Análisis de deserción escolar con minería de datos

José Luis Aguirre Mendiola[1], Rosa María Valdovinos Rosas[2], Juan Alberto Antonio Velazquez[1,3], Roberto Alejo Eleuterio[3], José Raymundo Marcial Romero[2]

[1] Universidad de Ixtlahuaca CUI, Ixtlahuaca de Rayón, Estado de México, México

[2] Universidad Autónoma del Estado de México, Facultad de Ingeniería, Estado de México, México

[3] Tecnológico de Estudios Superiores de Jocotitlán, Estado de México, México

Resumen. En el presente trabajo se analiza una base de datos para identificar las causas de la deserción escolar en la carrera de ingeniería en computación de la Universidad de Ixtlahuaca CUI (UICUI) a través de técnicas de minería de datos. Para ello, se aplicaron reactivos para obtener información relacionada al semestre, número de cuenta, edad, estado civil, conocer si trabaja para solventar sus estudios o cuenta con el apoyo de sus padres y forma de elección de la carrera. Una vez finalizado el proceso de minería de datos fue posible identificar algunas causas de la deserción escolar en la UICUI.

Palabras clave: minería de datos, deserción escolar, arboles de decisión.

1. Introducción

La deserción estudiantil se da cuando un número determinado de estudiantes matriculados no siguen sus estudios universitarios, bien sea por abandono producido por la insatisfacción, por repetir semestres, por causas familiares, por el grado de complejidad, entre otras. Este problema, tiene efectos de tipo financiero, académico y social que implican la pérdida de esfuerzos y recursos en donde más de la mitad de los estudiantes que comienzan una carrera universitaria no terminan sus estudios [1].

En los últimos años, la carrera de Ingeniería en computación de la Universidad de Ixtlahuaca (CUI) ha mostrado alarmantes índices de bajas temporales y totales en los primeros semestres de la carrera, con una pobre tasa de egresados.

Las causas centrales viables de estudio son cuestiones tanto familiares como económicas, además de considerar el grado de dificultad que la carrera tiene, de acuerdo al grado del uso del razonamiento lógico y matemático.

Al respecto, la minería de datos (MD) [2] proporciona una alternativa de solución para el análisis de los fenómenos no explícitos en bases de datos. Es considerada una herramienta que permite analizar grandes bases de datos de forma más detallada, facilitando la toma de decisiones. En este sentido, el objetivo principal de la MD es integrar, analizar datos y extraer modelos que forman determinados patrones a partir de los datos analizados, con los que permite obtener una descripción del comportamiento de los datos, tendencias y correlaciones [3].

La Minería de Datos se constituye la etapa de descubrimiento en el proceso de KDD [4] el cual consiste en el uso de algoritmos concretos los cuales generan una enumeración de patrones a partir de los datos anteriormente procesados, apoyándose con algoritmos de Aprendizaje Automático [2].

El estudio aquí presentado, aplica la metodología de MD con la intención de identificar vulnerabilidades en el área escolar de Ingeniería en computación para que la carrera tenga más potencial y se desarrollen mejor los profesionistas dentro de esta área que es muy demandada actualmente.

2. Caso de estudio

La Universidad de Ixtlahuaca CUI lleva un transcurso de 38 años, la escuela fue fundada en el año de 1977 iniciando originalmente como la Preparatoria Regional "Químico José Donaciano Morales", 15 años después en 1992 cuando se tuvo la necesidad de que los alumnos egresados de las escuelas preparatorias ingresaran a un nivel superior se inició la gestión para incorporar a la UAEM carreras universitarias. El 25 de Julio de 1996 el Consejo Universitario de la UAEM determinó incorporar los estudios de la licenciatura en Ingeniería en Computación y no fue sino hasta finales del 2011 que se otorga el nombre de Universidad de Ixtlahuaca CUI (http://uicui.edu.mx/).

En esta carrera ingresan cada semestre alrededor de 45 a 50 alumnos, de los cuales terminan en rededor de 20 alumnos llegando a ser siempre un sólo grupo, es difícil formar grupos de matrícula por estudiantes ya que uno de las sistemas que maneja la escuela es tomar materias de otros semestres y cursar las que repruebe para así no retrasarse en la carrera, también por eso se reducen los grupos y las matriculas pierden un orden.

La deserción que ha surgido a lo largo de la carrera ha causado que también las líneas de acentuación no sean cursadas, existen tres líneas las cuales son:

1. Administración de Proyectos Informáticos.
2. Redes y Comunicaciones.
3. Interacción Hombre-Maquina
4. Desarrollo de Software de Aplicación.

3. Desarrollo experimental

Para realizar el estudio aquí mostrado, se aplicó el proceso general de minería de datos, el universo de estudio fueron un total de 497 estudiantes del CUI, en tanto que para el análisis y construcción del modelo de minería de datos, se utilizó el software WEKA (http://www.cs.waikato.ac.nz/ml/weka/).

3.1. Adquisición de datos

Con la intención de realizar el estudio aquí mostrado, se desarrolló un instrumento para la recolección de datos (Figura 1) que se aplicó a un total de 497 alumnos de todos los semestres, con reactivos que permitan identificar las posibles causas que pudieran propiciar la deserción de los estudios universitarios.

¿Por qué elegiste la carrera?		**¿Con quién vives?**	
1.-Me gusta		1.- Papas	
2.-Me Obligaron		2.Parientes/vecinos	
3.-Ganar Dinero		3.-Solo	
¿Quién solventa tus gastos?		**¿Trabajas?**	
1.-Papa		1.-Si	
2.-Mama		2.-No	
3.-Ambos			
4.-Yo			
¿Cuál es el tiempo que trabajas?		**¿En que trabajas?**	
1.-Medio Tempo		(Pregunta abierta)	
2.-Tiempo completo			
3.-Fines de semana			
¿Qué otra carrera elegirías?			
1.-Psicologia		**Se te pide que leas cuidadosamente este cuestionario con el fin de realizar una investigación, en las preguntas con diferentes opciones, marcar con una (X), responde lo más honestamente que se te sea posible, por tu participación ¡Gracias!.**	
2.-Arquitectura			
3.-Diseño Grafico			
4.-Derecho			
5.-Administracion			
6.-Contaduria			
7.-Criminologia			
8.-Agronomia			
9.-QFB			
10.-Lenguas			
11.-Comuniacion			
12.- Otra			

Fig. 1. Formato para la aplicación de la encuesta.

3.2. Pre-procesamiento de Datos

En esta etapa se debe realizar una limpieza a los datos, i.e. obtener datos sin valores nulos o anómalos que pudieran obtener patrones de calidad. Los datos obtenidos de las encuestas, fueron analizados para identificar inconsistencia en ellos utilizando el sistema Weka. Este proceso solo se realizó para las preguntas con opción múltiple dándole valor 1 a la respuesta elegida y 0 a las demás opciones. Los atributos

seleccionados en su mayoría no contenían valores nulos ni anómalos *(outliers)*, pero en aquellos casos que se presentaban, estos fueron reemplazados utilizando técnicas estadísticas, tales como la media y la moda o derivando sus valores a través de otros, dependiendo del tipo de datos [4].

3.3. Análisis de datos

La Figura 2 muestra el análisis de frecuencias de algunas de las variables estudiadas. Como se puede observar la mayoria de los alumnos que ingresan en la carrera de Ingenieria en computacion son hombres con un 69.4% mientras que el 30.6% restante son mujeres. De éstos, la mayor parte son solteros y solo 9 son casados, lo que de algún modo indica que los compromisos familiares no pudieran ser el factor determinante para la deserción.

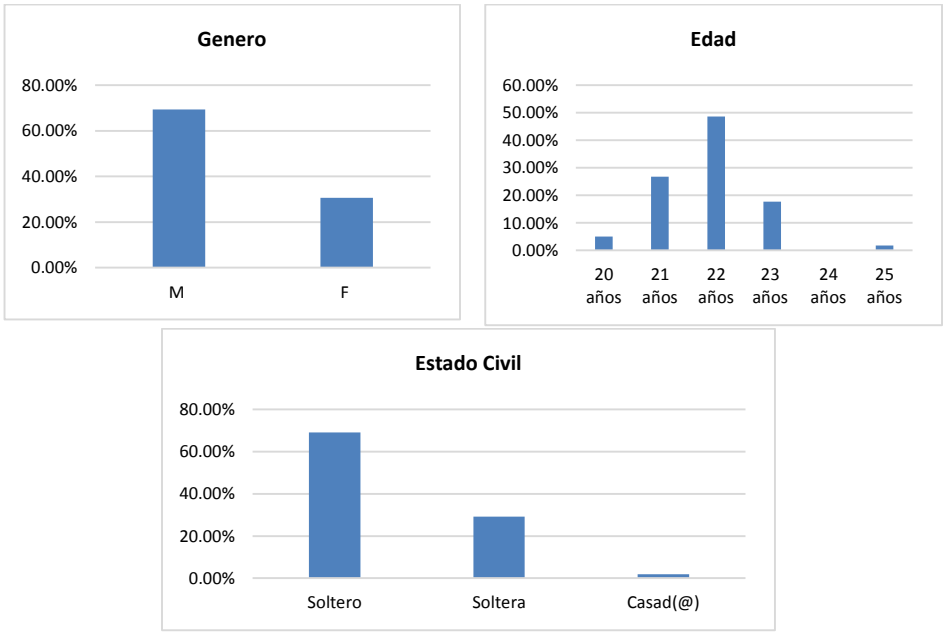

Fig. 2. Análisis de frecuencia de las variables Género, Edad y Estado civil.

Al preguntar el motivo de elección de la carrera, los encuestados manifiestan que en su mayoría (94.4%) la eligieron por agrado personal, en tanto que el porcentaje restante por considerar que sería redituable una vez finalizada.

Por otro lado, para identificar el impacto potencial que pudiera tener el aspecto económico en la continuidad de los estudios, la encuesta reveló que la mayoría de los

estudiantes no trabaja (Figura 3b), a los cuales los padres les solventan sus gastos, predominantemente ambos o el padre (Figura 3a). Respecto a los trabajos que tienen los estudiantes, es difícil establecer un predominio de éstos, ya que es muy diverso.

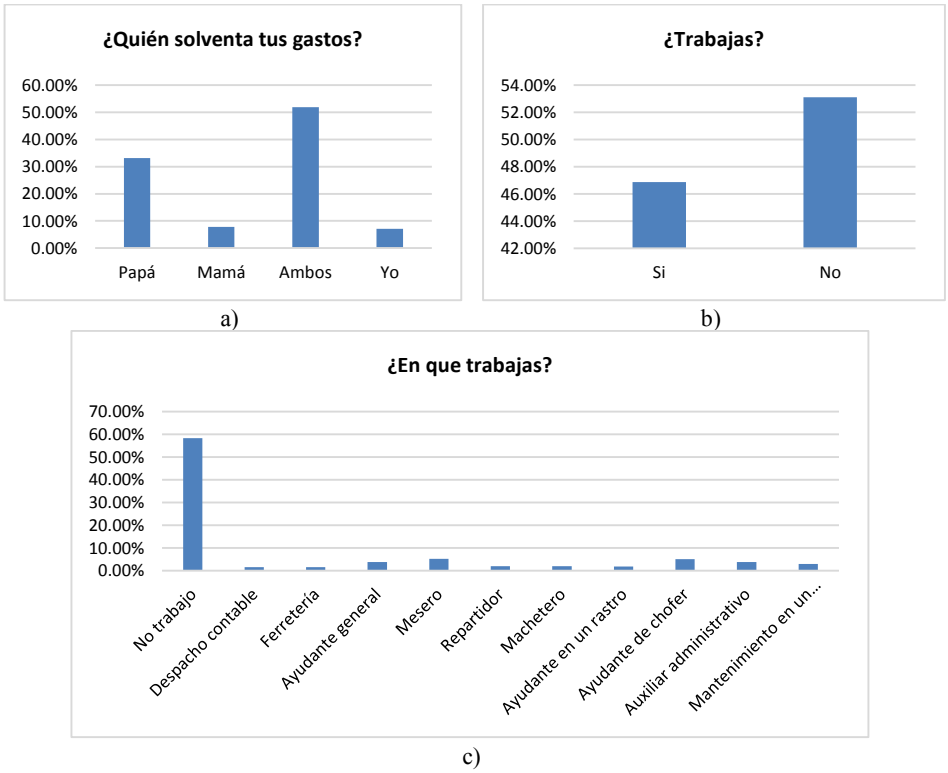

Fig. 3. Fuente de ingresos de los estudiantes encuestados.

3.4. Clasificación

Como se mencionó anteriormente, tanto para el proceso de limpieza, como para la clasificación se utilizó WEKA. Dado que los datos incluidos en el conjunto de datos son en su mayoría categóricos, se optó por utilizar árboles de decisión. De los algoritmos disponibles, se utilizó el programa J48 correspondiente al algoritmo C4.5 [5]. En su ejecución se utilizaron las especificaciones que por default tiene WEKA, así como el método de validación cruzada estratificada.

El procedimiento para generar el árbol consiste en seleccionar un atributo como raíz, y crear una rama con cada uno de los valores posibles de dicho atributo; con cada

rama resultante se realiza el mismo proceso. En cada nodo se debe seleccionar un atributo para seguir dividiendo, y para ello se selecciona aquel que mejor separe los ejemplos de acuerdo a la clase.

En el análisis de la Figura 4 se observa que los alumnos por grupo de edad que entraron a la carrera por decisión propia o porque ven beneficios económicos, de igual modo, se incluyen las formas de solventar sus gastos (por sus padres o de forma personal).

```
Gastos = PAPA: ME GUSTA (165.0)
Gastos = AMBOS: ME GUSTA (258.0/8.0)
Gastos = YO
|   Edad <= 22: GANAR DINERO (20.0)
|   Edad > 22: ME GUSTA (15.0)
Gastos = MAMA: ME GUSTA (39.0)
```

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

Correctly Classified Instances         489               98.3903 %
Incorrectly Classified Instances         8                1.6097 %
Kappa statistic                          0.8251
Mean absolute error                      0.0315
Root mean squared error                  0.1255
Relative absolute error                 29.1252 %
Root relative squared error             54.4193 %
Coverage of cases (0.95 level)          98.3903 %
Mean rel. region size (0.95 level)      50      %
Total Number of Instances              497

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.286      0.983     1         0.992      0.879    ME GUSTA
                 0.714     0          1         0.714     0.833      0.879    GANAR DINERO
Weighted Avg.    0.984     0.27       0.984     0.984     0.983      0.879
```

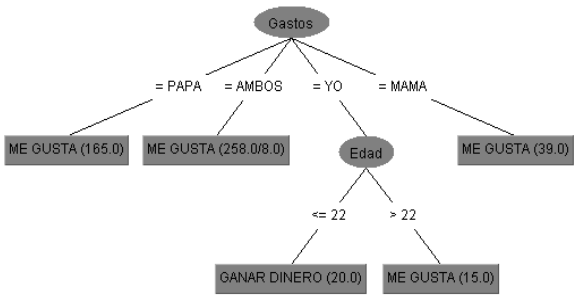

Fig. 4. Estudiantes que les gusta la carrera y quienes solventan sus propios gastos.

En el árbol resultante (Figura 4) se puede observar que, con un 1.6% de error, la predicción indica que los alumnos que solventan sus gastos de la carrera buscan

terminar la carrera para un mejor beneficio económico ya que se limita al tener que trabajar y son cuestiones por las que pueden abandonar sus estudios.

Por otro lado, el árbol de la Figura 5 muestra que los estudiantes que dependen económicamente tan solo de su mamá y su edad está en el rango de 20 a 23 años siguen solteros y los que rebasan los 23 años ya son casados pero siguen dependiendo del apoyo de su madre. Esto último permite indicar que casarse a mitad de la carrera es un factor ya que si se desea continuar con los estudios el apoyo depende de su madre, siendo esto un factor de riesgo para la deserción.

```
Genero = M
|   Gastos = PAPA: Soltero (90.0)
|   Gastos = AMBOS: Soltero (203.0)
|   Gastos = YO: Soltero (35.0)
|   Gastos = MAMA
|   |   Edad <= 23: Soltero (15.0)
|   |   Edad > 23: Casado (9.0)
Genero = F: Soltera (145.0)

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

Correctly Classified Instances         495               99.5976 %
Incorrectly Classified Instances         2                0.4024 %
Kappa statistic                          0.9909
Mean absolute error                      0.0027
Root mean squared error                  0.0464
Relative absolute error                  0.915  %
Root relative squared error             12.1387 %
Total Number of Instances              497

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.994     0          1         0.994     0.997      1        Soltero
                 1         0          1         1         1          1        Soltera
                 1         0.004      0.818     1         0.9        1        Casado
Weighted Avg.    0.996     0          0.997     0.996     0.996      1
```

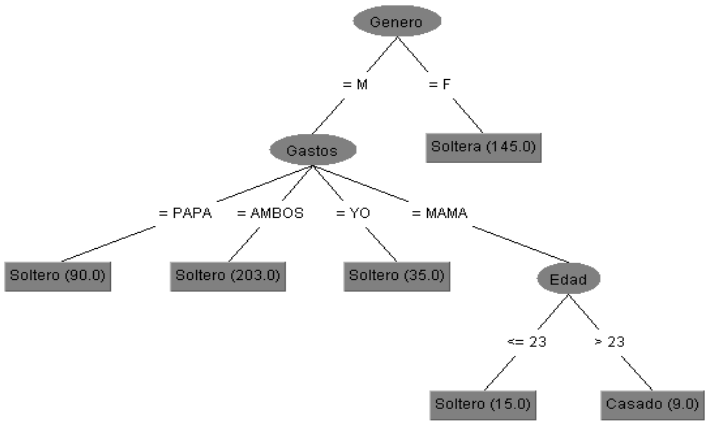

Fig. 5. Clasificación por género, estado civil y forma de solventar sus gastos.

Con la intención de analizar la relación existente entre las variables de "Estado Civil", "Trabajan" y "Carreras", el árbol de la Figura 6 muestra que de un total de 20 alumnos entre hombres y mujeres solo 11 escogerían la carrera de Diseño Gráfico, y los 9 restantes son mujeres, si escogieran la carrera de Lenguas la mayoría serian solo las mujeres y trabajan, siendo un total de 15 mujeres y 33 hombres que también estarían en la carrera pero en este caso no trabajarían y siendo el caso de que escogieran alguna otra carrera solo 119 hombres y 24 mujeres trabajarían medio tiempo, dando a entender que estando en otra carrera estarían realizando el trabajo para solventar sus gastos personales y continuar sus estudios.

```
Carreras = COMUNICACION: Soltero (44.0)
Carreras = OTRA
|   Trabajan = NO: Soltera (116.0/50.0)
|   Trabajan = SI: Soltero (132.0/37.0)
Carreras = DISEÑO GRAFICO: Soltera (20.0/11.0)
Carreras = LENGUAS
|   Trabajan = NO: Soltero (33.0)
|   Trabajan = SI: Soltera (15.0)
Carreras = QFB: Soltero (87.0)
Carreras = ARQUITECTURA: Soltera (35.0/17.0)
Carreras = CONTADURIA: Soltero (15.0)
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         371               74.6479 %
Incorrectly Classified Instances       126               25.3521 %
Kappa statistic                          0.434
Mean absolute error                      0.1891
Root mean squared error                  0.3096
Relative absolute error                 64.4932 %
Root relative squared error             80.9931 %
Total Number of Instances              497

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.816     0.331      0.846     0.816     0.831      0.836    Soltero
                 0.607     0.196      0.561     0.607     0.583      0.815    Soltera
                 0.333     0.012      0.333     0.333     0.333      0.984    Casado
Weighted Avg.    0.746     0.286      0.753     0.746     0.749      0.833
```

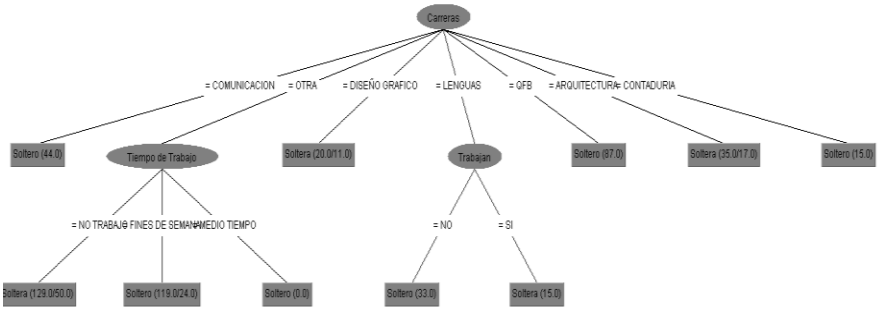

Fig. 6. Clasificación de los alumnos en caso de haber escogido otra carrera viendo en qué carrera solventarían sus gastos ellos mismos.

En el análisis de la Figura 7 se observa que los alumnos que tal vez estudiarían otra carrera serian 145 solteros que mantuvieran algún trabajo, mientras que 50 no trabajan, En el caso de las mujeres 103 mujeres tendrían algún trabajo y 37 no trabajan. Al existir un mínimo de alumnos casados, es posible indicar que el abandono de la carrera no sería especialmente por cuestiones de matrimonio, sino por otras razones que surjan en el transcurso de la carrera.

```
Carreras = COMUNICACION: SI (44.0/11.0)
Carreras = OTRA
|   Estado Civil = Soltero: SI (145.0/50.0)
|   Estado Civil = Soltera: NO (103.0/37.0)
|   Estado Civil = Casado: SI (0.0)
Carreras = DISEÑO GRAFICO: NO (20.0)
Carreras = LENGUAS
|   Estado Civil = Soltero: NO (33.0)
|   Estado Civil = Soltera: SI (15.0)
|   Estado Civil = Casado: NO (0.0)
Carreras = QFB: SI (87.0/34.0)
Carreras = ARQUITECTURA: NO (35.0)
Carreras = CONTADURIA: NO (15.0)

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

Correctly Classified Instances         365               73.4406 %
Incorrectly Classified Instances       132               26.5594 %
Kappa statistic                          0.4744
Mean absolute error                      0.3461
Root mean squared error                  0.4176
Relative absolute error                 69.4856 %
Root relative squared error             83.6843 %
Total Number of Instances              497

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.64      0.159      0.82      0.64      0.719      0.776    NO
                 0.841     0.36       0.674     0.841     0.748      0.776    SI
Weighted Avg.    0.734     0.253      0.752     0.734     0.733      0.776
```

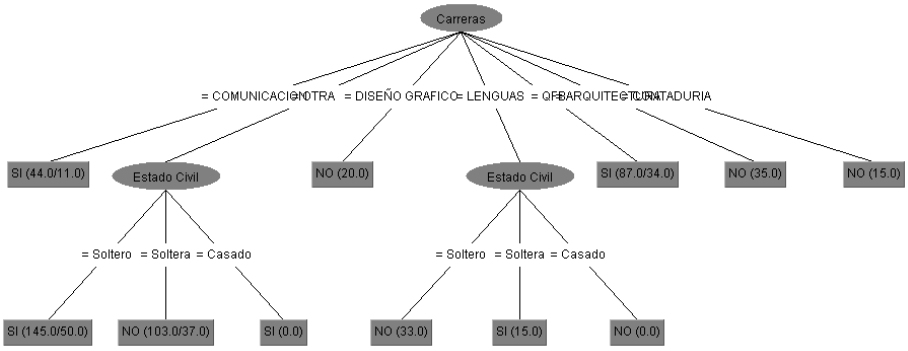

Fig. 7. Casos de haber escogido otra carrera si trabajan o no.

En el árbol de la Figura 8 se analiza la relación existente entre estudiar otra carrera y la forma de solventar sus gastos. Este diagrama muestra cómo se solventan los gastos en caso de darse de baja la carrera. Los resultados muestran con un 2.0% de error que, los gastos seguirían solventados por sus padres en la mayoría de los casos.

```
Carreras = COMUNICACION
|   Gastos = PAPA
|   |   Trabajan = NO: PAPAS (11.0)
|   |   Trabajan = SI: SOLO (19.0)
|   Gastos = AMBOS: PAPAS (14.0)
|   Gastos = YO: PAPAS (0.0)
|   Gastos = MAMA: PAPAS (0.0)
Carreras = OTRA: PAPAS (248.0)
Carreras = DISEÑO GRAFICO: PAPAS (20.0)
Carreras = LENGUAS: PAPAS (48.0)
Carreras = QFB: PAPAS (87.0/10.0)
Carreras = ARQUITECTURA: PAPAS (35.0)
Carreras = CONTADURIA: PARIENTES (15.0)
```

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

Correctly Classified Instances         487               97.9879 %
Incorrectly Classified Instances        10                2.0121 %
Kappa statistic                          0.8638
Mean absolute error                      0.0239
Root mean squared error                  0.1097
Relative absolute error                 21.2651 %
Root relative squared error             46.7918 %
Total Number of Instances              497

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.227      0.978     1         0.989      0.973    PAPAS
                 0.655     0          1         0.655     0.792      0.961    SOLO
                 1         0          1         1         1          1        PARIENTES
Weighted Avg.    0.98      0.207      0.98      0.98      0.978      0.973
```

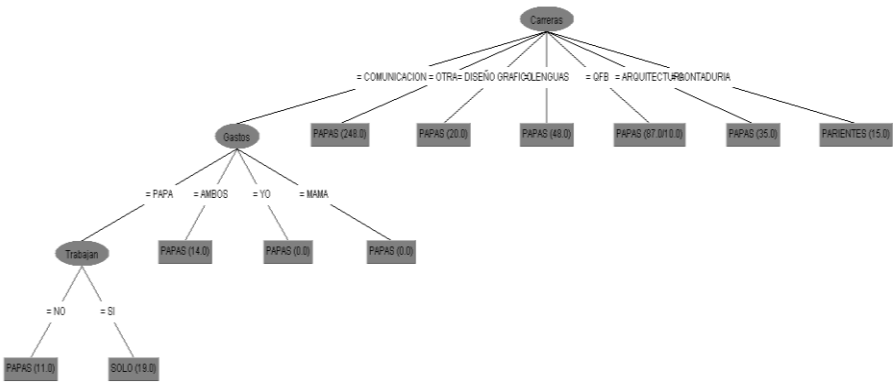

Fig. 8. Clasificación de alumnos que eligieran en caso de haber escogido otra carrera viendo en qué carrera y clasificación de gastos tanto por papá.

4. Conclusiones

Con base a los análisis realizados es posible mencionar que los problemas que afectan la deserción de alumnos en la carrera de Ing. En Computación son los siguientes:

1. Un 46% de los alumnos trabaja en el cual solo un 7% es para solventar sus propios gastos, tomando en cuenta que si los alumnos también tienen que aportar dinero a su familia casi a mediados de semestre abandonarían la carrera por seguir ayudando económicamente a su familia. No obstante, en la mayoría del total de los alumnos los padres son quienes solventas sus gastos, al parecer con independencia del hecho de que trabajen o no.

2. Los alumnos abandonan la carrera ya que también no era lo que ellos esperaban, lo cual no cumple con sus expectativas a futuro y en determinado momento deserten y escojan otra carrera que entre las que más les atraen son arquitectura, Diseño Gráfico, Lenguas, Comunicación y Contaduría.

3. La mayoría de los alumnos que ingresa a la carrera cuenta solo con el apoyo económico de su papá, su mamá o solventa sus gastos el mismo, es la cuestión por la que abandonarían la carrera ya que la carga económica en una persona suele ser muy grande y difícil, tomando en cuenta que se tienen gastos externos como luz, agua, gas entre otros.

Es fundamental se sigan buscando formas en las cuales se apoyen más a los alumnos en cuestiones economías, esto con la finalidad de que puedan terminar la carrera más alumnos y así se gradúen más ingenieros en Computación.

Agradecimientos. Este trabajo ha sido financiado parcialmente por los proyectos TESJo/CC/001 y SDMAIA-014, del Tecnológico de Estudios Superiores de Jocotitlán, y el 3834/2014/CIA de la UAEM.

Referencias

1. Rodríguez-Lagunas, J., Hernández-Vázquez, J. M.: La deserción escolar universitaria en México. La experiencia de la universidad autónoma metropolitana campus Iztapalapa. Actualidades Investigativas en Educación (2008)
2. Ian, H.W., Eibe, F.: Data mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco, CA (2005)
3. Fawcett, T., Provost, F.: Adaptive Fraud Detection. Data Mining and Knowledge Discovery, Vol. 1, No. 3, pp. 291–316 (1997)

4. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, USA (2006)
5. Quinlan, J. R.: C4.5 Programs for Machine Learning. San Mateo: Morgan Kaufmann. (1993)