# 

# Índice

[**Índice**](#_ngxwj72uif2o) **1**

[**Elaboración de un Indicador para el Programa de Reválidas**](#_49p4ou8e0hr) **2**

[Introducción](#_r4g2bexn6cea) 2

[Procedimiento](#_ct24555v8qsy) 3

[Datos](#_k6j0zrv3ppqv) 4

[Datos iniciales](#_o51dvyn7qb1y) 4

[Datos excluidos o derivados](#_kxszbvm4o444) 4

[Datos agregados](#_dgpy95lvr0yq) 5

[Datos utilizados](#_12qfp1z543mu) 5

[Métodos de clasificación](#_jemoeh19op5g) 6

[OneR](#_9rr89znm22hk) 6

[Pseudocódigo](#_z4be1uyh38h4) 7

[Ejemplo](#_259z8ugv12s) 7

[Referencias](#_cphsjoovgxo7) 8

[J48](#_oi5lhlubyz3w) 8

[Ejemplo](#_yot1hbi5xp45) 9

[Referencias](#_3jb9w580rciv) 9

[Atributos incluidos para ambos métodos](#_h78empz1athw) 10

[Resultados](#_a1hwdngsvc8p) 11

[OneR](#_5l3et7apyeso) 11

[J48](#_3r496zymz81x) 15

[Árbol de decisión (sin podar) con atributo plan incluido](#_1qc2lfr4e1lg) 15

[Árbol de decisión (sin podar) sin atributo plan](#_bkddw4wr12sm) 16

[Breve análisis inicial](#_egkkkqht4vi7) 20

# 

# Elaboración de un Indicador para el Programa de Reválidas

## 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 recurse, 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[[1]](#footnote-1). 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.

## Procedimiento

Como sugerencia de la cátedra, se tomó como base un trabajo similar[[2]](#footnote-2) 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 tratamos de reutilizar en nuestro 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](https://es.wikipedia.org/wiki/Extract,_transform_and_load) (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[[3]](#footnote-3).

* + 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 par poder volcarlos en la base de datos.
  + En la **etapa de carga** se insertaron los datos en la tabla reválidas, para analizarlos con R.

1. Se instaló y configuró CRan[[4]](#footnote-4) 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.
2. 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 Fac. Cs. Exactas, se sugirieron incorporaciones de datos y se dejaron de lado otros, a partir del análisis de los resultados.

### Datos iniciales

|  |  |  |
| --- | --- | --- |
| Nro | Columna | 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. |

### 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.

Luego de conversar con la comisión de tutorías, se concluyó que el atributo de *“Tutor”*, tanto en R1 como en R2, no debía tenerse en cuenta, ya que en R1 la mayoría de los alumnos se inscribió en las tutorías pensando que era obligatorio, pero luego no acompañó el desarrollo del programa con el tutor. En cambio, en R2 la mayoría no se inscribió en el programa de tutorías.

### 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

### 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\_anios\_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 rangos de 5 años. (1 si esta entre 0-4 años, 2 entre 5-9, 3 entre 10-14, etc) |
| 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 | fecha\_ultima\_act\_academica => tiempo\_desde\_última\_actividad\_académica | A partir de la fecha de última actividad académica, se calculó el tiempo transcurrido desde la última actividad académica realizada antes de inscribirse al plan de reválidas en <meses/años/rango>. |
| 10 | graduado | Valor booleano si se recibió estando anotado en el plan. |
| 11 | fecha\_graduación => tiempo\_transcurrido\_graduacion | A partir de la fecha de graduación se calculó el tiempo transcurrido en <definir meses/años> entre que el alumno se anotó en el plan de reválidas y la fecha en la que se recibió. |

## Métodos de clasificación

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 124 estudiantes considerados, 54 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 43,54% de los casos. Modelos alternativos de predicción del resultado deberán al menos alcanzar la misma precisión.

### OneR

OneR viene de One Rule, es un algoritmo de clasificación que genera un árbol de decisión de un único nivel. OneR es capaz de inferir reglas de clasificación a partir de un conjunto de instancias. El algoritmo crea una regla para cada atributo en los datos de entrenamiento, luego escoge la regla con la *tasa de error*[[5]](#footnote-5) 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.

#### Pseudocódigo

Para cada atributo o “predictor”[[6]](#footnote-6)

Para cada valor del atributo, crear una regla como sigue:

Contar las apariciones de cada valor del target[[7]](#footnote-7) (atributo a predecir: recibido)

Seleccionar la clase[[8]](#footnote-8) 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

#### 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 cuál 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 un accuracy 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 accuracy). 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

#### Referencias

* http://www.saedsayad.com/oner.htm

### J48

Este algoritmo construye árboles de decisión desde un grupo de datos de entrenamiento usando el concepto de [entropía de](https://es.wikipedia.org/w/index.php?title=Entropia_(teor%C3%ADa_de_informaci%C3%B3n)&action=edit&redlink=1) 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 sublistas 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.

#### 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.

#### Referencias

* An Introduction to Data Mining, [Dr. Saed Sayad](http://www.saedsayad.com/author.htm):<http://www.saedsayad.com/decision_tree.htm>

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

### Atributos incluidos para ambos métodos

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

Se nos había sugerido que los atributos *tutor\_r1* y *tutor\_r2* no se estaba seguros si tenerlo en cuenta, 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-incribieron luego de R1 en R2 cambiaron su condición respecto de este valor.

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 espera que la mayoría de los casos siempre va a tener plan S95 en adelante, dado que quedan pocos alumnos de planes anteriores inscriptos en la carrera sin graduarse.

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 año de ingreso fue reemplazado por la cantidad de años desde la inscripción, ya que nos parecía más significativo el valor. Por un lado, el indicador se calcula con datos históricos correspondientes a varios ciclos de Plan de Reválidas y a medida que pase el tiempo van a ser cada .

## 

## 

## Resultados

### OneR

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

#### Resultados Iniciales de One Rule

Atributo Precisión

**1 \* año\_ingreso 63.71%**

2 adeuda\_trab\_final\_inicio 62.9%

3 finales\_adeudados 60.48%

4 plan 56.45%

4 cursadas\_adeudadas 56.45%

4 optativas\_adeudadas 56.45%

4 tutor\_r1 56.45%

4 tutor\_r2 56.45%

**Atributo seleccionado dada la precisión:** '\*' (año\_ingreso)

**Reglas generadas a partir del atributo seleccionado (adeuda\_trabajo\_final\_inicio**)**:**

If año\_ingreso= 1986 then graduado = TRUE

If año\_ingreso= 1987 then graduado = FALSE

If año\_ingreso= 1988 then graduado = FALSE

If año\_ingreso= 1989 then graduado = FALSE

If año\_ingreso= 1991 then graduado = FALSE

If año\_ingreso= 1992 then graduado = FALSE

If año\_ingreso= 1993 then graduado = TRUE

If año\_ingreso= 1994 then graduado = FALSE

If año\_ingreso= 1995 then graduado = FALSE

If año\_ingreso= 1996 then graduado = FALSE

If año\_ingreso= 1997 then graduado = TRUE

If año\_ingreso= 1998 then graduado = FALSE

If año\_ingreso= 1999 then graduado = FALSE

If año\_ingreso= 2000 then graduado = FALSE

If año\_ingreso= 2001 then graduado = TRUE

If año\_ingreso= 2002 then graduado = FALSE

If año\_ingreso= 2003 then graduado = TRUE

If año\_ingreso= 2004 then graduado = FALSE

If año\_ingreso= 2005 then graduado = FALSE

If año\_ingreso= 2006 then graduado = FALSE

**Precisión:**

79 de 124 instancias clasificadas correctamente (63.71%)

#### Análisis Resultados Iniciales One Rule

##### Atributo año\_ingreso

A partir de los resultados de la corrida inicial, se analizó la selección del atributo ***año\_ingreso*** 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 el **año de ingreso a la carrera** con una certeza del **63.71%.**

Si bien este atributo es el que mejor precisión tiene, se puso en duda su utilización para definir las reglas, ya que los datos sobre el estado de las actividades académicas del alumno no se consideran para nada.

Por un lado, a medida que pasen los años se seguirán agregando valores posibles de *año\_ingreso* aumentando la cantidad de reglas considerablemente. Además, si bien en este contexto el porcentaje de certeza es el mayor, 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 estos motivos, se decidió probar en su lugar con un dato derivado que es la cantidad de tiempo transcurrido desde el año de ingreso hasta la inscripción al plan de reválidas al cual se denominó ***dif\_año\_ingreso***.

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

Attribute Accuracy

**1 \* adeuda\_trab\_final\_inicio 62.9%**

2 finales\_adeudados 60.48%

3 dif\_anios\_ingreso 57.26%

4 plan 56.45%

4 cursadas\_adeudadas 56.45%

4 optativas\_adeudadas 56.45%

4 tutor\_r1 56.45%

4 tutor\_r2 56.45%

Se observa que, al dejar de considerar el año calendario de ingreso, el nuevo atributo seleccionado por OneR es *adeuda\_trab\_final\_inicio*; mientras que el tiempo transcurrido en años desde el ingreso *dif\_anios\_ingreso* pasó al tercer lugar de precisión para este método.

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)

Attribute Accuracy

**1 \* adeuda\_trab\_final\_inicio 62.9%**

2 finales\_adeudados 60.48%

3 dif\_anios\_ingreso\_rangos\_de\_4 58.87%

4 dif\_anios\_ingreso\_total 58.06%

4 dif\_anios\_ingreso\_rangos\_de\_2 58.06%

6 dif\_anios\_ingreso\_rangos\_de\_3 57.26%

6 dif\_anios\_ingreso\_rangos\_de\_5 57.26%

8 plan 56.45%

8 cursadas\_adeudadas 56.45%

8 optativas\_adeudadas 56.45%

8 tutor\_r1 56.45%

8 tutor\_r2 56.45%

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. Aún así, el algoritmo sigue eligiendo como mejor clasificador el atributo *adeuda\_trab\_final\_inicio*.

Como conclusión del análisis sobre el año de ingreso, se puede decir que ...

lo que más me confunde/sorprende, es que el atributo este sigue siendo el de > precisión, incluso vs dif, rangos y demás.

##### Atributo adeuda\_trabajo\_final\_inicio

A partir de estos nuevos resultados, se analiza la salida para la selección del atributo ***adeuda\_trab\_final\_inicio.***

**Reglas generadas a partir del atributo seleccionado (adeuda\_trabajo\_final\_inicio**)**:**

If adeuda\_trab\_final\_inicio = FALSE then recibido = TRUE

If adeuda\_trab\_final\_inicio = TRUE then recibido = FALSE

**Precisión:**

78 de 124 instancias clasificadas correctamente (62.9%)

Significando que para One Rule, los alumnos que adeuden trabajo final no se graduarán y los que no lo adeuden si van a hacerlo con un **62.9%** de certeza.

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 One Rule.

### J48

#### Árbol de decisión (sin podar) *con atributo plan incluido*

finales\_adeudados = 0

| plan = 1995

| | adeuda\_trab\_final\_inicio = FALSE

| | | optativas\_adeudadas = 0: TRUE (3.0)

| | | optativas\_adeudadas = 1: FALSE (1.0)

| | | optativas\_adeudadas = 2: TRUE (0.0)

| | | optativas\_adeudadas = 3: TRUE (0.0)

| | | optativas\_adeudadas = 4: TRUE (0.0)

| | | optativas\_adeudadas = 5: TRUE (0.0)

| | adeuda\_trab\_final\_inicio = TRUE: FALSE (3.0)

| plan = 88: FALSE (8.0/1.0)

| plan = S95: TRUE (70.0/25.0)

finales\_adeudados = 1

| cursadas\_adeudadas = 0: FALSE (5.0)

| cursadas\_adeudadas = 1

| | plan = 1995: FALSE (0.0)

| | plan = 88: FALSE (2.0)

| | plan = S95

| | | optativas\_adeudadas = 0

| | | | adeuda\_trab\_final\_inicio = FALSE: TRUE (1.0)

| | | | adeuda\_trab\_final\_inicio = TRUE: FALSE (3.0/1.0)

| | | 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

| | cursadas\_adeudadas = 0: TRUE (1.0)

| | cursadas\_adeudadas = 1: FALSE (1.0)

| | cursadas\_adeudadas = 2: FALSE (0.0)

| | cursadas\_adeudadas = 4: FALSE (0.0)

| adeuda\_trab\_final\_inicio = TRUE: FALSE (12.0/1.0)

finales\_adeudados = 3: FALSE (6.0)

finales\_adeudados = 4: FALSE (5.0)

finales\_adeudados = 5: TRUE (1.0)

Número de hojas: 29

Tamaño del árbol : 39

Instancias correctamente clasificadas: 86 (69.3548 % de precisión)

Instancias incorrectamente clasificadas: 38 (30.6452 % de error)

=== Matriz de Confusión ===

|  |  |  |
| --- | --- | --- |
| a | b | <-- clasificado como |
| 41 | 29 | a = FALSE |
| 9 | 45 | b = TRUE |

#### Árbol de decisión (sin podar) *sin atributo plan*

finales\_adeudados = 0

| adeuda\_trab\_final\_inicio = FALSE

| | optativas\_adeudadas = 0: TRUE (12.0)

| | optativas\_adeudadas = 1: FALSE (1.0)

| | optativas\_adeudadas = 2: TRUE (0.0)

| | optativas\_adeudadas = 3: TRUE (0.0)

| | optativas\_adeudadas = 4: FALSE (1.0)

| | optativas\_adeudadas = 5: TRUE (0.0)

| adeuda\_trab\_final\_inicio = TRUE

| | tutor\_r2 = FALSE: TRUE (70.0/33.0)

| | tutor\_r2 = TRUE: FALSE (1.0)

finales\_adeudados = 1: FALSE (13.0/2.0)

finales\_adeudados = 2

| adeuda\_trab\_final\_inicio = FALSE

| | cursadas\_adeudadas = 0: TRUE (1.0)

| | cursadas\_adeudadas = 1: FALSE (1.0)

| | cursadas\_adeudadas = 2: FALSE (0.0)

| | cursadas\_adeudadas = 4: FALSE (0.0)

| adeuda\_trab\_final\_inicio = TRUE: FALSE (12.0/1.0)

finales\_adeudados = 3: FALSE (6.0)

finales\_adeudados = 4: FALSE (5.0)

finales\_adeudados = 5: TRUE (1.0)

Número de hojas : 17

Tamaño del árbol : 23

Instancias correctamente clasificadas: 79 (63.7097 % de precisión)

Instancias incorrectamente clasificadas: 45 (36.2903 % de error)

Si no se analiza el atributo plan baja la precisión del mismo, esto se debe a que los alumnos que pertenecían a planes como el 88 que se anotaron en r1 solo porque en paralelo estaba el plan Delta y no se recibieron se clasifiquen erróneamente.

Árbol de decisión (sin podar) con dif\_año\_ingresocon rango de 5 años y sin plan

finales\_adeudados = 0

| adeuda\_trab\_final\_inicio = FALSE

| | optativas\_adeudadas = 0: TRUE (12.0)

| | optativas\_adeudadas = 1: FALSE (1.0)

| | optativas\_adeudadas = 2: TRUE (0.0)

| | optativas\_adeudadas = 3: TRUE (0.0)

| | optativas\_adeudadas = 4: FALSE (1.0)

| | optativas\_adeudadas = 5: TRUE (0.0)

| adeuda\_trab\_final\_inicio = TRUE

| | dif\_año\_ingreso= 2: TRUE (16.0/7.0)

| | dif\_año\_ingreso= 3

| | | optativas\_adeudadas = 0: TRUE (32.0/15.0)

| | | optativas\_adeudadas = 1: FALSE (1.0)

| | | optativas\_adeudadas = 2: FALSE (4.0/2.0)

| | | optativas\_adeudadas = 3: TRUE (0.0)

| | | optativas\_adeudadas = 4: TRUE (0.0)

| | | optativas\_adeudadas = 5: TRUE (0.0)

| | dif\_año\_ingreso= 4: TRUE (10.0/2.0)

| | dif\_año\_ingreso= 5: FALSE (6.0)

| | dif\_año\_ingreso= 6: FALSE (2.0/1.0)

finales\_adeudados = 1

| cursadas\_adeudadas = 0: FALSE (5.0)

| cursadas\_adeudadas = 1

| | optativas\_adeudadas = 0

| | | dif\_año\_ingreso= 2: FALSE (3.0/1.0)

| | | dif\_año\_ingreso= 3: TRUE (1.0)

| | | dif\_año\_ingreso= 4: FALSE (0.0)

| | | dif\_año\_ingreso= 5: FALSE (2.0)

| | | dif\_año\_ingreso= 6: 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

| dif\_año\_ingreso= 2

| | adeuda\_trab\_final\_inicio = FALSE: TRUE (1.0)

| | adeuda\_trab\_final\_inicio = TRUE: FALSE (5.0/1.0)

| dif\_año\_ingreso= 3: FALSE (6.0)

| dif\_año\_ingreso= 4: FALSE (1.0)

| dif\_año\_ingreso= 5: FALSE (1.0)

| dif\_año\_ingreso= 6: FALSE (0.0)

finales\_adeudados = 3: FALSE (6.0)

finales\_adeudados = 4: FALSE (5.0)

finales\_adeudados = 5: TRUE (1.0)

Número de hojas : 38

Tamaño del árbol : 48

Instancias correctamente clasificadas: 74 (59.6774 % de precisión)

Instancias incorrectamente clasificadas: 50 (40.3226 % de error)

Baja la cantidad de instancias que se clasificaron correctamente.

Árbol de decisión (sin podar) con dif\_año\_ingresocon rango de 5 años y con plan

finales\_adeudados = 0

| plan = 1995

| | adeuda\_trab\_final\_inicio = FALSE

| | | optativas\_adeudadas = 0: TRUE (3.0)

| | | optativas\_adeudadas = 1: FALSE (1.0)

| | | optativas\_adeudadas = 2: TRUE (0.0)

| | | optativas\_adeudadas = 3: TRUE (0.0)

| | | optativas\_adeudadas = 4: TRUE (0.0)

| | | optativas\_adeudadas = 5: TRUE (0.0)

| | adeuda\_trab\_final\_inicio = TRUE: FALSE (3.0)

| plan = 88: FALSE (8.0/1.0)

| plan = S95

| | adeuda\_trab\_final\_inicio = FALSE: TRUE (9.0)

| | adeuda\_trab\_final\_inicio = TRUE

| | | dif\_año\_ingreso= 2: TRUE (14.0/5.0)

| | | dif\_año\_ingreso= 3

| | | | optativas\_adeudadas = 0: TRUE (31.0/14.0)

| | | | optativas\_adeudadas = 1: FALSE (1.0)

| | | | optativas\_adeudadas = 2: FALSE (4.0/2.0)

| | | | optativas\_adeudadas = 3: TRUE (0.0)

| | | | optativas\_adeudadas = 4: TRUE (0.0)

| | | | optativas\_adeudadas = 5: TRUE (0.0)

| | | dif\_año\_ingreso= 4: TRUE (10.0/2.0)

| | | dif\_año\_ingreso= 5: FALSE (1.0)

| | | dif\_año\_ingreso= 6: TRUE (0.0)

finales\_adeudados = 1

| cursadas\_adeudadas = 0: FALSE (5.0)

| cursadas\_adeudadas = 1

| | plan = 1995: FALSE (0.0)

| | plan = 88: FALSE (2.0)

| | plan = S95

| | | optativas\_adeudadas = 0

| | | | dif\_año\_ingreso= 2: FALSE (3.0/1.0)

| | | | dif\_año\_ingreso= 3: TRUE (1.0)

| | | | dif\_año\_ingreso= 4: FALSE (0.0)

| | | | dif\_año\_ingreso= 5: FALSE (0.0)

| | | | dif\_año\_ingreso= 6: 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

| | dif\_año\_ingreso= 2: TRUE (1.0)

| | dif\_año\_ingreso= 3: FALSE (0.0)

| | dif\_año\_ingreso= 4: FALSE (0.0)

| | dif\_año\_ingreso= 5: FALSE (1.0)

| | dif\_año\_ingreso= 6: FALSE (0.0)

| adeuda\_trab\_final\_inicio = TRUE: FALSE (12.0/1.0)

finales\_adeudados = 3: FALSE (6.0)

finales\_adeudados = 4: FALSE (5.0)

finales\_adeudados = 5: TRUE (1.0)

Número de hojas : 43

Tamaño del árbol : 56

Instancias correctamente clasificadas: 84 (67.7419 % de precisión)

Instancias incorrectamente clasificadas: 74 (32.2581 % de error)

## Breve análisis inicial

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 recibido, J48 considera todos los atributos que brinden información de decisión para llegar a un valor de *“recibido”* y predecir con qué certeza se puede afirmar que el alumno se va a recibir o no.

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

Creemos 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 nos parece que 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 ej, 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 ritmo de estudio (su última actividad académica es reciente), y tal vez otro que adeude cursadas y finales tenga las mismas chances. Por esto, pensamos 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 vemos al momento de definir un método para obtener el/los indicadores, es 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. Esto significa que, en el peor de los casos, la precisión va a mejorar porque van a existir más datos de entrenamiento y serán más precisos sobre la situación final de los alumnos.

Mejora

* ultima\_actividad\_academica

Comentar que vimos que se podía mejorar con esta info y realizar de nuevo los cálculos, mostrar resultados y conclusiones.

1. Los créditos son la suma de las actividades académicas adeudadas. Por ejemplo, un final adeudado suma 1 crédito, una cursada adeudada 2 (1 por la cursada y 1 por el final), una optativa 1 crédito y la tesis 1 crédito. [↑](#footnote-ref-1)
2. Introducción a la Ciencia de Datos - Alumnos: Viviana Ferraggine y Sebastián Torcida. Prof. Dr. Gustavo Illescas. [↑](#footnote-ref-2)
3. **R** (<https://www.r-project.org/>) es un entorno de software libre para informática estadística y gráficos. Se compila y se ejecuta en una amplia variedad de plataformas/sistemas operativos. [↑](#footnote-ref-3)
4. **CRan** (<https://cran.r-project.org/>) es una herramienta para correr consultas de R sobre una base de datos en modo gráfico o de consola. En nuestro caso elegimos la consola para poder programar los scripts necesarios y dejarlos a disposición para futuras actualizaciones. [↑](#footnote-ref-4)
5. Tasa de error: número de instancias de los datos de entrenamiento en los que la clase del valor de un atributo no concuerda con la asociación que la regla le da al valor de ese atributo. [↑](#footnote-ref-5)
6. Atributos/predictores: son los datos de entrada que ayudarán a predecir el target,en nuestro caso cant\_finales\_adeudados, adeuda\_trabajo\_final\_inicio, cant\_cursadas\_adeudadas, etc. [↑](#footnote-ref-6)
7. Target: es el atributo a predecir, en nuestro caso “recibido”. [↑](#footnote-ref-7)
8. Clase: es la cantidad de veces que aparece cada valor de un atributo. [↑](#footnote-ref-8)