ISSN: 0798-1015 DOI: 10.48082/espacios-a21v42n07p05

REVISTA ESPACIOS

Educación • Education • Educação • Vol. 42 (07) 2021 • Art. 5

Recibido/Received: 04/03/2021 • **Aprobado**/Approved: 10/03/2021 • **Publicado**/Published: 15/04/2021

Análisis comparativo de técnicas de clasificación para determinar la deserción estudiantil de la facultad de ingeniería de la Universidad de Antioquia, Colombia

Comparative analysis of classification techniques to determine student attrition in the faculty of engineering of the University of Antioquia

CHAPARRO MESA, Jorge E. [1]
CUATINDIOY IMBACHI, Jenny [2]
BARRERA LOMBANA, Nelson [3]

Resumen
El objetivo de este estudio es comparar diferentes algoritmos de clasificación de aprendizaje supervisado como las redes neuronales artificiales, métodos probabilísticos como regresión logística multinomial, métodos de ensamble como *random forest*, *bagging, boosting* y las máquinas de soporte vectorial, con el fin de identificar perfiles de posibles estudiantes desertores de la facultad de ingeniería de la Universidad de Antioquia, a partir de dos targets; número de créditos inscritos en último semestre y semestre en el cual el estudiante abandona la universidad.
Palabras clave: modelos de clasificación, aprendizaje supervisado, deserción estudiantil, facultad de ingeniería udea.

Abstract
The objective of this study is to compare different supervised learning classification algorithms such as artificial neural networks, probabilistic methods such as multinomial logistic regression, ensemble methods such as random forest, bagging, boosting and support vector machines, in order to identify profiles of possible dropouts from the engineering faculty of the University of Antioquia, based on two targets; number of credits enrolled in the last semester and semester in which the student leaves the university.
Key words: classification models, supervised learning, student dropout, udea school of engineering

1. Introducción

La deserción en educación superior es un problema que afecta a los individuos, sus familias, a las instituciones de educación superior (IES) y a la sociedad en general. En Colombia el fenómeno alcanza niveles cercanos al 50%,

[1] Ingeniero Electrónico, Especialista en Redes de Alta Velocidad, Magister en Tecnología Informática. Estudiante de doctorado en ingeniería electrónica y de computación, Universidad de Antioquia. Docente investigador Fundación Universitaria Internacional del Trópico Americano, Unitrópico. jorge.chaparro1@udea.edu.co
[2] Ingeniera Electrónica y en Telecomunicaciones, Especialista en Redes y Servicios, Magister en Ingeniería de Telecomunicaciones. Estudiante de doctorado en ingeniería electrónica y de computación, Universidad de Antioquia. Docente Investigador Universidad de Medellín, jecuantindioy@udem.edu.co
[3]Ingeniero Electrónico, Especialista en Telemática, Magister en Educación, Doctor en Ciencias de la Educación. Universidad Pedagógica y Tecnológica de Colombia UPTC. Docente Investigador Grupo de Investigación en Robótica y Automatización Industrial GIRA-UPTC. nelson.barrera@uptc.edu.co

es decir, casi la mitad de los estudiantes que inician un programa no culminan (Burgos Mantilla et al., 2009). En este sentido es necesario conocer los factores que están asociados con la deserción en las IES para poder diseñar políticas públicas e institucionales que busquen la retención y graduación de los estudiantes.

Este articulo busca identificar perfiles de estudiantes desertores de la facultad de ingeniería de la Universidad de Antioquia; para esto se contó con un dataset compuesto por 2761 registros y 58 variables con información de estudiantes desertores de las cohortes 2010-1 a 2018-1. La información del dataset se puede clasificar en dos grandes grupos, por la línea de tiempo del estudiante y por factores determinantes de la deserción estudiantil. En cuanto a línea del tiempo del estudiante se puede subclasificar en datos de preingreso, datos de ingreso e historial académico en la UdeA; mientras que por factores determinantes de la deserción estudiantil se subclasifican en cuatro categorías, socioeconómicas, individuales, institucionales y académicas. el trabajo se encaminó en dos factores académicos determinantes en la deserción de estudiantes, como son el nivel de pregrado que corresponde el semestre de la carrera en que es más probable que deserte el estudiante y el número de créditos, correspondiente a la cantidad de créditos académicos inscritos por el estudiante en el momento de la deserción.

Para realizar la clasificación de los estudiantes que han desertado de la facultad de ingeniería de la Universidad de Antioquia, se han trabajado algunos algoritmos de aprendizaje automático, utilizando modelos de aprendizaje supervisado, más exactamente modelos de clasificación multiclase, los cuales desarrollan tareas de clasificación con más de dos clases (Mortaz, 2020). Para implementar los algoritmos se utilizó el software R, así mismo para escoger los algoritmos se tuvo en cuenta que el algoritmo clasificara múltiples clases, por lo tanto, se trabajó con los siguientes: (Regresión Logística Multinomial, (*Support Vector Machines, SVM*) - Maquina de Soporte Vectorial, (*Decision Tree Classifier*) - Árbol de decisión, - *Boosting C5.0, Boosting XGBOOST, Random Forests* y Red Neuronal Artificial). Una vez realizados los modelos se definió la métrica Accuracy para evaluar los resultados obtenidos. Para este caso se separaron las observaciones en un conjunto de entrenamiento y un conjunto de validación. Finalmente se evaluó la capacidad de los modelos con el conjunto de prueba, con el fin de conocer la capacidad que tiene cada modelo cuando predice nuevas observaciones.

1.1. Estado del Arte

Según (Ramírez & Grandón, 2018), en la Universidad Católica de Chile la mayor deserción se presenta en los tres primeros semestres y una de las variables más significativas está relacionada con la clasificación de los colegios, presentando mayor deserción en los estudiantes provenientes de colegios financiados por el estado chileno. Entre las variables que presenta más significancia frente al tema de deserción se encuentran los siguientes aspectos: Edad de ingreso debido a la falta de orientación profesional y bajos recursos (Burgos Mantilla. De otra parte, en (Himmel, 2012) la deserción presentada en la Facultad de Ingeniería de la Universidad Nacional está altamente relacionada con el rendimiento académico de los estudiantes y la falta de recursos económicos. En la fase de ingreso a la universidad, se presenta una alta heterogeneidad en el nivel académico de los estudiantes de primer semestre. En la Facultad de Ingeniería de la Universidad de Cartagena, se identifican factores económicos, familiares y personales como los que potencializan la deserción en la comunidad estudiantil universitaria de esta facultad. Otro de los factores relevantes es el nivel económico de las familias que respaldan los estudiantes de esta facultad, según Ramírez & Grandón, (2018) de acuerdo con el estudio realizado para el periodo entre 2009 y 2013. Según Montes, et al,(2010), de la Universidad EAFIT, el estudio propone una división entre semestres, observándose que el rendimiento académico es una de las variables más significativas que inciden en la deserción de estudiantes de últimos semestres. En trabajo realizado por el Ministerio de Educación Nacional titulado “Deserción estudiantil en la educación superior colombiana” Burgos Mantilla et al., (2009), plantean 4 factores determinantes de la deserción estudiantil (socioeconómicos, académicos, individuales e institucionales).

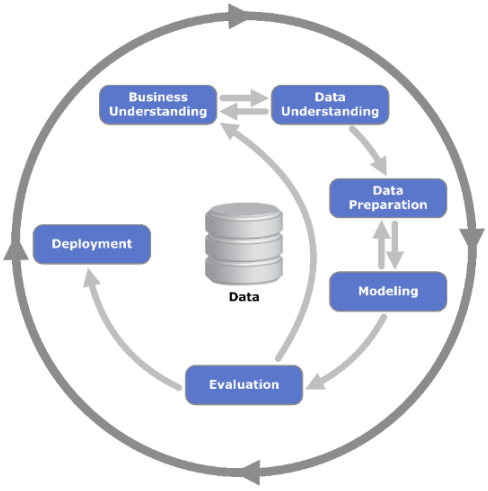
Urrego (2019), recoge los resultados más relevantes de una revisión documental de 28 investigaciones realizadas en Colombia entre el 2006 y el 2016, sobre deserción en educación superior. En parte importante de los estudios analizados en este articulo resalta la preocupación por el carácter estructural de la deserción en Colombia, con mayor incidencia en estratos socioeconómicos bajos, en los que los factores económicos se conjugan y reflejan en capital cultural y simbólico precario para la formación académica. Por otra parte Himmel (2012), divide los enfoques del análisis de la deserción y retención en cinco grandes categorías, dependiendo del énfasis que otorgan a las variables explicativas, ya sea individuales, institucionales o del medio familiar, de la siguiente forma: psicológicos, económicos, sociológicos, organizacionales y de interacciones. Referente a los modelos de clasificación para la deserción en (Olaya et al., 2020), presentan un modelo Uplift en el cual se muestran los resultados y las virtudes del modelado de mejora en la adaptación de los esfuerzos de retención en la educación superior sobre los enfoques convencionales de modelado predictivo.

Viloria et al., (2019), presentan un clasificador bayesiano aplicado a la deserción en la educación superior, la investigación propone un nuevo clasificador bayesiano simple (SBND) con Márkov de la variable de clase a una estructura de red donde se utiliza la herramienta "*Weka*" para realizar la clasificación y el modelo propuesto se compara estadísticamente con otros clasificadores bayesianos. En los resultados de investigación presentados por los autores anteriormente mencionados, se observa que los aspectos de más incidencia en la deserción están relacionados con los recursos económicos y el rendimiento académico.

2. Metodología

Para guiar el proceso de desarrollo de este trabajo se ha seguido la metodología CRISP-DM (del inglés *Cross Industry Standard Process for Data Mining*), la cual consta de 6 fases (ver fig. 1).

Figura 1
Metodología CRISP-DM para el desarrollo de los modelos



Fuente: (Schröer et al., 2021)

Como se observa en la figura 1, la metodología CRISP-DM consta de seis etapas: Comprensión del problema o negocio, Comprensión de datos, Preparación de datos, Modelado, Evaluación del modelo e Implementación del modelo. La metodología *CRISP-DM* es una de las más empleadas actualmente para el desarrollo de proyectos de minería de datos. En 1997, se puso en marcha bajo el financiamiento del Programa de Investigación y Desarrollo en Tecnologías de Información de la Unión Europea (ESPRIT) (Huber et al., 2019; Schröer et al., 2021).

3. Resultados y discusión

De acuerdo con la metodología se realiza un análisis del dataset a fin de comprender las características de los datos, como segunda instancia se seleccionan las *features* argumentando los criterios de selección. Se implementan distintos modelos con algoritmos de redes neuronales y se compara métodos probabilísticos y métodos de ensamble (*random forest*, C5.0*, bagging* y *boosting*). En el paso 4 se realiza el tuning de las variables de tal forma que el alcanzado por los modelos logren tener un *Accuracy* por encima de 70%.

3.1. Comprensión del problema

El problema en particular se trata de comparar diferentes modelos de clasificación que permitan identificar perfiles de estudiantes desertores de la facultad de ingeniería de la Universidad de Antioquia, a partir de datos históricos, individuales, académicos y socioeconómicos de estudiantes desertores.

3.2. Comprensión de los datos

3.2.1. Caracterización descriptiva de la base de datos

Los modelos fueron elaborados a partir de datos históricos con distintas variables de tipo social, económicas, académicas de ingreso, académicas de permanencia e individuales de estudiantes desertores de las cohortes 2010-1 a 2018-1 de la facultad de ingeniería de la Universidad de Antioquia. Los datos se separaron de acuerdo con los lineamientos del Ministerio de Educación Nacional (Burgos Mantilla et al., 2009), sin embargo se trabajan solo los componentes, académicos, individuales, y socioeconómicos, teniendo en cuenta que el dataset no cuenta con información institucional (ver fig. 2).

Figura 2
Clasificación de los factores determinantes
en la deserción estudiantil.

Fuente: Elaboración propia con base en (Burgos Mantilla et al., 2009)

3.2.2. Exploración de datos

En esta fase se realizan las gráficas de caja y bigotes, histogramas y análisis de correlación, además se escogen las variables respuesta. El objetivo es desarrollar modelos de clasificación que permitan identificar perfiles de estudiantes desertores de la facultad de ingeniería de la Universidad de Antioquia, a partir del número de créditos inscritos en el último semestre cursado y el semestre en el cual el estudiante abandona su proceso de formación. El dataset está compuesto por 2761 registros y 58 variables, de las cuales se escogen 2 variables respuesta para los modelos de clasificación:

Variable Respuesta $V1 =$ Numero de créditos inscritos en el último semestre cursado. Esta variable se agrupa en 4 grupos por rangos de 8 créditos, teniendo en cuenta que nadie matricula un único crédito debido que las asignaturas tienen mínimo 2 créditos académicos.

Variable Respuesta $V2 =$ Semestre en el cual el estudiante abandona su proceso de formación. En esta variable se agruparon los semestres 7 a 11 teniendo en cuenta que el número de registros es demasiado bajo.

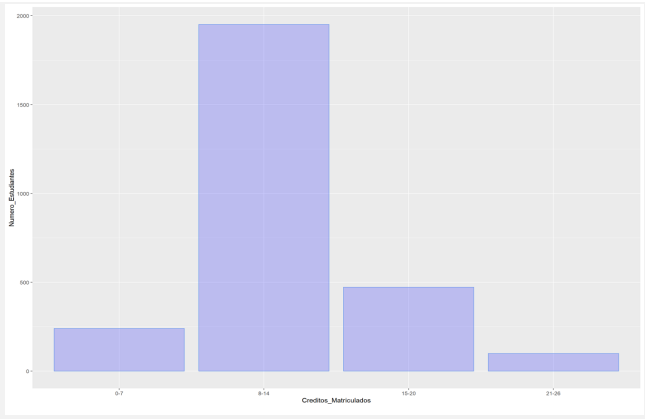
A. Exploración de variables cuantitativas

A continuación, se observa un gráfico de barras de las variables respuesta V1 y V2.

Figura 3

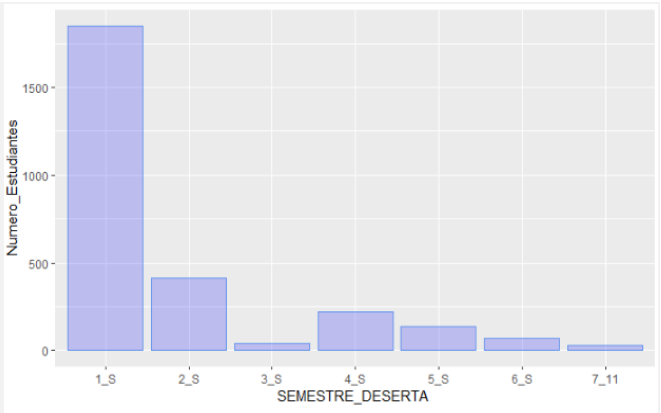
Gráficos de barras variables V1 y V2

a) Gráfico de barras variable respuesta V1



En la Figura 3-a) se observa que la mayoría de los estudiantes desertores tenían inscritos entre 7 y 14 créditos al momento de la deserción.

b) Gráfico de barras variable respuesta V1



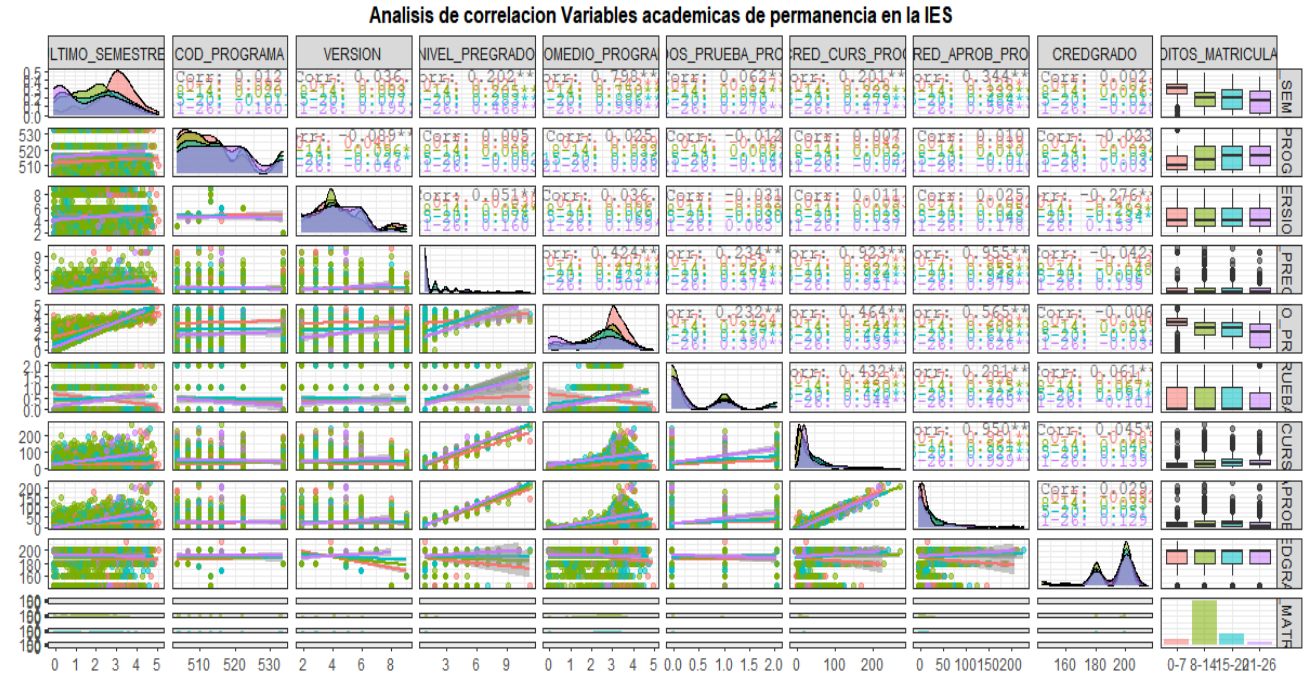
En la Figura 3-b), se observa que la mayoría de los estudiantes desertan de la educación superior en los primeros dos semestres, especialmente en el semestre 1

B. Análisis de variable V1 - correlación y *Boxplot*

El análisis de correlación consiste en un procedimiento estadístico para determinar si dos o más variables están relacionadas o no. El resultado del análisis es un coeficiente de correlación que puede tomar valores entre -1 y +1. A continuación, en la Figura 4 se observa el grafico de correlación y *boxplot* de las variables cuantitativas.

Figura 4

Correlación y *Boxplot* de variables académicas de permanencia en la IES, vs. número de créditos matriculados.

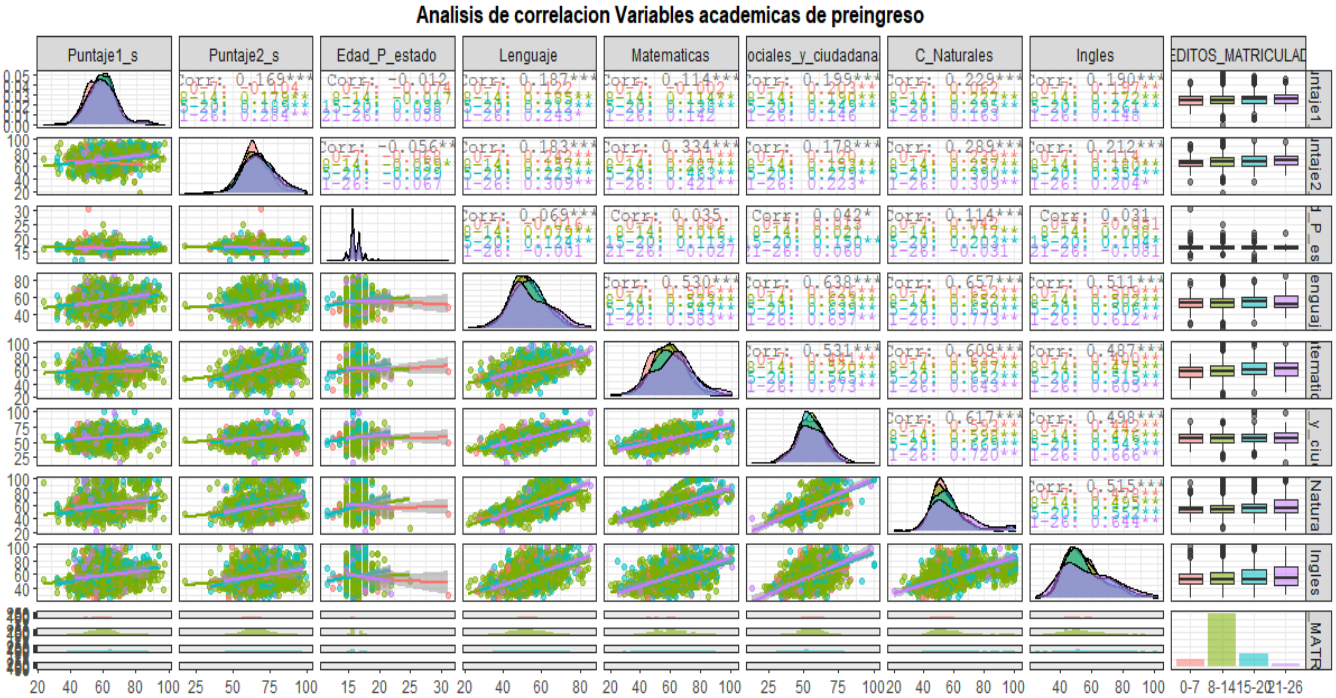


Fuente: autores

En la Figura 4, en la caja de bigotes se observa una diferencia de los estudiantes que tenían entre 0 y 7 créditos inscritos, esto podría ser interesante ya que según los datos arrojados en el estado del arte esta es una característica de estos estudiantes. A continuación, en la Figura 5, se muestra la gráfica de correlación entre variables académicas de preingreso y el número de créditos matriculados.

Figura 5

Análisis de correlación entre variables académicas de preingreso y el número de créditos matriculados



Fuente: autores

En la Figura 5 no se observan diferencias significativas entre las variables académicas y el número de créditos matriculados. Existe alta correlación entre estas variables y se podría decir que el nivel de inglés se relaciona con los alumnos que tienen más créditos inscritos.

Exploración de variables cuantitativas con la variable respuesta V2

Respecto a la variable $\boldsymbol{V2}$ = Semestre en el cual el estudiante abandona su proceso de formación se agruparon los semestres 7 a 11 teniendo en cuenta que el número de registros es demasiado bajo.

En la Figura 6, se observa que la mayoría de los estudiantes desertan de la educación superior en los primeros dos semestres, especialmente en el semestre 1.

Figura 6

Análisis de correlación entre variables académicas
de preingreso y el número de créditos matriculados.

Fuente: autores

En la Figura 6 se puede observar promedio del último semestre terminado, tiene una diferencia significativa respecto al semestre de deserción, es decir que los promedios más bajos desertan en el primer semestre. Por otra parte, entre menos créditos haya cursado mayor es el número de deserción.

Figura 7

Correlación y *Boxplot* de variables académicas de preingreso, vs. semestre de deserción

Fuente: autores

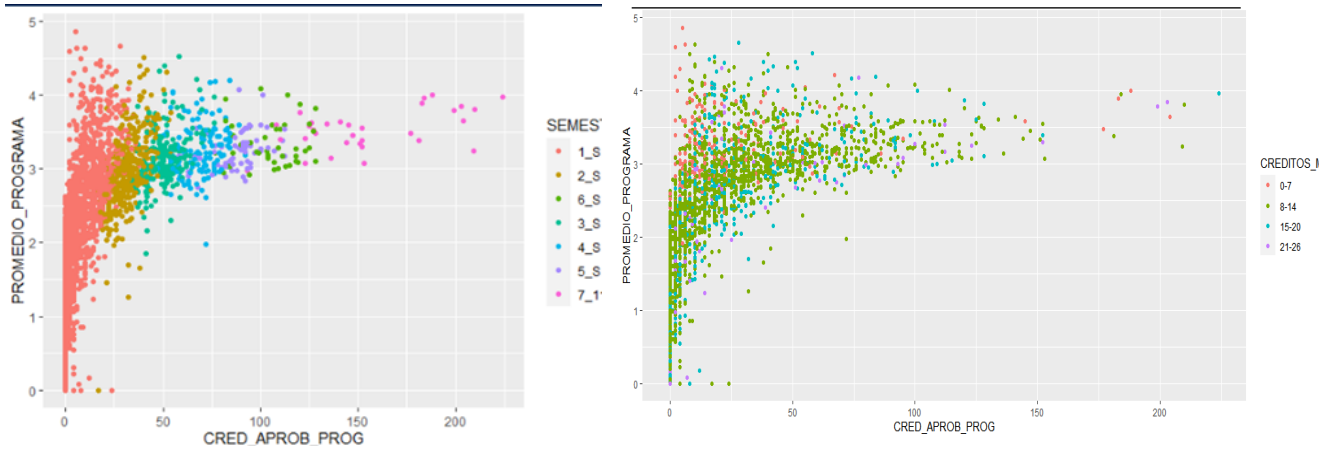
En la Figura 7 no se observan diferencias significativas entre las variables académicas y el nivel del semestre en el cual el alumno decide desertar. Existe alta correlación entre estas variables y se podría decir que el nivel de inglés se relaciona con los alumnos que desertan en los últimos semestres.

A continuación, se realiza una gráfica de las variables respuesta V1 y V2, respecto a 2 de las variables más significativas analizadas en las Figuras anteriores, el promedio y el número de créditos aprobados.

Figura 8

Correlación de las variables respuesta V1 y V2 con dos de las variables más significativas (créditos aprobados y el promedio)

$V2 =$ Variable Nivel $V1 =$ Variable Créditos último semestre



Fuente: autores

En la Figura 8, se observa la correlación de las variables respuesta V1= Número de créditos inscritos en el último semestre y la variable V2= Semestre en el cual el estudiante abandona su proceso de formación, respecto a dos de las variables más significativas (créditos aprobados y el promedio) que se encontró en el análisis exploratorio de los datos. Por otra parte, la variable V2 separa mejor los datos que la variable V1. Esto es importante porque se esperaría que los modelos de clasificación respecto a la Variable V2 tengan mejor rendimiento a la hora de clasificar que los modelos implementados con la variable V1.

3.3. Preparación de datos

Una vez identificadas las dos variables respuesta se procede a realizar la preparación de los datos para la implementación de los modelos. En esta fase se realiza la agrupación de los datos y el escalamiento de las variables. Por otro lado, se realiza la separación de datos en conjunto de entrenamiento y validación. Para entrenar y validar los modelos propuestos se separaron los datos en proporciones de 80% para entrenamiento y 20% para validación.

El desarrollo de cada uno de los modelos implementados en este trabajo se elaboró con el software R. El proceso es similar para cada uno de los modelos, inicialmente se corre el modelo, se entrena con los datos de entrenamiento, se evalúa con los datos de testeo, se calcula la matriz de confusión y las métricas de desempeño y de acuerdo con el resultado se optimizan los valores de los hiper parámetros y se corren nuevamente hasta lograr el máximo rendimiento.

3.4. Implementación de Modelos de clasificación multiclase

Para la implementación de los modelos se utilizó el software R (R-project.org, n.d.). A continuación se desarrollan los modelos de clasificación a través del aprendizaje supervisado, el cual se basa en entrenar a un modelo o método por medio de diferentes datos para poder predecir una variable partiendo de estos mismos datos (Li et al., 2020). Los problemas de clasificación parten de un conjunto de datos la cual tiene un conjunto de características y conocemos la clase a la cual pertenece llamándose a este conjunto de entrenamiento o aprendizaje, creando un conjunto de reglas el cual nos permiten validar con un conjunto de datos diferente, permitiendo estimar la precisión del modelo de clasificación.

Los algoritmos escogidos para realizar el proceso de clasificación son multiclase y se trabajó con los siguientes algoritmos: Regresión Logística Multinomial, (*Support Vector Machines, SVM*) - Maquina de Soporte Vectorial, *(Decision Tree Classifier)* - Árbol de decisión, - *Boosting C5.0*, *Boosting XGBOOST*, *Random Forests* y Red Neuronal Artificial.

Métrica de desempeño

La métrica de desempeño utilizada para comparar los modelos y para evaluar la eficiencia y el ajuste de los hiper parámetros será el *Accuracy*.

Esta métrica indica el número de elementos clasificados correctamente en comparación con el número total de elementos. Así mismo *Accuracy* funciona muy bien siempre y cuando las clases estén equilibradas, siendo este ese caso ya que se cuenta con igual número de observaciones para cada clase.

$$\boldsymbol{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

3.4.1. Regresión Logística Multinomial

Para trabajar este modelo logístico multinomial se utiliza GLM multivariante usando $multinom\ ()$ desde el paquete $nnet$. En estadística, la regresión logística multinomial generaliza el método de regresión logística para problemas multiclase, es decir, con más de dos posibles resultados discretos (Rahman & Baker, 2018).

A. Modelo Regresión Logística – Variable V1

Este modelo obtiene una medida de *Accuracy* de `70.21014 %`

```
# Model accuracy
mean(predicted.classes == test_set$CREDITOS_MATRICULADOS)
## [1] 0. 7021014
```

B. Modelo Regresión Logística – Variable V2

Este modelo obtiene una medida de *Accuracy* de `79.8913 %.`

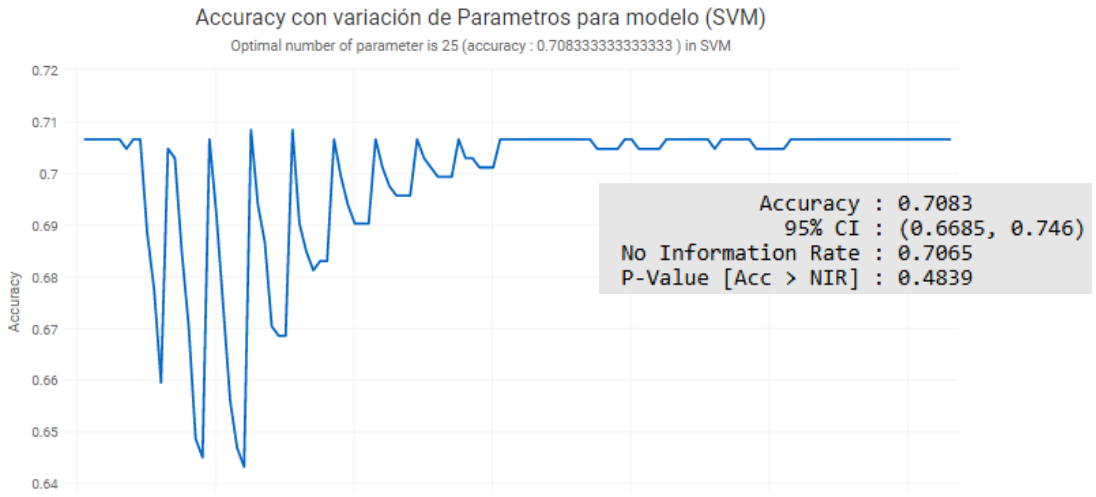
```
# Model accuracy
mean(predicted.classes == test_set$SEMESTRE_DESERTA)
## [1] 0.798913
```

3.4.2. SVM Maquina de Soporte Vectorial con Kernel lineal

Las máquinas de soporte vectorial o máquinas de vector soporte (del inglés *Support Vector Machines, SVM*) son un conjunto de algoritmos de aprendizaje supervisado desarrollados por Vladimir Vapnik y su equipo en los laboratorios AT&T (Liu et al., 2020). Estos métodos están propiamente relacionados con problemas de clasificación y regresión. Dado un conjunto de ejemplos de entrenamiento (de muestras) podemos etiquetar las clases y entrenar una SVM para construir un modelo que prediga la clase de una nueva muestra (Liu et al., 2020).

A. Modelo SVM - Variable V1

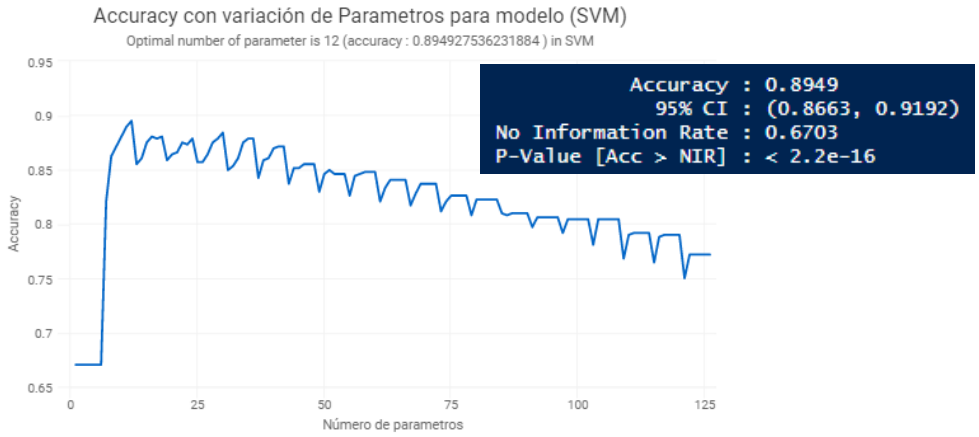
Figura 9
Modelo SVM Variable V1



Fuente: autores

B. Modelo SVM - Variable V2

Figura 10
Modelo SVM Variable V2



Fuente: autores

3.4.3. Modelo Random Forests

Random Forests: Los bosques aleatorios (RF) es un poderoso algoritmo de modelo de distribución de especies (SDM). Este modelo de conjunto por defecto puede producir mapas de distribución de especies categóricos y numéricos basados en sus algoritmos de árbol de clasificación (CT) y árbol de regresión (RT), respectivamente (Zhang et al., 2019).

Cuadro 1

Medidas de desempeño modelos
Random Forest variables V1 y V2

Modelo Random Forest - Variable V1	Modelo Random Forest - Variable V2
Accuracy : 0.7609 95% CI : (0.723, 0.7959) No Information Rate : 0.933 P-Value [Acc > NIR] : 1	Accuracy : 0.8659 95% CI : (0.8346, 0.8932) No Information Rate : 0.692 P-Value [Acc > NIR] : < 2.2e-16

Fuente: autores

3.4.4. Modelo Decision Tree Classifier

Los árboles de decisión o de clasificación son un modelo surgido en el ámbito del aprendizaje automático (*Machine Learning*) y de la Inteligencia Artificial que, partiendo de una base de datos, crea diagramas de construcciones lógicas que nos ayudan a resolver problemas. A esta técnica también se la denomina segmentación jerárquica. (Sadeghi et al., 2012).

Los árboles de clasificación se asemejan mucho a los árboles de regresión, con la diferencia de que predicen variables respuestas cualitativas en lugar de continuas. Este modelo se corre con la librería *library(rpart)* de R.

Cuadro 2

Medidas de desempeño modelos decisión
Tree Classifier variables V1 y V2

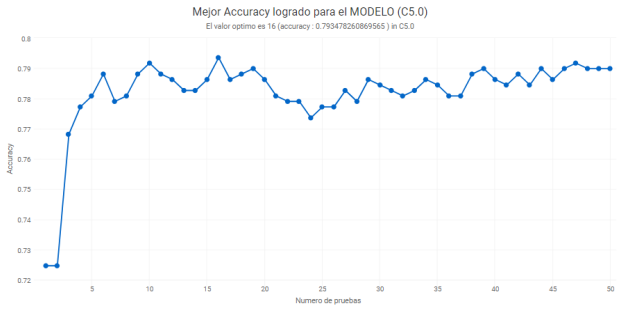
Modelo *Decision Tree Classifier* - V1	Modelo *Decision Tree Classifier* - V2
Accuracy : 0.7862 95% CI : (0.7496, 0.8197) No Information Rate : 0.8442 P-Value [Acc > NIR] : 0.9999	Accuracy : 0.8804 95% CI : (0.8504, 0.9063) No Information Rate : 0.663 P-Value [Acc > NIR] : < 2.2e-16

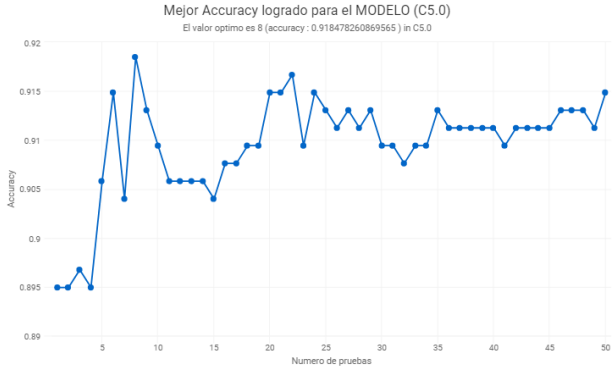
Fuente: autores

3.4.5. Modelo C5.0.

El árbol de decisión de inducción de arriba hacia abajo (TDIDT, también llamado algoritmo C5.0), es uno de los métodos más utilizados. La construcción del árbol continúa comenzando con el conjunto completo de ejemplos de capacitación. En cada paso, el atributo más informativo se selecciona como la raíz del árbol y el conjunto de entrenamiento actual se divide en subconjuntos de acuerdo con los valores de los atributos seleccionados (Sadeghi et al., 2012).

Cuadro 3
Medidas de desempeño
modelo C5.0. variables V1 y V2






```
          Accuracy : 0.7935
            95% CI : (0.7573, 0.8265)
No Information Rate : 0.7065
P-Value [Acc > NIR] : 2.263e-06
```

```
          Accuracy : 0.9312
            95% CI : (0.9067, 0.9508)
No Information Rate : 0.6703
P-Value [Acc > NIR] : < 2.2e-16
```

Fuente: autores

3.4.6. Modelo (*BOOSTING*) – *XGBOOST*

Trabaja sobre árboles de decisiones, pero potenciando los resultados de estos, debido al procesamiento secuencial de la data con una función de pérdida o coste, la cual, minimiza el error iteración tras iteración, haciéndolo de esta manera, un pronosticador fuerte (Zhong et al., 2020).

A. Modelo *XGBoost* V1

```
accuracy_XGBOOST <-mean(predicted.classes == test_set$CREDITOS_MATRICULADOS)
accuracy_XGBOOST
```

```
## [1] 0.8097826
```

B. Modelo *XGBoost*=V2

```
accuracy_XGBOOST <-mean(predicted.classes == test_set$SEMESTRE_DESERTA)
accuracy_XGBOOST
```

```
## [1] 0.9202899
```

3.4.7. Modelo Red Neuronal Artificial (RNA)

Existen diferentes algoritmos para programar una red neuronal, depende inicialmente del problema si es clasificación o regresión. Para este caso se trabarán con algoritmos de clasificación (Toğaçar et al., 2020).

A. Cálculo de la medida de desempeño *Accuracy* para la RNA Variable 1

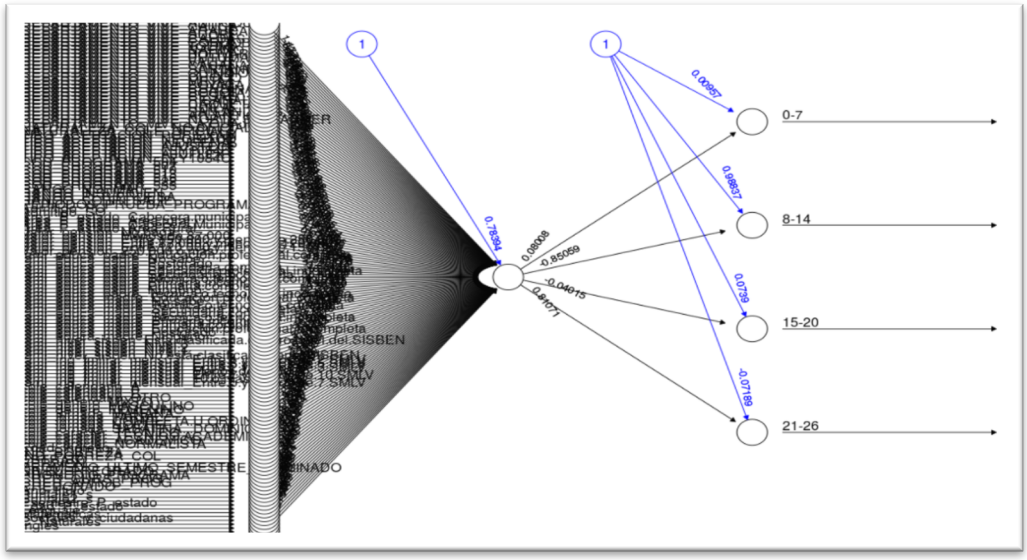
A continuación, se muestra el resultado de la medida de *Accuracy* del modelo RNA para la variable 1

```
accuracy_RNA <-round((sum(diag(matrizConfusion))/sum(matrizConfusion)),4)
accuracy_RNA
```

```
## [1] 0.6848
```

Figura 11

Red Neuronal Artificial Variable V1



Fuente: autores

B. Cálculo de la medida de desempeño *Accuracy* para la RNA Variable 2

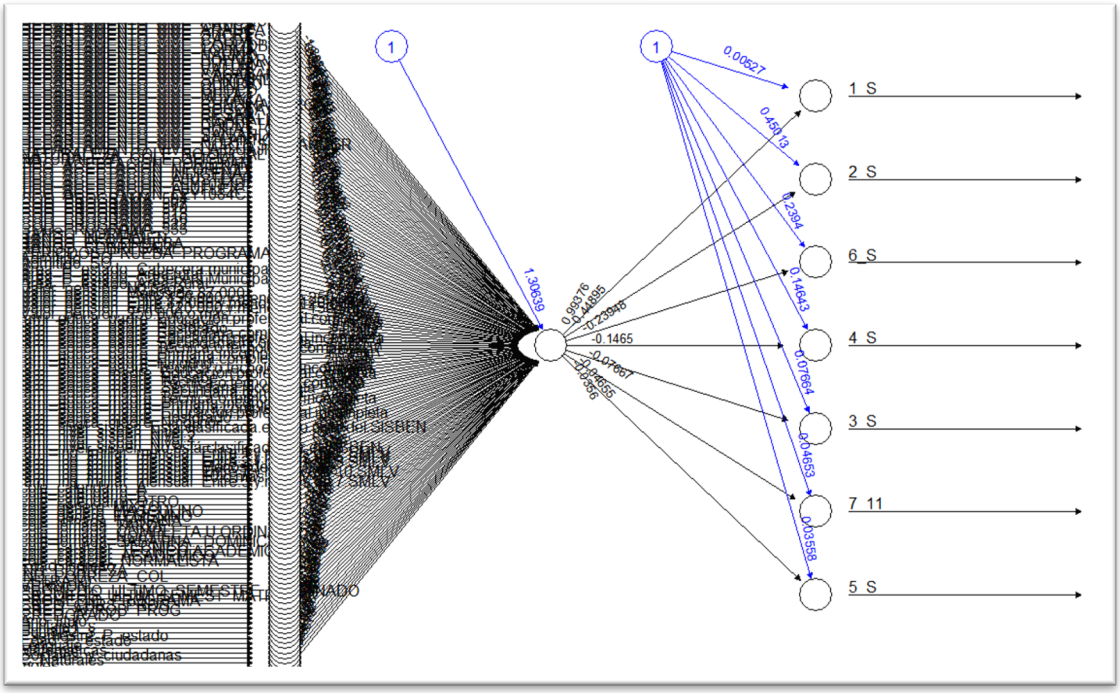
A continuación, se muestra el resultado de la medida de *Accuracy* del modelo RNA para la variable 2

```
accuracy_RNA <-round((sum(diag(matrizConfusion))/sum(matrizConfusion)),4)
accuracy_RNA
```

```
## [1] 0.7681
```

Figura 12

Red Neuronal Artificial Variable V1



Fuente: autores

3.4.8. Discusión

Una vez que se han entrenado y optimizado los diferentes modelos, se tiene que identificar cuál de ellos consigue mejores resultados para el problema en cuestión; en este caso, predecir el Número de créditos inscritos en el último semestre y el Semestre en el cual el estudiante abandona su proceso de formación.

Con los datos disponibles, existen dos formas de comparar los modelos. Si bien las dos no tienen por qué dar los mismos resultados, son complementarias a la hora de tomar una decisión final.

Una de las principales fases en todo el proceso de análisis de datos es la evaluación del modelo. Se trata de evaluar la calidad del modelo aprendido, cuantificado mediante unas varias métricas que evalúan el rendimiento del modelo. Esta comparación permite determinar cuáles modelos presentan mejores resultados.

A. Análisis de resultados para la variable V1

Con base en la métrica escogida a continuación en el cuadro 4 se presentan los resultados de la métrica de desempeño *Accuracy* de los 7 modelos implementados para la **V1** = Número de créditos inscritos en el último semestre

Cuadro 4

Tabla comparativa del *Accuracy* variable V1

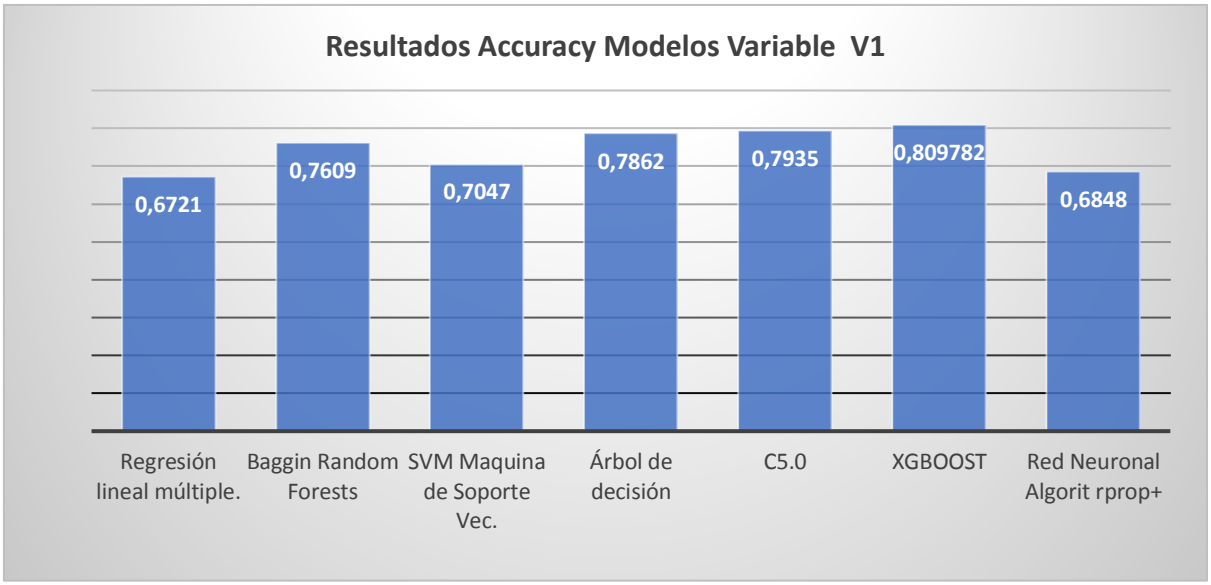
#	Modelo	*Accuracy*
1	Regresión lineal múltiple.	0.67210
2	Baggin Random Forests	0.7609
3	SVM Maquina de Soporte Vec.	0.7047
4	Árbol de decisión	0.7862
5	C5.0	0.7935
6	**XGBOOST**	**0.809782**
7	Red Neuronal Algorit rprop+	0.6848

Fuente: autores

En la Figura 13 se observa la gráfica del resultado de los *Accuracy* de los modelos para V1.

Figura 13

Grafica del *Accuracy* de los modelos realizados con la variable V1



Fuente: autores

A continuación, en el cuadro 5 se presenta el *Accuracy* de los 7 modelos implementados para la variable **V2**.

Cuadro 5

Tabla comparativa del *Accuracy* variable V2

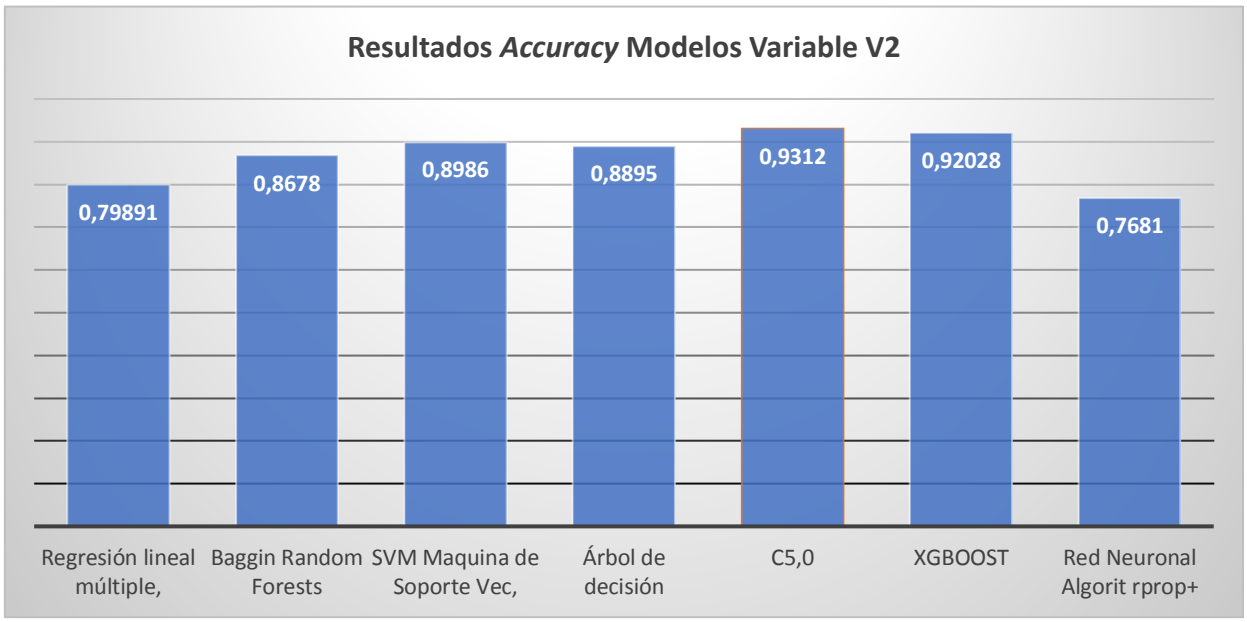
#	Modelo	*Accuracy*
1	Regresión lineal múltiple.	0.79891
2	Baggin Random Forests	0.8678
3	SVM Maquina de Soporte Vec.	0.8986
4	Árbol de decisión	0.8895
5	**C5.0**	**0.9312**
6	XGBOOST	0.92028
7	Red Neuronal Algorit rprop+	0.7681

Fuente: autores

En la Figura 14 se observa la gráfica del resultado de los *Accuracy* de los modelos para la variable V2.

Figura 14

Grafica del *Accuracy* de los modelos realizados con la variable V2



Fuente: autores

Con base en los resultados obtenidos en los cuadros 4 y 5 se puede ver que en **V1** = Número de créditos inscritos en el último semestre, se obtuvo un *Accuracy* máximo de **0.809782** con el algoritmo (*BOOSTING*) - *XGBOOST* - (*Extra Gradient Boosting*) Árboles de clasificación potenciados.

Este método *Boosting* trabaja sobre árboles de decisiones, pero potenciando los resultados fuertes de estos, debido al procesamiento secuencial de los datos con una función de pérdida o coste, la cual, minimiza el error iteración tras iteración, haciéndolo de esta manera, un pronosticador.

De otra parte, para la variable **V2** = Semestre en el cual el estudiante abandona su proceso de formación, se obtuvo un *Accuracy* máximo de **0.9312** con el algoritmo (*BOOSTING*) – c5.0.

Estos resultados nos permiten concluir que los métodos de ensamble resultaron ser los mejores modelos para clasificación de estudiantes desertores. Un hallazgo interesante es que en las dos variables analizadas los métodos de ensamble obtuvieron mejores rendimientos. Es interesante ver también que modelos como *Random Forest* que no requieren gran capacidad de cómputo y entregan los resultados muy rápido también obtuvieron muy buenos valores de rendimiento comparado con los otros métodos.

Otro de los modelos muy eficientes con la variable V2 fue la máquina de soporte vectorial, ya que obtuvo un *Accuracy* muy alto comprado con otros métodos como la regresión logística. Se debe tener en cuenta que algunos datos presentan muchos puntos atípicos, por lo tanto, los mejores modelos que clasificaron son aquellos que tiene poca afectación de los puntos atípicos.

Analizando las 2 variables en conjunto podemos concluir que los estudiantes tienden a desertar en el primer semestre y el promedio de créditos para la mayoría de los desertores esta entre 8 y 16, esto corresponde a la cantidad de créditos que un alumno matricula en un semestre académico.

Como se observa en los modelos realizados la deserción en los estudiantes es el resultado de la combinación y efecto de distintas variables. En éstas se encuentran características preuniversitarias, familiares, individuales y socioeconómicas. El trabajar estos dos modelos de forma conjunta con las variables académicas se puede observar que es más fácil clasificar a los estudiantes por el semestre en el cual desertan de la institución que por el número de créditos académicos matriculados en el último semestre.

4. Conclusiones

Los modelos de clasificación son herramientas interdisciplinarias muy importantes para entender la complejidad que caracteriza los sistemas académicos. El trabajar con diferentes variables categóricas fue un reto interesante sin embargo se logro con base en el estado del arte organizar las variables de tal forma que los modelos tuvieron un excelente desempeño

Los modelos utilizados fueron modelos de clasificación multi-clase, por lo tanto, no se pudo trabajar con aquellos modelos que funcionan para predecir sistemas binarios. Los métodos de ensamble de modelos o métodos combinados como el *XGBOOST* y *C5.0*, son métodos muy potentes y no dependen mucho del comportamiento de los datos ni de los puntos atípicos, esto logra medidas de rendimiento muy buenas.

Otro modelo que funcionó muy bien con pocos ajustes de hiper parámetros fue la Maquina de Soporte Vectorial. En cuanto a la RNA clasifica muy bien siempre y cuando se entrene con los parámetros adecuados.

El presente trabajo tuvo por objetivo comparar siete técnicas de predicción para la deserción estudiantil, a la vez que comparar dos modelos de clasificación con dos variables relacionadas pero diferentes, resultando en un ejercicio académico bastante significativo ya que se puedo ver que los modelos trabajan sus medidas de rendimiento de acuerdo con las características de las variables a clasificar y el comportamiento de los datos.

El clasificar las variables en académicas, institucionales, socioeconómicas e individuales de acuerdo con los lineamientos del Ministerio de educación dan una certeza de que los datos arrojados por estos modelos siguen unos estándares nacionales.

Los resultados obtenidos con la medida de rendimiento *Accuracy* nos permiten concluir que los métodos de ensamble resultaron ser los mejores modelos para clasificación de estudiantes desertores.

Referencias bibliográficas

Burgos Mantilla, G., Victoria Angulo, M., Guzmán Ruiz, C., Guzmán Ruiz Diana Durán Muriel Jorge Franco Gallego, C., Castaño Vélez Santiago Gallón Gómez Karoll Gómez Portilla Johanna Vásquez Velásquez, E., & Gómez Díaz Fotografía de carátula, K. (2009). *Deserción estudiantil en la educación superior colombiana* (Ministerio de Educación Nacional (ed.); Primera). Ministerio de Educación Nacional. www.mineducacion.gov.co

Himmel, E. (2012). Modelo de análisis de la deserción estudiantil en la educación superior. *Calidad En La Educación*, *17*, 91. https://doi.org/10.31619/caledu.n17.409

Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications - A holistic extension to the CRISP-DM model. *Procedia CIRP*, *79*, 403–408. https://doi.org/10.1016/j.procir.2019.02.106

Li, J., Zhong, P. an, Yang, M., Zhu, F., Chen, J., Liu, W., & Xu, S. (2020). Intelligent identification of effective reservoirs based on the random forest classification model. *Journal of Hydrology*, *591*, 125324. https://doi.org/10.1016/j.jhydrol.2020.125324

Liu, M. Z., Shao, Y. H., Li, C. N., & Chen, W. J. (2020). Smooth pinball loss nonparallel support vector machine for robust classification. *Applied Soft Computing*, 106840. https://doi.org/10.1016/j.asoc.2020.106840

Mortaz, E. (2020). Imbalance accuracy metric for model selection in multi-class imbalance classification problems. *Knowledge-Based Systems*, *210*, 106490. https://doi.org/10.1016/j.knosys.2020.106490

Olaya, D., Vásquez, J., Maldonado, S., Miranda, J., & Verbeke, W. (2020). Uplift Modeling for preventing student dropout in higher education. *Decision Support Systems*, *134*, 113320. https://doi.org/10.1016/j.dss.2020.113320

R-project.org. (n.d.). *R: El Proyecto R para Computación Estadística*. Retrieved March 1, 2021, from https://www.r-project.org/

Rahman, M. L., & Baker, D. (2018). Modelling induced mode switch behaviour in Bangladesh: A multinomial logistic regression approach. *Transport Policy*, *71*, 81–91. https://doi.org/10.1016/j.tranpol.2018.09.006

Ramírez, P. E., & Grandón, E. E. (2018). Predicción de la Deserción Académica en una Universidad Pública Chilena a través de la Clasificación basada en Árboles de Decisión con Parámetros Optimizados Prediction of Student Dropout in a Chilean Public University through Classification based on Decision Trees with Optimized Parameters. *Versión Final Ene*, *11*(3), 3–10. https://doi.org/10.4067/S0718-50062018000300003

Sadeghi, R., Zarkami, R., Sabetraftar, K., & Van Damme, P. (2012). Application of classification trees to model the distribution pattern of a new exotic species Azolla filiculoides (Lam.) at Selkeh Wildlife Refuge, Anzali wetland, Iran. *Ecological Modelling*, *243*, 8–17. https://doi.org/10.1016/j.ecolmodel.2012.06.011

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Toğaçar, M., Ergen, B., & Cömert, Z. (2020). Classification of flower species by using features extracted from the intersection of feature selection methods in convolutional neural network models. *Measurement: Journal of the International Measurement Confederation*, *158*, 107703. https://doi.org/10.1016/j.measurement.2020.107703

Urrego, M. R. (2019). La investigación sobre deserción universitaria en Colombia 2006-2016. Tendencias y resultados. *Revista Pedagogía y Saberes - Universidad Pedagógica Nacional Facultad de Educación*, *51*(2019. pp. 49–66 La), 49–66.

Viloria, A., Lezama, O. B. P., & Varela, N. (2019). Bayesian classifier applied to higher education dropout. *Procedia Computer Science*, *160*, 573–577. https://doi.org/10.1016/j.procs.2019.11.045

Zhang, L., Huettmann, F., Zhang, X., Liu, S., Sun, P., Yu, Z., & Mi, C. (2019). The use of classification and regression algorithms using the random forests method with presence-only data to model species' distribution. *MethodsX*, *6*, 2281–2292. https://doi.org/10.1016/j.mex.2019.09.035

Zhong, R., Johnson, R., & Chen, Z. (2020). Generating pseudo density log from drilling and logging-while-drilling data using extreme gradient boosting (XGBoost). *International Journal of Coal Geology*, *220*, 103416. https://doi.org/10.1016/j.coal.2020.103416