INDICADOR DE GRADUACIÓN PARA ALUMNOS DEL PROGRAMA DE REVÁLIDAS AUTOMÁTICAS DE FAC. CS. EXACTAS, UNICEN

Facultad de Ciencias Exactas, Ingeniería de Sistemas
Universidad Nacional del Centro de la Provincia de Buenos Aires

ANA CAROLINA BAQUERO *baquero.carolina@gmail.com*

JUAN JOSÉ RODRÍGUEZ *jjoserodriguez@live.com*

TUTOR: Dr. Ing. Gustavo Illescas

RESUMEN

Este trabajo describe el proceso de análisis realizado durante la elaboración de un indicador para el programa de reválidas automáticas desarrollado por la Facultad de Ciencias Exactas de la UNICEN. Consiste en la evaluación de dos algoritmos diferentes para predecir la graduación o no de un alumno que se inscribe en el programa, basándose en los datos de sus actividades académicas adeudadas, la fecha de inscripción a la carrera y al plan de reválidas, entre otros datos.

El trabajo muestra la evaluación de los métodos One Rule y J48 como clasificadores de datos y cómo se decidió qué datos de entrada mantener y cuáles descartar para la predicción, además de fundamentar la elección de uno de los algoritmos de los que se evaluaron.

PALABRAS CLAVE

Ciencia de Datos, Minería de Datos, Métodos de clasificación de instancias, R, Algoritmo One Rule, Algoritmo J48.

ABSTRACT

This work describes the analysis process for the elaboration of a graduation prediction indicator for students registered on a signature revalidation program, developed by Facultad de Ciencias Exactas at UNICEN. The approach consists on the evaluation of two different algorithms to predict the ability to graduate for a student that has just signed up to the program, based on the student's academic deb, career inscription and revalidation program inscription dates, among other data.

The work shows the evaluation of One Rule and J48 algorithms using the given data by the program and how it was decided to include/exclude part of the mentioned data, while explaining why choosing one algorithm over the other.

KEY WORDS

Data Science, Data Mining, Métodos de clasificación de instancias, R, One Rule Algorithm, J48 Algorithm.

INTRODUCCIÓN

Este trabajo se plantea en el contexto de la existencia de un programa de reválidas automáticas para la carrera Ingeniería de Sistemas, dictada en la Facultad de Cs. Exactas de la Universidad Nacional del Centro de la Pcia. de Bs. As. Para todos los alumnos de esta carrera, las materias aprobadas pierden su validez al cumplirse diez años de rendido el examen final, momento en el cual es necesario realizar una revalidación de la materia para poder seguir avanzando y obtener el título de grado. Muchas veces, el profesor a cargo, puede pedir que la materia se vuelva a cursar, se vuelva a rendir el examen final o se realice un trabajo práctico similar al trabajo de cursada de ese año. En otros casos la revalidación sucede sin pedir requisito alguno al alumno más que la solicitud de reválida y un certificado o carta que justifique la demora en graduarse.

Desde el año 2013, la carrera cuenta con un Plan de Reválidas Automáticas, en respuesta a una necesidad concreta de los estudiantes que han postergado su graduación o abandonado la carrera y quieren retomarla para poder graduarse sin tener que rendir nuevamente estas materias vencidas.

El plan de reválidas automáticas permite a un alumno inscribirse adeudando una cantidad límite de actividades académicas, y comprometiendo un plan para rendir las actividades que le restan. Una vez inscripto, el alumno tiene un período de 2 (dos) años para recibirse y, finalizado este plazo, vuelve a la antigua modalidad de revalidación de materias. En algunos casos, pueden realizarse excepciones y tramitar una extensión del plazo para recibirse.

Hasta el momento, se han lanzado 2 ediciones del Programa de Reválidas, una en el período que va de Diciembre de 2013 a Diciembre de 2015 y otra que va desde Diciembre de 2015 a Diciembre de 2017.

El objetivo de este trabajo es evaluar los datos históricos de los planes de reválidas con el fin de predecir, para los inscriptos en futuras ediciones, la factibilidad que tienen de graduarse en el plazo otorgado de 2 (dos) años. Este indicador también serviría para evaluar si se le debería sugerir al alumno la inscripción al programa de tutorías para que sea guiado en el proceso.

En adelante, llamaremos R1 al primer período correspondiente a la primera edición del plan de reválidas (dic 2013 a dic 2015) y R2 al segundo período (dic 2015 a dic 2017).

Inicialmente, el plan R1 permitía a un alumno inscribirse adeudando 14 créditos. Actualmente (R2) el Plan de Reválidas es otorgado a alumnos que adeuden 9 créditos o menos, a los que se les da un plazo de 2 años para que puedan finalizar su carrera.

DESARROLLO

PROCEDIMIENTO

Como sugerencia de la cátedra, se tomó como base un trabajo similar [1] realizado para la materia Introducción a la Ciencia de Datos, donde se analizó la correctitud del orden de mérito elaborado en el pasado para el programa Delta G, de la misma unidad académica y carrera de grado de la UNICEN.

En dicho trabajo se propone un procedimiento que se reutilizó en este caso para el análisis de los datos con distintos algoritmos:

1. Se identificó el problema tal cual como se detalla en la introducción.

2. Se procesaron los datos recolectados durante las ediciones R1 y R2 (sin terminar) del programa de reválidas automáticas.

   Para el procesamiento de los datos se empleó un procedimiento ETL (Extract, Transform and Load) para facilitar la carga y transformación de los datos en una base de datos PostgreSQL para luego su posterior análisis con comandos de R [2].

   - En la **etapa de extracción** se creó una Base de Datos en PostgreSQL y se creó una tabla (reválidas) para almacenar los datos. Se utilizó el comando COPY de PostgreSQL para la carga de los datos en la tabla.

   - En la **etapa de transformación** se adaptaron los datos para poder volcarlos en la base de datos.

   - En la **etapa de carga** se insertaron los datos en la tabla reválidas, para analizarlos con R.

3. Se instaló y configuró CRan [3] en Windows para correr comandos de R desde una consola contra PostgreSQL. Se instalaron los paquetes “OneR” y “RWeka” lo cual permitió escribir el código necesario para correr los distintos algoritmos a evaluar.

4. Se analizaron los algoritmos propuestos (OneR y J48) y se corrieron con los datos proporcionados.

   Una vez obtenidos los resultados, se realizaron diversos análisis estadísticos exploratorios (gráficos, cálculo de medidas de asociación, etc.) con el paquete estadístico-gráfico abierto y libre R (https://www.r-project.org/) y se generaron distintos modelos de clasificación.

   En particular se apuntó a detectar qué variables o índices mostraron mayor vinculación con la posterior obtención o no del título, a la vez que se buscaron modelos descriptivos sencillos que permitieran predecir dicha obtención con un nivel de precisión aceptable.

DATOS

Se analizaron los datos de 124 alumnos inscriptos en las fases R1 y R2 del programa. Estos datos fueron provistos por la Comisión de Tutorías y reválidas de la Facultad de Ciencias Exactas, se sugirieron incorporaciones de datos y se dejaron de lado otros, a partir del análisis de los resultados.

Existieron algunos casos en los cuales los alumnos ya no debían actividades académicas pero tenían vencidos los finales y se inscribieron para poder obtener las reválidas y tramitar su título. Estos casos no se tuvieron en consideración, dado que no aportaban al análisis y predicciones que se intentan realizar en este trabajo, quedando un total de 117 registros útiles para el análisis de este trabajo.

1. DATOS INICIALES

Nro	Columna en BD	Descripción
1	nro_solicitud	Nro. de solicitud de inscripción al plan. Utilizado para que la comisión de tutorías pueda asociar la solicitud a un alumno y poder usar los datos de forma anónima en este análisis
2	año_ingreso	Año de ingreso a la carrera
3	finales_adeudados	Finales adeudados al momento de la inscripción al plan
4	cursadas_adeudadas	Cursadas adeudados al momento de la inscripción al plan
5	optativas_adeudadas	Optativas adeudados al momento de la inscripción al plan
6	adeuda_trab_final_inicio	Valor booleano (True/false) que indica si el alumno adeuda el trabajo final (Tesis) al momento de la inscripción
7	plan	Plan de carrera al que pertenece (Ej 1995, s95)
8	tutor_r1	Si tuvo tutor asignado en la edición r1 del plan.
9	tutor_r2	Si tuvo tutor asignado en la edición r1 del plan.
10	fecha_graduación	Fecha de graduación del alumno, en caso de haber completado la carrera.

2. DATOS EXCLUIDOS O DERIVADOS

Algunos datos fueron omitidos dado que no resultaban de relevancia para el análisis y otros resultaron de derivaciones de los originales.

El atributo *“año de ingreso”*, se reemplazó por la cantidad de años transcurridos desde el ingreso hasta el momento de la inscripción, dado que una cantidad de años resulta ser más representativo de la situación de un alumno particular al momento de inscribirse en el año “X”, que el año calendario de ingreso.

Desde la comisión de tutorías se había sugerido que los atributos *tutor_r1* y *tutor_r2* podrían no aportar información correcta para la predicción de valores de graduado, dado que en R1 la mayoría de los alumnos se inscribió con tutor sin saber por qué y de hecho no lo usaron. En cambio, en R2 pareciera haber menos alumnos con tutor y algunos alumnos que se re-inscribieron luego de R1 en R2 cambiaron su condición respecto de este valor. Se decidió analizar los resultados de los métodos con y sin estos atributos para decidir.

3. DATOS AGREGADOS

Fue necesario recolectar los siguientes datos para poder predecir un resultado:

- **graduado**: indica si el alumno se graduó o no. Es el atributo más importante, dado que es necesario para analizar el éxito del alumno en el plan de reválidas.

Se sugirió la incorporación de los siguientes datos para mejorar la precisión del indicador:

- **fecha de inscripción** al plan de reválidas
- **fecha de la última actividad académica** previa a la inscripción. Se utiliza el atributo derivado “meses desde la última actividad académica”.
- **fecha_graduación**: fecha en que el alumno se graduó, si lo hizo

4. DATOS UTILIZADOS

Nro	Columna	Descripción
1	nro_solicitud	Nro de solicitud de inscripción al plan.
2	fecha_inscripcion	Fecha de inscripción al plan de reválidas
3	año_ingreso=> dif_años_ingreso	Del año calendario de ingreso, se derivó la cantidad de años transcurridos desde que el alumno se inscribió en la carrera hasta que se inscribe al plan de reválidas. Se implementaron distintos rangos de 2, 3 ,4 y 5 años.
4	finales_adeudados	Finales adeudados al momento de la inscripción al plan
5	cursadas_adeudadas	Cursadas adeudados al momento de la inscripción al plan

6	optativas_adeudadas	Optativas adeudados al momento de la inscripción al plan
7	adeuda_trab_final_inicio	Valor booleano (True/false) que indica si el alumno adeuda el trabajo final (Tesis) al momento de la inscripción
8	plan	Plan de carrera al que pertenece (Ej. 1995, s95)
9	graduado	Valor booleano que indica si el alumno se graduó estando anotado en el plan. Se extrajo por deducción a partir de la fecha de graduación, siendo true cuando ésta no era nula.

MÉTODOS DE CLASIFICACIÓN

1. CLASIFICACIÓN DE BASE

El modelo de clasificación (o clasificador) más sencillo es aquel que asigna a cada individuo el resultado que más se repite: en este caso, consistiría predecir "se graduará en el transcurso de 2 (dos) años" para todos los individuos.

Este tipo de clasificador básico no utiliza la información disponible de las variables y por lo tanto su poder de predicción es el mínimo; sin embargo, permite obtener un valor de referencia para evaluar la precisión de la performance de modelos más complejos. Cualquier modelo de clasificación alternativo debería entonces igualar o superar el nivel de precisión de esta clasificación para ser considerado útil.

De los 117 estudiantes considerados, 47 lograron finalizar su carrera en el período de vigencia del programa. Así, predecir que un estudiante cualquiera se graduará de manera mecánica lograría una clasificación correcta en el **40,17%** de los casos. Modelos alternativos de predicción del resultado deberán al menos alcanzar la misma precisión.

2. ALGORITMO “ONE R”

2.1 Descripción

“OneR” viene de “One Rule”, es un algoritmo de clasificación que genera un árbol de decisión de un único nivel [4]. OneR es capaz de inferir reglas de clasificación a partir de un conjunto de instancias [5]. El algoritmo crea una regla para cada atributo en los datos de entrenamiento, luego escoge la regla con la *tasa de error* más pequeño como su "one rule". Para crear una regla para cada atributo debe determinarse la clase más frecuente para cada valor del atributo.

Las desventajas que podemos advertir en este algoritmo son:

i) El algoritmo trata todos los atributos numéricamente evaluados como continuos, usa un método directo para dividir el rango de valores en intervalos disjuntos. Esto introduce un riesgo de "overfitting" en el caso de atributos evaluados de forma continua, por ejemplo: números de teléfono, etc…

ii) El "overfitting" de atributos nominales con valores únicos tales como nombres de personas, direcciones de correo electrónico, etc…

iii) Selección aleatoria de un atributo cuando las tasas de error son iguales.

iv) Selección aleatoria de una clase cuando dos o más clases dan la misma tasa de error con un atributo.

Overfitting (o sobreajuste): es el efecto de sobreentrenar un algoritmo de aprendizaje con unos ciertos datos para los que se conoce el resultado deseado. Cuando un sistema se entrena demasiado (se sobreentrena) o se entrena con datos extraños, el algoritmo de aprendizaje puede quedar ajustado a unas características muy específicas de los datos de entrenamiento que no tienen relación causal con la función objetivo.

Accuracy: porcentaje de precisión de la predicción.

2.1 Pseudocódigo

Para cada atributo o "predictor"
Para cada valor del atributo, crear una regla como sigue:
Contar las apariciones de cada valor del target (atributo a predecir: recibido)
Seleccionar la clase más frecuente
Hacer que la regla asigne esa clase a este valor del atributo o "predictor"
Calcular el error total de las reglas de cada atributo predictor
Elegir el atributo "predictor" con el menor error total

2.2 Ejemplo

Al correr el algoritmo OneR con los datos recolectados, se calcula la precisión de cada atributo y en base a eso, se realiza la selección del atributo a partir del cual se van a generar las reglas.

A continuación se muestra, a modo de ejemplo, el análisis realizado para los dos atributos con más precisión dentro del conjunto de atributos (para ver todos los resultados ir más adelante, a la sección "Resultados").

Análisis de la precisión para el atributo adeuda_trabajo_final_inicio

adeuda_trab_final_inicio\recibido	1	0	Precisión
1	39	63	0,6290322581
0	15	7	

La precisión de este atributo es **0,6290322581** = (63 + 15) / 124

Análisis de la precisión para Finales adeudados

Finales_adeudados\Recibido	1	0	Precisión
0-1	51	47	0,6048387097
2	2	12	
3	0	6	
4	0	5	
5	1	0	

La precisión de este atributo es **0,6048387097** = (51 + 12 + 6 + 5 + 1) / 124

Criterio de selección

El atributo adeuda_trabajo_final_inicio tiene una precisión mayor, 0,629..., que finales_adeudados, 0,604. (Es lo mismo que decir que tiene menor tasa de fallo que finales_adeudados) por lo que el algoritmo escoge adeuda_trabajo_final_inicio.

Esto se hace con todos los atributos y se selecciona el que menor tasa de error tenga. (O mayor precisión). En base al atributo seleccionado, se generan las reglas, basado en los valores que éste puede tomar.

Como se mencionó anteriormente, luego de comparar todos los atributos, se encuentra que el atributo con mejor porcentaje de precisión es **adeuda_trabajo_final** y las reglas generadas para este atributo son:

If adeuda_trab_final_inicio = FALSE then recibido = TRUE
If adeuda_trab_final_inicio = TRUE then recibido = FALSE

3. ALGORITMO “J48”

3.1 Descripción

Este algoritmo es una implementación de WEKA del algoritmo C45 [6] y construye árboles de decisión desde un grupo de datos de entrenamiento usando el concepto de *entropía* de información. En cada nodo del árbol, el algoritmo elige el atributo de los datos que más eficazmente divida el conjunto de muestras en subconjuntos enriquecidos en una clase u otra. Su criterio es el normalizado para ganancia de información (diferencia de entropía) que resulta en la elección de un atributo para dividir los datos. El atributo con la mayor ganancia de información normalizada se elige como

parámetro de decisión. El algoritmo divide el conjunto de datos recursivamente en sub-listas más pequeñas.

Entropía: se puede considerar como la cantidad de información promedio que contienen los símbolos usados. Los símbolos con menor probabilidad son los que aportan mayor información; por ejemplo, si se considera como sistema de símbolos a las palabras en un texto, palabras frecuentes como «que», «el», «a» aportan poca información, mientras que palabras menos frecuentes como «corren», «niño», «perro» aportan más información. Cuando todos los símbolos son igualmente probables (distribución de probabilidad plana), todos aportan información relevante y la entropía es máxima.

3.2 Ejemplo

Una rama -> finales_adeudados = 0
| plan = 88: => FALSE (8.0/1.0)

Quiere decir que si no se adeudan finales y pertenece al plan 88, se predice que no se van a recibir y una sola predicción de cada 8 falla. (7/8 = 0,875, se predice con un 87,5% de acierto)

Como los otros atributos no suman información a esta rama, se corta ahí.

RESULTADOS

1. DE LOS ATRIBUTOS UTILIZADOS

En un primer momento, se ejecutaron los métodos de clasificación con todos los atributos a considerar, para poder realizar un primer análisis de las predicciones.

Se incluyeron los siguientes datos para cada alumno:

- finales_adeudados
- cursadas_adeudadas
- optativas_adeudadas
- adeuda_trab_final_inicio
- plan
- año_ingreso
- graduado

Para el atributo *plan*, se discutió el hecho de que la mayoría de los inscriptos tenía plan S95 y el resto no era un porcentaje significativo de alumnos, respecto del total. Se estima que la mayoría de los casos futuros van a tener plan S95 en adelante.

Si bien el dato no va a aportar mucho en adelante, cuando sólo se inscriban alumnos del plan S95, fue necesario usarlo para que se clasificaran bien las instancias históricas correspondientes a planes viejos y no perder precisión en los resultados. Además, se supone que en algún momento en el futuro podrían llegar a existir nuevos planes.

El uso del atributo *año de ingreso* fue analizado con distintas variantes, y en one Rule dio mejores resultados ya que el aporte del valor resultó más significativo. Finalmente, en One Rule fue reemplazado por la cantidad de años transcurridos entre el ingreso y la inscripción al plan de reválidas (dividiendo los valores en rangos de 4 años), mientras que en J48 se descartó.

2. RESULTADOS DE ONE RULE

En esta sección se detallarán y analizarán los resultados obtenidos para todos los atributos considerados con el método One Rule.

2.1. Resultados Iniciales

	Atributo	Precisión
1 *	**finales_adeudados**	**60.68%**
1	**adeuda_trab_final_inicio**	**60.68%**
3	año_ingreso	59.83%
3	plan	59.83%
3	cursadas_adeudadas	59.83%
3	optativas_adeudadas	59.83%
3	tutor_r1	59.83%
3	tutor_r2	59.83%

Atributo seleccionado dada la precisión: '*' (finales_adeudados)

Reglas (finales_adeudados):

If finales_adeudados = (-0.005,1] then graduado = FALSE
If finales_adeudados = (1,2] then graduado = FALSE
If finales_adeudados = (2,3] then graduado = FALSE
If finales_adeudados = (3,4] then graduado = FALSE
If finales_adeudados = (4,5] then graduado = TRUE

Precisión:
71 de 117 instancias clasificadas correctamente (60,68%)

2.2. Análisis de Resultados

2.2.1. Atributo finales_adeudados

A partir de los resultados de la corrida inicial, se analizó la selección del atributo ***finales_adeudados*** (podría haber sido *adeuda_trab_final_inicio* porque tienen el mismo porcentaje) como el de mejor precisión para predecir si un alumno se graduará o no.

One Rule nos dice en este caso, que puede predecir si un alumno va a graduarse basándose en la cantidad de finales adeudados tiene el alumno al momento de inscribirse con una certeza del **60.68%.**

2.2.2. Atributo adeuda_trabajo_final_inicio

El atributo *adeuda_trabajo_final_inicio* presentó la misma precisión que *finales_adeudados*, por lo cual se muestran también las reglas generadas a partir de éste:

Reglas:

If adeuda_trab_final_inicio = FALSE then graduado = TRUE
If adeuda_trab_final_inicio = TRUE then graduado = FALSE

One Rule nos dice en este caso, que puede predecir si un alumno va a graduarse basándose en el hecho de si debe el trabajo final (tesis) o no al inscribirse al plan con una certeza del **60.68%.** Con esta certeza, predice que un alumno que adeude la tesis no va a graduarse, mientras que uno que no la deba sí lo va a hacer.

2.2.3. Atributo año_ingreso

En el caso particular del atributo año de ingreso, que quedó en tercer lugar, se pensó que tomar el año calendario no era un dato muy representativo. Si bien en este contexto el porcentaje de certeza (59.83%) es superior al clasificador de base (40.17%), se presume que los datos históricos para los años más viejos no van a ser útiles más adelante dado que los alumnos que se inscriban van a ser de años posteriores; mientras que los años más recientes no tendrán suficiente información histórica para predecir a priori si el alumno va a graduarse o no.

Por este motivo se decidió analizar el comportamiento cuando se trataba como la cantidad total de años transcurridos desde el ingreso. En una segunda instancia, se analizó el comportamiento clasificando esta diferencia de años en distintos rangos.

A continuación se detalla el análisis realizado para los distintos valores considerados para el atributo ***dif_año_ingreso*.**

1.1 Diferencia en años como valor absoluto

dif_año_ingreso = año inscripción al plan de reválidas - año ingreso

Atributo		Precisión
1 * dif_años_ingreso		***60.68%***
1	**finales_adeudados**	**60.68%**
1	**adeuda_trab_final_inicio**	**60.68%**
4	plan	59.83%
4	cursadas_adeudadas	59.83%
4	optativas_adeudadas	59.83%
4	tutor_r1	59.83%
4	tutor_r2	59.83%

Reglas (dif_año_ingreso):

If dif_años_ingreso = (9.98,13.8] then graduado = FALSE
If dif_años_ingreso = (13.8,17.6] then graduado = FALSE
If dif_años_ingreso = (17.6,21.4] then graduado = TRUE
If dif_años_ingreso = (21.4,25.2] then graduado = FALSE
If dif_años_ingreso = (25.2,29] then graduado = FALSE

Se observa que, al dejar de considerar el año calendario de ingreso, el nuevo atributo seleccionado por OneR puede ser *dif_años_ingreso*; dado que iguala la precisión de *finales_adeudados* y *adeuda_trab_final_inicio.* Se puede decir, entonces, que es mejor tomar la diferencia de años que el año calendario.

1.2 Diferencia en años clasificado en rangos de X= 2, 3, 4 y 5 años

dif_año_ingreso = Floor((año inscripción al plan de reválidas - año ingreso) / X)

	Atributo	Precisión
*1 *	***dif_años_ingreso_rangos_de_4***	***62.39%***
2	*dif_años_ingreso_total*	*61.54%*
2	*dif_años_ingreso_rangos_de_2*	*61.54%*
4	*dif_años_ingreso_rangos_de_3*	*60.68%*
4	*dif_años_ingreso_rangos_de_5*	*60.68%*
4	finales_adeudados	60.68%
4	adeuda_trab_final_inicio	60.68%
8	plan	59.83%
8	cursadas_adeudadas	59.83%
8	optativas_adeudadas	59.83%
8	tutor_r1	59.83%
8	tutor_r2	59.83%

Reglas generadas para dif_años_ingreso_rangos_de_4:

If dif_años_ingreso_rangos_de_4 = (0.995,2] then graduado = FALSE
If dif_años_ingreso_rangos_de_4 = (2,3] then graduado = TRUE
If dif_años_ingreso_rangos_de_4 = (3,4] then graduado = FALSE
If dif_años_ingreso_rangos_de_4 = (4,5] then graduado = FALSE
If dif_años_ingreso_rangos_de_4 = (5,6] then graduado = FALSE

El rango de diferencia de años se define como año inscripción – año ingreso, se toma la parte entera y se arma la siguiente tabla

Valor de dif_años		Rango de años
1	=>	[10;11]
2	=>	[12;15]
3	=>	[16;19]
4	=>	[20;24]
5	=>	[25;29]
6	=>	[30;34]
7	=>	[35;39]

Se observa que al correr OneR con las distintas opciones de rangos para los años transcurridos desde el ingreso, la de mayor precisión fue la clasificación con rangos de 4 años, siguiéndole la de diferencia de años total y los rangos de 2 años. Los rangos de 3 y 5 años fueron los de menor precisión. De estos últimos resultados, se puede concluir que es mejor tomar la diferencia de años en rangos de cuatro (4), que tomando de a un (1) año.

Conclusiones de One Rule

Como conclusión del análisis sobre los atributos a incluir para la generación de reglas con One Rule, se puede decir que si bien el atributo *dif_años_ingreso_rangos_de_4* es el que mayor precisión presenta, se cree que no es conveniente utilizarlo a futuro. A medida que pasen los años se seguirán agregando valores posibles de *año de ingreso* aumentando la cantidad de reglas considerablemente.

Los siguientes clasificadores más precisos fueron *finales adeudados* y *adeuda trabajo final al inicio*, los cuales parecen más adecuados para este contexto ya que son actividades académicas adeudadas por el alumno.

Por último, se puede observar que los atributos plan, cursadas adeudadas, optativas adeudadas, tutor_r1 y tutor_r2, son los que menos precisión tienen para generar reglas con One Rule, presentando la misma precisión.

3. RESULTADOS DE J48

3.1. Resultados Iniciales

Inicialmente, se ejecutó el algoritmo J48 con todos los atributos brindados por la comisión de tutorías:

1. año_ingreso,
2. plan,
3. finales_adeudados,
4. cursadas_adeudadas,
5. optativas_adeudadas,
6. adeuda_trab_final_inicio,
7. tutor_r1,
8. tutor_r2,
9. graduado

Árbol de decisión generado

```
plan = 1995: FALSE (7.0)
plan = 88
|   año_ingreso = 1986: TRUE (1.0)
|   año_ingreso = 1987: FALSE (1.0)
|   año_ingreso = 1988: FALSE (1.0)
|   año_ingreso = 1989: FALSE (3.0)
|   año_ingreso = 1991: FALSE (3.0)
```

```
|   año_ingreso = 1992: FALSE (1.0)
|   año_ingreso = 1993: FALSE (0.0)
|   año_ingreso = 1994: FALSE (1.0)
|   año_ingreso = 1995: FALSE (0.0)
|   año_ingreso = 1996: FALSE (0.0)
|   año_ingreso = 1997: FALSE (0.0)
|   año_ingreso = 1998: FALSE (0.0)
|   año_ingreso = 1999: FALSE (0.0)
|   año_ingreso = 2000: FALSE (0.0)
|   año_ingreso = 2001: FALSE (0.0)
|   año_ingreso = 2002: FALSE (0.0)
|   año_ingreso = 2003: FALSE (0.0)
|   año_ingreso = 2004: FALSE (0.0)
|   año_ingreso = 2005: FALSE (0.0)
|   año_ingreso = 2006: FALSE (0.0)
plan = S95
|   finales_adeudados = 0
|   |   año_ingreso = 1986: TRUE (0.0)
|   |   año_ingreso = 1987: TRUE (0.0)
|   |   año_ingreso = 1988: TRUE (0.0)
|   |   año_ingreso = 1989: TRUE (0.0)
|   |   año_ingreso = 1991: TRUE (0.0)
|   |   año_ingreso = 1992: FALSE (1.0)
|   |   año_ingreso = 1993: TRUE (1.0)
|   |   año_ingreso = 1994: TRUE (1.0)
|   |   año_ingreso = 1995: TRUE (1.0)
|   |   año_ingreso = 1996
|   |   |   tutor_r1 = FALSE: FALSE (1.0)
|   |   |   tutor_r1 = TRUE: TRUE (3.0/1.0)
|   |   año_ingreso = 1997: TRUE (3.0)
|   |   año_ingreso = 1998
|   |   |   adeuda_trab_final_inicio = FALSE: TRUE (2.0)
|   |   |   adeuda_trab_final_inicio = TRUE
|   |   |   |   tutor_r1 = FALSE: TRUE (1.0)
|   |   |   |   tutor_r1 = TRUE: FALSE (1.0)
|   |   año_ingreso = 1999: TRUE (6.0/1.0)
|   |   año_ingreso = 2000
|   |   |   optativas_adeudadas = 0: FALSE (5.0/1.0)
|   |   |   optativas_adeudadas = 1: FALSE (0.0)
|   |   |   optativas_adeudadas = 2: TRUE (1.0)
|   |   |   optativas_adeudadas = 3: FALSE (0.0)
|   |   |   optativas_adeudadas = 4: FALSE (0.0)
|   |   |   optativas_adeudadas = 5: FALSE (0.0)
|   |   año_ingreso = 2001
|   |   |   tutor_r1 = FALSE
|   |   |   |   optativas_adeudadas = 0: FALSE (3.0/1.0)
```

```
|   |   |   optativas_adeudadas = 1: FALSE (0.0)
|   |   |   optativas_adeudadas = 2: TRUE (1.0)
|   |   |   optativas_adeudadas = 3: FALSE (0.0)
|   |   |   optativas_adeudadas = 4: FALSE (0.0)
|   |   |   optativas_adeudadas = 5: FALSE (0.0)
|   |   |   tutor_r1 = TRUE: TRUE (6.0)
|   |   año_ingreso = 2002
|   |   |   optativas_adeudadas = 0
|   |   |   |   adeuda_trab_final_inicio = FALSE: TRUE (1.0)
|   |   |   |   adeuda_trab_final_inicio = TRUE
|   |   |   |   |   tutor_r1 = FALSE: TRUE (7.0/3.0)
|   |   |   |   |   tutor_r1 = TRUE: FALSE (4.0/1.0)
|   |   |   optativas_adeudadas = 1: FALSE (1.0)
|   |   |   optativas_adeudadas = 2: FALSE (2.0)
|   |   |   optativas_adeudadas = 3: FALSE (0.0)
|   |   |   optativas_adeudadas = 4: FALSE (0.0)
|   |   |   optativas_adeudadas = 5: FALSE (0.0)
|   |   año_ingreso = 2003: TRUE (9.0/2.0)
|   |   año_ingreso = 2004
|   |   |   tutor_r1 = FALSE: FALSE (1.0)
|   |   |   tutor_r1 = TRUE: TRUE (1.0)
|   |   año_ingreso = 2005: FALSE (2.0)
|   |   año_ingreso = 2006: TRUE (1.0)
|   finales_adeudados = 1
|   |   cursadas_adeudadas = 0: FALSE (5.0)
|   |   cursadas_adeudadas = 1
|   |   |   optativas_adeudadas = 0
|   |   |   |   año_ingreso = 1986: FALSE (0.0)
|   |   |   |   año_ingreso = 1987: FALSE (0.0)
|   |   |   |   año_ingreso = 1988: FALSE (0.0)
|   |   |   |   año_ingreso = 1989: FALSE (0.0)
|   |   |   |   año_ingreso = 1991: FALSE (0.0)
|   |   |   |   año_ingreso = 1992: FALSE (0.0)
|   |   |   |   año_ingreso = 1993: FALSE (0.0)
|   |   |   |   año_ingreso = 1994: FALSE (0.0)
|   |   |   |   año_ingreso = 1995: FALSE (0.0)
|   |   |   |   año_ingreso = 1996: FALSE (0.0)
|   |   |   |   año_ingreso = 1997: FALSE (0.0)
|   |   |   |   año_ingreso = 1998: TRUE (1.0)
|   |   |   |   año_ingreso = 1999: FALSE (0.0)
|   |   |   |   año_ingreso = 2000: FALSE (0.0)
|   |   |   |   año_ingreso = 2001: FALSE (0.0)
|   |   |   |   año_ingreso = 2002: FALSE (0.0)
|   |   |   |   año_ingreso = 2003: FALSE (3.0/1.0)
|   |   |   |   año_ingreso = 2004: FALSE (0.0)
|   |   |   |   año_ingreso = 2005: FALSE (0.0)
```

```
|   |   |   |   año_ingreso = 2006: FALSE (0.0)
|   |   |   optativas_adeudadas = 1: FALSE (1.0)
|   |   |   optativas_adeudadas = 2: FALSE (1.0)
|   |   |   optativas_adeudadas = 3: FALSE (0.0)
|   |   |   optativas_adeudadas = 4: FALSE (0.0)
|   |   |   optativas_adeudadas = 5: FALSE (0.0)
|   |   cursadas_adeudadas = 2: FALSE (0.0)
|   |   cursadas_adeudadas = 4: FALSE (0.0)
|   finales_adeudados = 2
|   |   adeuda_trab_final_inicio = FALSE: TRUE (1.0)
|   |   adeuda_trab_final_inicio = TRUE
|   |   |   tutor_r1 = FALSE
|   |   |   |   año_ingreso = 1986: FALSE (0.0)
|   |   |   |   año_ingreso = 1987: FALSE (0.0)
|   |   |   |   año_ingreso = 1988: FALSE (0.0)
|   |   |   |   año_ingreso = 1989: FALSE (0.0)
|   |   |   |   año_ingreso = 1991: FALSE (0.0)
|   |   |   |   año_ingreso = 1992: FALSE (0.0)
|   |   |   |   año_ingreso = 1993: FALSE (0.0)
|   |   |   |   año_ingreso = 1994: FALSE (0.0)
|   |   |   |   año_ingreso = 1995: FALSE (0.0)
|   |   |   |   año_ingreso = 1996: FALSE (0.0)
|   |   |   |   año_ingreso = 1997: FALSE (0.0)
|   |   |   |   año_ingreso = 1998: FALSE (0.0)
|   |   |   |   año_ingreso = 1999: FALSE (0.0)
|   |   |   |   año_ingreso = 2000: FALSE (0.0)
|   |   |   |   año_ingreso = 2001: FALSE (1.0)
|   |   |   |   año_ingreso = 2002: FALSE (0.0)
|   |   |   |   año_ingreso = 2003: TRUE (1.0)
|   |   |   |   año_ingreso = 2004: FALSE (1.0)
|   |   |   |   año_ingreso = 2005: FALSE (0.0)
|   |   |   |   año_ingreso = 2006: FALSE (0.0)
|   |   |   tutor_r1 = TRUE: FALSE (7.0)
|   finales_adeudados = 3: FALSE (5.0)
|   finales_adeudados = 4: FALSE (5.0)
|   finales_adeudados = 5: TRUE (1.0)
```

Número de hojas: 116
Tamaño del árbol: 136
Instancias Correctamente Clasificadas: 80 **68.3761 %**
Instancias Incorrectamente Clasificadas: 37 31.6239 %

El árbol generado es muy grande, tanto en profundidad como en cantidad de ramas (genera un total de 116 hojas) y con una precisión del 68.37%, clasifica correctamente 80 instancias.

Se observa que el año de ingreso es el que genera muchas ramas lo cual pareciera hacer el árbol más preciso, en principio. Sin embargo, a medida que transcurra el tiempo y se agreguen años aumentará la cantidad de ramas, y a su vez los datos más antiguos ya no van a aportar información para las nuevas inscripciones. En base a esto y en la experiencia obtenida con One Rule, se decidió analizar el comportamiento de este atributo como diferencia de años entre la fecha de ingreso y la fecha de inscripción a reválidas, y también generando rangos de valores para dicha diferencia.

Los valores que se tomaron fueron:

- Antigüedad Total: fecha de inscripción al plan de reválidas - fecha de inicio en la carrera
- Antigüedad en rangos de 2 años
- Antigüedad en rangos de 3 años
- Antigüedad en rangos de 4 años
- Antigüedad en rangos de 5 años
- Sin año de ingreso

Además, se analizó el comportamiento de todos estas posibilidades combinadas con el atributo *plan* y sin él, y con los atributos *tutor_r1* y *tutor_r2* y sin ellos.

3.2. Resultados con variación de parámetros

A continuación se puede ver la tabla de resultados al variar los datos de entrada antes mencionados, y un análisis de los mismos.

	Finales, Optativas, Cursadas y Adeuda Trabajo Final	Año de ingreso	Total	Rangos de 2	Rangos de 3	Rangos de 4	Rangos de 5	Plan	Tutor	Precisión	Observaciones
Año Ingreso Calendario	X	X						X	X	68.37%	Este es el **valor a mejorar**, dado que es el mejor con año de ingreso calendario. Ningún rango con plan y tutor superó esta precisión Ningún rango con ninguna combinación superó este valor
	X	X						X		65.81%	
	X	X							X	62.39%	
	X	X								64.95%	

Antigüedad Total	X		X					X	X	62.39%	La diferencia total de años no alcanza ni mejora al año calendario de ingreso con ninguna combinación de atributos.
	X		X					X		62.39%	
	X		X						X	54.70%	
	X		X							58.11%	
Antigüedad en Rangos de 2	X			X				X	X	68.37%	Con plan y con tutor, el único igual es rangos de 2, los demás no mejoran en nada el valor de año de ingreso calendario
	X			X				X		62.39%	No mejoran en nada al año de ingreso calendario
	X			X					X	65.81%	
	X			X						60.68%	
Antigüedad en Rangos de 3	X				X			X	X	64.95%	En rangos de 3, no se equipara ni mejora al año calendario de ingreso con ninguna combinación de atributos
	X				X			X		64.95%	
	X				X				X	65.82%	
	X				X					61.53%	
Antigüedad en Rangos de 4	X					X		X	X	60.68%	En rangos de 4, no se equipara ni mejora al año calendario de ingreso con ninguna combinación de atributos
	X					X		X		65.81%	
	X					X			X	60.68%	
	X					X				61.53%	
Antigüedad en Rangos de 5	X						X	X	X	60.68%	No supera el valor a mejorar
	X						X	X		68.37%	Rangos de 5, con Plan y sin tutor equipara al original de año calendario con todos los atributos

	X						X		X	67.52%	No supera el valor a mejorar
	X						X			60.68%	No supera el valor a mejorar
Sin información de año de ingreso	X							X	X	65.82%	No supera el valor a mejorar
	X							X		70.94%	El único que supera al año ingreso calendario es este. Además de mejorar la precisión cumple con la hipótesis que se había planteado de excluir los atributos tutor_r1 y tutor_r2
	X								X	64.10%	No supera el valor a mejorar
	X									57.26%	No supera el valor a mejorar

Tabla 1 - Comparación de resultados de J48 para distintos datos de entrada

Filtrando los resultados con precisión baja y separando los resultados más altos, se puede ver que la combinación de atributos con mejor precisión fue **sin el año de ingreso y sin tutor**, con un **70.94%** de precisión

Como resultado de todo este análisis sobre el *año de ingreso*, se decidió descartar el atributo porque si bien brinda una muy buena precisión sin ser la mejor, no es conveniente utilizarlo a futuro. Por un lado, a medida que pasen los años se seguirán agregando valores posibles de *año_ingreso* aumentando la cantidad de ramas del árbol considerablemente. Por otra parte, si bien en este contexto el porcentaje de certeza es alto, se presume que los datos históricos para los años más viejos no van a ser útiles más adelante dado que los alumnos que se inscriban van a ser de años posteriores; mientras que los años más recientes no tendrán suficiente información histórica para predecir a priori si el alumno va a graduarse o no.

La mejor solución encontrada con el algoritmo J48 fue seleccionando los siguientes atributos:

1. plan,
2. finales_adeudados,
3. cursadas_adeudadas,
4. optativas_adeudadas,
5. adeuda_trab_final_inicio
6. graduado

A continuación, se observa el árbol generado para el primer grupo de datos:

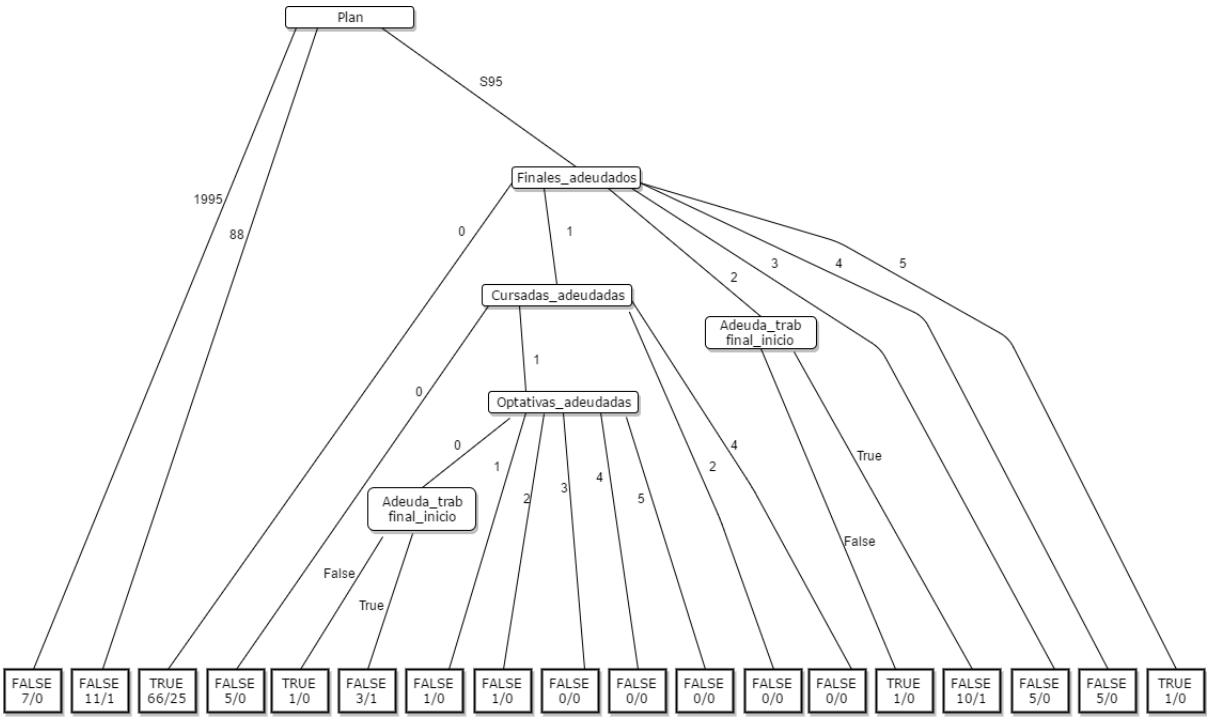

Figura 1 - Árbol de decisión generado por J48 para el primer grupo de datos

Número de hojas: 18
Tamaño del árbol: 24
Instancias Correctamente Clasificadas: 83 **70.9402 %**
Instancias Incorrectamente Clasificadas: 34 29.0598 %

Con el resultado anterior, dado un alumno que se inscribe al plan de reválidas, se puede navegar el árbol según el plan y las actividades académicas adeudadas hasta llegar a una hoja. Allí, se obtiene una predicción sobre si éste va a graduarse o no, con una determinada precisión.

Por ejemplo:

Atributo	Valor
• plan	S95
• finales_adeudados	1
• cursadas_adeudadas	1
• optativas_adeudadas	0
• adeuda_trab_final_inicio	TRUE

Para este conjunto de datos, se predice que no se va a recibir con una precisión del 66% (3.0/1.0 significa que de 3 registros falla en la clasificación de 1). A continuación se puede observar el camino marcado en color rojo en el árbol de decisión.

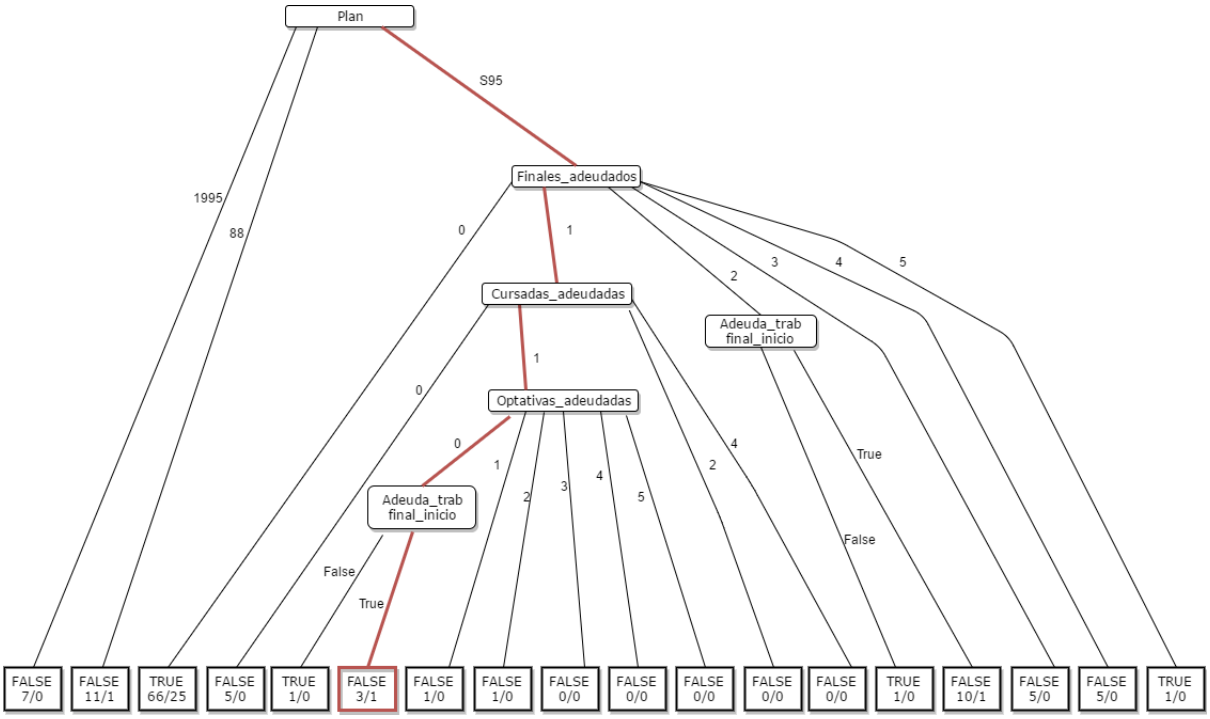

Figura 2 – Camino a la hoja que indica la predicción del ejemplo en el Árbol de decisión

3.3. Conclusiones para J48

El árbol generado a partir de los atributos seleccionados es un árbol con menor cantidad de hojas y menos reglas, por lo tanto la clasificación será más rápida. Por otra parte, cuantos menos atributos presentes haya en el árbol, habrá menos información para que el algoritmo entrene. Por eso, si bien en este caso fue necesario quitar los atributos relacionados al tutor porque al representar escenarios falsos del problema derivó en resultados erróneos; en un futuro, cuando los datos relacionados al tutor reflejen un uso real de las tutorías, se podrían volver a considerar.

Acerca del atributo plan, se observa que si bien el dato no va a aportar mucho en adelante, cuando sólo se inscriban alumnos del plan S95, fue necesario usarlo para que se clasificaran bien las instancias históricas correspondientes a planes viejos y no perder precisión en los resultados. Además, se supone que en algún momento en el futuro podrían llegar a existir nuevos planes y será necesario considerarlo.

A diferencia de One Rule, el algoritmo J48, permite utilizar varios atributos combinados y cuantos más atributos haya, mejor es la precisión.

CONCLUSIONES

Mientras que OneR considera los atributos de forma aislada y genera reglas en base al atributo que más se correlaciona al valor “true” o “false” de *graduado*, J48 considera todos los atributos que brinden información de decisión para llegar a un valor de *graduado* y predecir con qué certeza se puede afirmar que el alumno se va a recibir o no.

Si bien se reconoce que el algoritmo OneR puede ser útil para evaluar índices ad-hoc formados por la combinación de múltiples atributos ponderados y determinar su precisión, estos índices no dejan de ser especulaciones que no están basadas en el análisis estadístico de los datos históricos.

Se concluye que J48 representa mejor la situación de un alumno que se inscribe en el programa de reválidas, ya que no es suficiente saber cuál es el factor que más influye al momento de recibirse, sino que en este caso varios factores -tal vez de menor influencia- conjuntamente pueden afectar al resultado final de igual manera que un factor solo de mayor influencia. Muy importante en este punto es el dinamismo con el que se forman las ramas, porque puede que para cierto valor de un atributo *X* influya más el atributo *Z* y para otro valor de *X* sea otro el atributo *W* que más influya. Por ejemplo, para un alumno que no adeuda finales, el atributo de cursadas adeudadas no influye para nada, mientras que cuando se adeuda un final el atributo de cursadas adeudadas es el más influyente.

Es decir, un alumno que al comenzar el plan de reválidas adeuda 5 finales, pero hace mucho que no rinde nada, quizá tenga la misma posibilidad de recibirse que un alumno que adeuda el trabajo final únicamente y se encuentra en buen ritmo de estudio (su última actividad académica es reciente), y tal vez otro que adeude cursadas y finales tenga las mismas chances. Por esto, se piensa que un método que evalúe todas las alternativas es mejor que uno que genere reglas por cada atributo o dato a considerar.

Una limitación que se encontró al momento de definir un método para obtener el/los indicadores, fue que los datos de R1 están “completos” porque el período del ciclo ya terminó, mientras que los de R2 están incompletos, dado que al finalizar el período es posible que haya más alumnos recibidos que al momento de recolectar estos datos. Sin embargo, se procedió a realizar el análisis con los datos existentes ya que, en el peor de los casos, una vez completada R2 la precisión va a mejorar porque van a existir más datos de entrenamiento y serán más precisos sobre la situación final de los alumnos.

MEJORAS PROPUESTAS

Se recomienda disponer de un atributo que registre la fecha de la **última actividad académica** realizada por el alumno al momento de la inscripción al plan de reválidas, considerando que cuanto más tiempo transcurre desde esta última actividad más difícil es retomar el ritmo de estudio.

REFERENCIAS

[1] Ferraggine V., Torcida S., Introducción a la Ciencia de Datos.

[2] R (https://www.r-project.org/)

[3] CRan (https://cran.r-project.org/)

[4] Dr. Saed Sayad, An Introduction to Data Mining: http://www.saedsayad.com/decision_tree.htm

[5] Dr. Saed Sayad, An Introduction to Data Mining: http://www.saedsayad.com/oner.htm

[6] Francisco Ferrero Mateos, Minería De Datos En Weka: http://www.it.uc3m.es/~jvillena/irc/practicas/03-04/20.mem.pdf