Introducción a la Ciencia de Datos

Alumnos: Viviana Ferraggine y Sebastián Torcida

Prof. Dr. Gustavo Illescas

Objetivo

El objetivo de este trabajo es aplicar algunas herramientas básicas de Ciencias de Datos a un problema concreto. Específicamente:

- elaborar una base de datos a partir de la información disponible en torno al problema;
- llevar a cabo un procesamiento del tipo *data mining* y listar sus conclusiones.

Las etapas que tuvieron lugar fueron las siguientes:

1) Se identificó el problema de evaluar los resultados de la primera edición del programa *Delta G*, proyecto de la Secretaría de Políticas Universitarias (SPU) que impulsaba la graduación en el término de un año de estudiantes avanzados de Ingeniería y Agronomía insertos laboralmente. El programa contemplaba la adjudicación de becas-estímulo de $25.000 y el período de vigencia fue desde abril 2014 a marzo 2015.

Los datos de la implementación del *Delta G* para la carrera de Ingeniería en Sistemas de la Facultad de Ciencias Exactas (FCEX) de la UNCPBA utilizados en éste análisis fueron facilitados por la Lic. Laura Rivero. El programa establecía las siguientes condiciones:

- ser estudiante avanzado de Ingeniería o Agronomía inserto laboralmente (ya sea en relación de dependencia o de manera independiente) cuya finalización de carrera ha sido discontinuada o retrasada.
- adeudar hasta 4 (cuatro) actividades académicas, incluyendo el Trabajo Final de Grado.
- recibirse en el término de un año.

2) Para el procesamiento de los datos se utilizó un proceso **ETL** (*Extract, Transform, Load*: extraer, transformar y cargar), que permitió mover datos desde múltiples fuentes (planillas excel), reformatearlos y limpiarlos, y cargarlos en una base de datos PostgreSQL para analizarlos desde otros sistemas operacionales como R y Weka.

En la etapa de **extracción**, en la mayoría de los proyectos ETL, los datos provenientes de diferentes fuentes de origen se fusionan. En esta parte hay que analiza que cada sistema por separado puede usar una organización diferente de los datos o formatos distintos, dichos formatos están fuertemente atados a su soporte ya sea otras bases de datos o archivo planos (planillas excel, documentos word, archivos csv, u otros formatos). En esta etapa se convierten los datos a un formato preparado para iniciar el proceso de transformación.

- se creo una base de datos especialmente dedicada a contener los datos de Delta G en un servidor PostgreSQL que utiliza la cátedra de Bases de Datos. Dicha tarea la realizó el Ig. Gustavo Correa Reina, responsable del mantenimiento de dicho servidor.
- se asignaron permisos de acceso a esta base de datos a 3 usuarios (Ferraggine, V.; Rivero, L. y Del Fresno, M)
- se utilizó el comando copy de POstgreSQL para ingresar los datos originales a la base de datos (https://www.postgresql.org/docs/9.2/static/sql-copy.html)
- se creo una tabla para contener todos los datos migrados y transformados (alumno_delta_g_i)

En la etapa de **transformación** se requirió realizar una serie de adaptaciones de los datos migrados en la que se aplicaron las siguientes transformaciones:

- se descartaron algunas columnas que eran redundantes.
- se limpian los datos que consideraban 2 años de ingreso diferentes considerando los que se encontraban etre paréntesis.
- se agrega la información referida a la fecha de egreso.
- se agrega una columna egresado con valores verdadero/falso y se obtiene este valos calculado a partir de la fecha de egreso

La etapa de **carga** incorporó los datos de la fase anterior (transformación) en la tabla alumno_delta_g_i, dejándola lista para poder ser utilizados desde R y Weka y proceder a los análisis correspondientes.

3) 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 con el software abierto de *machine learning* Weka 3.8 (*http://www.cs.waikato.ac.nz/ml/weka/downloading.html*). 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

El presente análisis utilizó los datos de 61 estudiantes inscriptos en el programa. Esos datos, que incluían diversas variables y también índices elaborados específicamente, fueron oportunamente evaluados por una comisión ad hoc de la FCEX. La comisión utilizó dicha información para generar un Orden de Mérito que posteriormente se elevó al Ministerio de Educación a fin de asignar las becas del programa.

La tabla *alumno_delta_g_i* incluye las variables e índices utilizados en los diferentes análisis:[1]

Nro.	***Columna***	**Descripción**
1.	*cantidad_aaa_real*	Cantidad de actividades académicas adeudadas (cursadas, finales, tesis) por el estudiante. Valores enteros entre 1 y 4.
2.	*difa4_cantidad_aaa_real*	La diferencia entre 4 y el valor de la variable anterior. Valores enteros entre 3 y 0.
3.	*dias_desde_ult_aa*	Días contados desde la última actividad académica realizada por el estudiante. Valores enteros.
4.	*cociente_desde_cant_aaa_real*	El cociente la cantidad de días de su última actividad académica y la cantidad de actividades académicas adeudadas
5.	*nota_promedio*	Promedio de carrera. Valores continuos entre 1 y 10.
6.	*cantidad_aplazos*	Número de aplazos. Valores enteros.
7.	*aplazos_div_10*	Número de aplazos dividido 10.
8.	*factor_rendim_acad*	Factor de rendimiento académico= Promedio de carrera menos número de aplazos dividido 10
9.	*categoria_proyecto_final*	Trabajo final finalizado =3 Plan de trabajo final aprobado=2 No tiene plan de trabajo presentado=1
10.	*suma_aaa_plan*	*Actividades académicas adeudadas más categoría de plan de trabajo final.*
11.	*dias_desde_normal*	Días contados desde la última actividad académica realizada normalizada
12.	*suma_aaa_plan_normal*	*Actividades académicas adeudadas más plan*
13.	*factor_rendim_acad_normal*	Factor de rendimiento académico normalizado
14.	*indice_1*	Una combinación ponderada de las variables 10, 11 y 13 con pesos 2.5, 5 y 2.5 respectivamente, utilizada para rankear estudiantes
15.	*indice_2*	Una combinación ponderada de las variables 10, 11 y 13 con pesos 2, 4 y 4 respectivamente, utilizada para rankear estudiantes
16.	*om_delta_g_i*	Posición final en el Orden de Mérito
17.	*egresado*	Si el alumno culminó ("TRUE" ó "1") o no ("FALSE" ó "0") los estudios en el transcurso del año. Variable lógica o dicotómica

.

[1]Los datos de variables que no resultaron relevantes en los análisis o que contenían información confidencial de los postulantes no fueron incluidos en este informe.

Conclusiones

Los resultados más interesantes del análisis realizado se presentan a continuación.

Asociación entre variables

Para estudiar la asociación entre las distintas variables y la posterior obtención (o no) del título se consideró la *correlación biserial* entre cada variable/índice y la variable binaria *egresado*. La correlación biserial mide el grado de la asociación entre una variable numérica cualquiera y una variable *binaria* ó con sólo dos valores, típicamente "0" (no se graduó, en este caso) y "1" (se graduó). Como la mayoría de los coeficientes de correlación, sus valores se encuentran entre -1 y 1 donde valores absolutos cercanos a 1 indican fuerte asociación, mientras que valores cercanos a 0 sugieren no asociación.

Para estudiar la asociación entre cualquier otro par de variables se consideraron tanto la típica *correlación de Pearson*, que mide asociación del tipo lineal entre las variables, como la *correlación de Spearman* o correlación por rankings, que es más general y robusta que Pearson ya que mide la relación de monotonía no necesariamente lineal entre las variables. (Nota: es importante recordar que ninguna correlación establece relaciones de causalidad, sino que se limita a medir patrones de co-ocurrencia).

Comandos en R

```
biserial(var_name,egresado)
cor(dataframe_name, [,var_range], use="pairwise.complete.obs", method="Pearson")
cor(dataframe_name, [,var_range], use="pairwise.complete.obs", method="Spearman")
```

Resultados

- Ninguno de los dos índices elaborados por la comisión, *indice_1* e *indice_2*, mostraron una correlación biserial importante con la obtención del título. Esto relativiza la capacidad de tales índices de brindar información sobre la culminación de la carrera; más aún, se verá luego que el *indice_1* resulta un clasificador bastante preciso aunque *no monótono* de la obtención del título, lo que pondría en duda su utilidad.
- Tampoco fue importante la correlación biserial entre el Orden de Mérito (*om_delta_g_i*) elaborado y la posterior obtención del título. Queda en duda entonces la eficacia de dicho Orden de Mérito en brindar información sobre el posible cumplimiento del objetivo final del programa.
- La categoría del proyecto final (*categoria_proy_final*) presentó la única correlación biserial importante con la obtención del título, con un valor de -0.7: así, una categoría "3" ocurrió en general conjuntamente con la no graduación posterior mientras que una categoría "2" se dio en gral. con la graduación; este resultado se confirmó luego al considerar el algoritmo de clasificación OneR. Como conclusión, esta variable por sí sola resultó mucho más eficaz para brindar información sobre la posterior obtención del título que cualquier

otra variable. En particular, mucho más eficaz que los índices y el Orden de Mérito elaborados a tal fin.

- En gral., el Orden de Mérito elaborado por la comisión mostró fuertes asociaciones inversas (correlaciones de Pearson y de Spearman negativas y superiores a 0.7 en valor absoluto) con ambos índices, con el cociente entre días transcurridos desde la última actividad académica realizada y la cantidad de actividades académicas adeudadas (variable 4, otro indicador ad hoc) y con los días transcurridos normalizados (variable 11, también ad hoc).

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 del año*" 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 61 estudiantes considerados, 34 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 **55.73%** de los casos. Modelos alternativos de predicción del resultado deberán al menos alcanzar la misma precisión.

OneR ("One Rule")

Se trata de un algoritmo de clasificación simple y preciso, que genera inicialmente un clasificador (regla) por cada variable considerada en los datos. Para ello construye una tabla de frecuencias *niveles de la variable vs. clases*, asignando a cada nivel su clase más frecuente y calculando luego el error total respectivo. La variable con el menor error total es finalmente seleccionada por el algoritmo como la regla individual con el mayor poder predictivo (nota: si la variable es cuantitativa, se la discretiza de manera dinámica y el procedimiento que se sigue es análogo al descripto).

OneR produce resultados ligeramente menos precisos que muchos algoritmos de clasificación más complejos, pero al estar basado en una sola regla o variable es fácil de interpretar para cualquier usuario.

indice_1:

< 4.0641525 -> FALSE (*no se graduará*)

< 4.904644 -> TRUE (*se graduará*)

< 5.245241 -> FALSE

>= 5.245241 ->TRUE

(Stratified cross-validation)

Correctly Classified Instances	52	**85.2459 %**
Incorrectly Classified Instances	9	14.7541 %

CONFUSION MATRIX

a	b	<-- classified as
29	5	a = TRUE
4	23	b = FALSE

Sorpresivamente, la variable indice_1 (utilizada por la comisión para confeccionar un Orden de Mérito preliminar no considerado en éste análisis) es la que tiene el menor error pero no resulta un indicador monótono en términos de predecir la graduación. Teniendo en cuenta que la correlación entre este índice y el Orden de Mérito es importante y negativa (-0.76, indicando que en gral. a mayor valor del indice_1 menor será el ranking del individuo y por lo tanto más alta su posición en el Orden de Mérito final) se evidencian algunas inconsistencias: para un estudiante con valor de indice_1 de aproximadamente 4.5 se predice que se graduará, mientras que para un valor 5 se predice que no lo hará. Y este clasificador alcanza una precisión de más del 85%.

Como información complementaria se analizó entonces cuál de las variables originales (esto es, descartando ambos índices) tenía la mayor precisión. La respuesta fue la variable *categoria_proyecto_final*, que sin estar relacionada con el *indice_1* en su construcción consigue la misma precisión que aquel.

categoria_proyecto_final:

< 2.5 -> TRUE

>= 2.5 ->FALSE

(Stratified cross-validation)

Correctly Classified Instances	52	**85.2459 %**
Incorrectly Classified Instances	9	14.7541 %

CONFUSION MATRIX

a	b	<-- classified as
29	5	a = TRUE
4	23	b = FALSE

Así, considerando únicamente la categoría del proyecto final (variable que sólo toma los valores "2" y "3") se puede predecir si el estudiante se graduará en el transcurso del año: con un 85% de precisión, terminará graduándose si su proyecto final es categoría 2 mientras que no lo hará si su proyecto es categoría 3.

Este resultado es realmente interesante si se tiene en cuenta que la comisión generó varios indicadores para obtener un Orden de Mérito que permitiera pensar que los mejor posicionados lograrían alcanzar el título. Ninguno de ellos consiguió el mismo poder predictivo que la categoría del proyecto.

J48

Este algoritmo recursivo genera árboles de clasificación basados en conceptos de la *teoría de la información*, utilizando una parte de los datos como entrenamiento y otra parte para su testeo. En cada paso, el algoritmo intenta encontrar los nodos (variables) cuyas categorías sean "lo más puras posibles": es decir, que cada categoría esté en preferentemente asociada a una sola clase de las que se pretende predecir.

El criterio habitual para la selección de los nodos es la *ganancia de información normalizada* (variación de la entropía) de modo que la variable con la mayor ganancia de información es la elegida en cada paso de la recursión. Para evitar que el árbol final resulte demasiado dependiente de los datos de entrenamiento (overfitting), diversos parámetros permiten controlar la cantidad mínima de casos alcanzados por cada hoja. Cuando se utilizan variables numéricas, éstas son discretizadas de manera dinámica buscando la menor pérdida de información; en caso de datos faltantes (missing data) las variables respectivas no son utilizadas para el cálculo de la entropía. Parámetros de "poda" permiten además mantener el árbol en un nivel aceptable de complejidad. En síntesis, el método es robusto y los árboles que produce son de fácil interpretación.

categoria_proyecto_final<= 2

cantidad_aplazos<= 11: TRUE (30.0 correct/2.0 incorrect)

cantidad_aplazos> 11: FALSE (3.0/1.0)

categoria_proyecto_final> 2

nota_promedio<= 7.33: FALSE (23.0/1.0)

nota_promedio> 7.33: TRUE (5.0/1.0)

Number of Leaves : 4
Size of thetree : 7
(Stratifiedcross-validation)

CorrectlyClassifiedInstances	52	**85.2459 %**
IncorrectlyClassifiedInstances	9	14.7541 %

CONFUSION MATRIX

a	b	<-- classified as
30	4	a = TRUE
5	22	b = FALSE

Resulta interesante observar la elevada precisión que se consigue empleando un árbol relativamente simple, de 4 hojas (y 7 nodos en total). Notar que la variable raíz (la de mayor ganancia de información) coincide con la seleccionada por el algoritmo OneR como mejor predictora en ausencia del *indice_1*. Pero el árbol generado por el algoritmo J48 tiene, además de la misma precisión, un mayor poder predictivo (esto es, resulta más aplicable a nuevos conjuntos de datos) porque utiliza más información que la de una sola variable. Y se mantiene en un nivel de complejidad bajo y es fácilmente comprensible. Por ej., si la categoría del proyecto fuera 2 y la cantidad de aplazos no superara 11 el individuo se graduaría en 30 de 32 casos. En cambio, si la categoría del proyecto fuera 3 y la nota promedio no superara 7.33, el individuo no se graduaría en 23 de 24 casos.

Consideraciones finales

Se desprende del análisis realizado que la mayor parte de los indicadores ad hoc diseñados para ordenar a los estudiantes interesados en el programa *Delta G*, incluyendo el Orden de Mérito mismo elaborado por la comisión para la asignación de las becas, resultaron poco eficaces para predecir la posterior obtención (o no) del título. En cambio, tanto una variable aislada (la categoría del proyecto final) como un árbol sencillo que involucra sólo unas pocas variables diferentes evidenciaron una alta precisión para predecir ese resultado. Y entre ambos, el árbol resulta la herramienta de clasificación con mayor poder predictivo por su flexibilidad para aplicarse a nuevos datos.

Sería deseable entonces que la información que este análisis arroja pueda estar disponible de existir nuevas ediciones del programa. Se intentaría así contribuir a una mejor selección de indicadores que resulten informativos respecto a la obtención del título, el objetivo de *Delta G*.