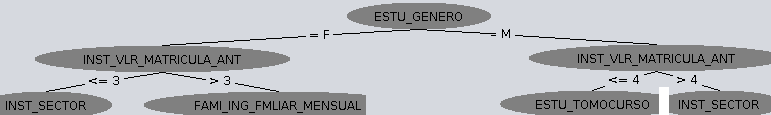
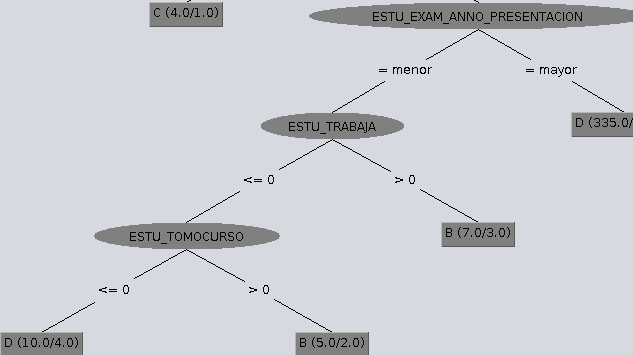


Gráfico 3. Raíz del árbol construido

  
Gráfico 4. Hojas del árbol construido, clasificando al final por atributo: *Tomo curso*.

# Conclusiones

La predicción de un factor tan complejo como lo es el puntaje que se obtiene no es una tarea fácil.

En esta interactúan muchas variables de las cuales no se tienen mediciones, y los patrones que puedan surgir son débilmente notados, por esto es normal encontrar tan baja confianza en el producto final del presente proyecto.   
Sin embargo, el ejercicio es útil para analizar relaciones que puedan surgir entre áreas: Los esfuerzos laborales que se dan paralelamente a lo académico y también familiar. No es nuevo que la situación laboral de un estudiante interfiere en sus quehaceres académicos, pero aquí se refuerza lo dicho.

Sin embargo, sería necio negar el desarrollo de otras virtudes para el estudiante trabajador, virtudes no medidas en el presente estudio.

Agradecimientos

Curso Minería de Datos, Elizabeth León, Departamento de Sistemas e Industrial. Universidad Nacional.

1. Primer Autor: wiarodriguezmo@unal.edu.co, estudiante de Ingeniería de Sistemas y Computación, Universidad Nacional de Colombia.

   Segundo Autor: yasmin.sarbaoui@gmail.com, Estudiante .... [↑](#footnote-ref-0)
2. Guía General V.2.1 - GUÍA DE ACCESO A BASES DE DATOS ICFES. VERSIÓN FEBRERO DE 2013. [↑](#footnote-ref-1)