

Un modelo de minería de datos para la predicción de la deserción estudiantil en la Facultad de Ingeniería Fisicomecánicas de la Universidad Industrial de Santander en el periodo 2015-2019.

Juliana Hernández Mosquera^{1,2,3,·}

¹ Escuela de Estudios Industriales y Empresariales (EEIE).

² Facultad de Ingeniería Fisicomecánicas.

³ Universidad Industrial de Santander.

RESUMEN

Por medio de esta investigación se busca encontrar un modelo de minería de datos que clasifique a los estudiantes en dos grupos, el grupo de estar en normalidad académica o estar en el grupo de riesgo de deserción estudiantil. Para el desarrollo de los modelos se usaron datos suministrados por la Dirección de Admisiones y Registro Académico de la Universidad Industrial de Santander, de estudiantes asociados a la Facultad de Ingeniería Fisicomecánicas en el periodo de 2015 a 2019. Para el modelamiento se utilizó el lenguaje de programación Python y la aplicación de código abierto Jupyter Notebook; se eligieron los modelos según la naturaleza de los datos, que para esta investigación son modelo de Árbol de Decisión (Decision Tree), modelo de Bosques Aleatorios (Random Forest) y modelo de Regresión logística (Logistic Regression). Con los tres modelos se obtuvieron resultados de entre 98 % a 99 % de precisión en la clasificación de los estudiantes, al optimizar los modelos utilizando el método de validación cruzada (Cross Validation) el único modelo que no tuvo una disminución en su precisión fue el de Regresión Logística. Por esto se concluye que el modelo que mejor desempeño tiene es el de Regresión Logística.

Palabras Clave: minería de datos, predicción, estudiantes, árboles de decisión, minería de datos educativos, deserción, estudiantes de pregrado, instituciones de educación superior, modelo predictivo.

ABSTRAC

Through this research, we aim to find a data mining model that classifies students into two groups: the group of being in academic normality or being in the group at risk of dropping out. For the development of the models, data provided by the Dirección de Admisiones y Registro Académico de la Universidad Industrial de Santander was used. The data pertains to students associated with the Faculty of Physicomechanical Engineering from the period 2015 to 2019. Python programming language and the open-source Jupyter Notebook application were used for modeling. The models were chosen based on the nature of the data. For this research, Decision Tree model, Random Forest model and Logistic Regression model were employed. With all three models, classification results ranging from 98% to 99% accuracy were obtained. Optimizing the models using the Cross Validation method, the only model that did not

experience a decrease in accuracy was the Logistic Regression model. Therefore, it is concluded that the Logistic Regression model performs the best.

Keywords: data mining, prediction, students, decision trees, educational data mining, dropout, undergraduate students, higher education institutions, predictive model.

CONTENIDO

I.	Introducción.....	2
II.	Revisión de la literatura.....	3
	1°. Antecedentes.....	3
	2°. Hipótesis.....	4
III.	Planteamiento del problema.....	4
IV.	Análisis preliminar de los datos.....	5
	1°. Análisis por características.....	5
	2°. Análisis multivariable.....	7
V.	Modelado.....	10
	1°. Modelo de Árbol de Decisión.....	11
	2°. Modelo de Bosques Aleatorios.....	12
	3°. Modelo de Regresión Logística.....	14
VI.	Optimización y validación de los modelos.....	16
VII.	Elección del modelo predictivo.....	19
VIII.	Conclusiones.....	19
IX.	Recomendaciones.....	20
X.	Bibliografía.....	20

I. INTRODUCCIÓN

““Aprender es como remar contra corriente: en cuanto se deja, se retrocede.”. – Edward Benjamin Britten.

La deserción en la educación superior es un fenómeno que afecta directamente el desarrollo social, económico, cultural y político de un país; a nivel mundial los investigadores intentan identificar las causas de la deserción, aunque se ha llegado a hallazgos favorables para las instituciones, aún no existe una explicación que pueda aplicarse a todos los casos y en ocasiones

las investigaciones son inconclusas, esto se debe en parte a la diversidad cultural y de condiciones de cada región.

La educación superior es un privilegio al que muy pocos colombianos tienen acceso, aun así, existen altos índices de abandono a los estudios técnicos, tecnológicos y profesionales; el gobierno nacional en el afán de comprender este fenómeno ha creado instituciones y programas especializados para dar apoyo a los estudiantes, sin embargo, los esfuerzos del estado han resultado favorables, pero con baja cobertura.

El intento de encontrar una solución a este problema en la Universidad Industrial de Santander impulsa a realizar la presente investigación; la deserción estudiantil en Santander puede causar efectos negativos en la calidad de vida de los santandereanos, encontrando baja mano de obra calificada, aumento en la delincuencia e incluso aumento de enfermedades mentales tales como la depresión.

Esta investigación propone encontrar un modelo predictivo para la detección temprana del riesgo de abandono escolar en estudiantes de la Universidad Industrial de Santander (UIS), con el fin de disminuir las brechas de desigualdad e inequidad causadas por la no culminación de los estudios de pregrado; trabajando con información sobre las características socioeconómicas y sociodemográficas de estudiantes de la Facultad de Ingeniería Fisicomecánicas de la UIS en el rango de tiempo de 2015 a 2019; la información reposa en las bases de datos de la Dirección de Admisiones y Registro académico, considerada una fuente de información secundaria.

Inicialmente se realiza una revisión literaria dentro de las bases de datos de la Biblioteca de la UIS, para encontrar referentes dentro del tema investigativo; seguido se hace el análisis exploratorio de la información para organizar, limpiar, depurar y asignar variables a los datos correspondientes; así mismo encontrar posibles correlaciones entre las variables, que sirvan de punto de inicio en la elección del software a utilizar y el método.

El problema de investigación es un problema de clasificación, se intenta dar una respuesta binaria sobre la posible deserción de un estudiante; después del procesamiento de los datos se prueban modelos predictivos existentes como

árbol de decisión, regresión logística, Naive Bayes, entre otros; las estadísticas nos ayudarán a elegir el modelo que tenga mayor precisión con respecto a los demás, para realizar la evaluación de los modelos, estudiando minuciosamente la matriz de confusión, la exactitud, la precisión, la sensibilidad y la puntuación F1; que son las características de confiabilidad de un modelo.

Se espera encontrar un modelo que se ajuste a los datos y tenga un buen margen de confiabilidad para así recomendar acciones correctivas a la dependencia encargada de la disminución de la deserción estudiantil en la UIS.

II. REVISIÓN DE LA LITERATURA

1º. Antecedentes

Según Tinto (1975) la deserción se define como el procedimiento que lleva a cabo un estudiante universitario cuando abandona voluntaria o forzosamente sus estudios, debido a la influencia negativa o positiva de factores internos o externos; esta definición de deserción estudiantil aún tiene vigencia en la actualidad y por ello se trae a colación.

Así mismo el autor resalta que la deserción puede estar ligada a las metas individuales del estudiante, las cuales tienen influencia de los procesos sociales e intelectuales que experimente el estudiante a lo largo del desarrollo de su programa académico. Sólo algunas deserciones son causadas por bajo rendimiento académico, puesto que la mayoría son voluntarias y los estudiantes que las realizan pueden tener un mayor rendimiento académico que los estudiantes que persisten en sus estudios; lo que resaltan diversos autores con respecto a esto es que el estudiante que abandonan, lo hacen debido a la baja integración personal con los ambientes intelectual y social de la comunidad institucional; esto por lo general causa deserciones tempranas, dado que las relaciones sociales e intelectuales

4

con la institución suelen darse en etapas tempranas del programa académico (Tinto, 1975).

Lo anteriormente descrito está ligado a los factores internos que llevan a un estudiante a tomar la decisión de desertar de sus estudios, esto se denomina deserción desde el punto de vista individual; para definirlo desde la parte institucional podemos observar los efectos económicos negativos, contemplando el escenario de las instituciones privadas, la deserción es una inestabilidad a en la fuente de ingresos puesto que el pago de los estudiantes es la principal fuente; con respecto a las IES públicas, la deserción constituye presupuestos insuficientes que pueden entorpecer las actividades misionales de la misma (Tinto, 1975).

2º. Hipótesis

Hipótesis 1a (H1a). La deserción está directamente relacionada con la existencia de inequidades e inversamente relacionada con los programas y políticas destinados a compensarlas.

Hipótesis 1b (H1b). Las personas con entornos socioeconómicos más vulnerables tienen más probabilidades de abandonar los estudios que las personas con entornos más desfavorecidos.

Hipótesis 1c (H1c). Cuanto mayor sea la capacidad académica del individuo en la escuela secundaria, menor será la probabilidad de abandonar los estudios terciarios.

Hipótesis 1d (H1d). La implementación de programas de ayuda promueve una disminución de las tasas de deserción

Hipótesis 2 (H2). La existencia de desigualdades regionales influye en las tasas de abandono de los estudios terciarios.

Hipótesis 3 (H3). Existen diferencias institucionales por ciclo de programa y campo de estudio que afectan las tasas de deserción.

III. PLANTEAMIENTO DEL PROBLEMA

La educación superior es el eje fundamental para el progreso de las naciones; el nivel educativo de los ciudadanos influye en la economía, la industria, el desarrollo tecnológico, la producción científica e intelectual, avances médicos y demás ejes que contribuyen al progreso integral de los países; las principales potencias mundiales se encuentran dentro de las clasificaciones de mejores sistemas nacionales de educación superior según Datos Mundial (2021) y Villanueva (2019). Por ello los países con mejor sistema educativo logra retener a los estudiantes y aumentar la tasa de profesionales con título universitario.

He ahí donde los gobiernos fijan esfuerzos para intervenir en los principales problemas que afectan su sistema de educación superior; dentro de estos problemas se encuentra la deserción estudiantil; revisando los porcentajes de deserción estudiantil, en la educación superior de Colombia se visualiza que se mantiene en un rango entre 8.19 % y 9.89 % en programas universitarios en el periodo del 2010 al 2018 según el Sistema para la Prevención y Análisis de la Deserción en las Instituciones de Educación Superior (SPADIES) (2020).

SPADIES es un sistema de información especializado para el análisis de la permanencia

en la educación superior colombiana a partir del seguimiento a la deserción estudiantil, que consolida y clasifica la información para facilitar el acompañamiento a las condiciones que desestiman la continuidad en el sistema educativo (Mineducación, 2002).

La Universidad Industrial de Santander desarrolla desde la Vicerrectoría Académica el programa SEA (Sistema de excelencia académica) que cuenta con distintos programas de apoyo en cuatro momentos:

Momento 1. Antes del ingreso a la educación superior.

En esta etapa SEA se encarga de la divulgación de la oferta académica y de articular la universidad con la educación media.

Momento 2. En la transición a la educación superior.

Se da la caracterización del estudiante, ofrece cursos introductorios de matemáticas y lectura universitaria.

Momento 3. Durante la trayectoria académica.

Acompaña al estudiante en los ámbitos académicos, cognitivos, socioeconómicos, salud y ofrece clubes de lectura.

Momento 4. Transición a la vida laboral.

En este momento SEA ofrece la Prueba Saber UIS, charlas preparatorias Saber Pro y Saber TyT, acompañamiento trabajo de grado y talleres de preparación para la vida laboral.

Después de revisar la literatura y consultar múltiples investigaciones se refleja que los programas ofrecidos por SEA se limitan a estar disponibles y no buscan encontrar la raíz y brindar apoyo oportuno a los estudiantes en riesgo.

Teniendo en cuenta la información anteriormente mencionada, el tema de la deserción en la educación superior es de alto interés para el ámbito investigativo; por ello se realiza la presente investigación, específicamente en la facultad de Ingenierías Fisicomecánicas de la Universidad Industrial de Santander, utilizando datos suministrados por la Dirección de Admisiones y Registro Académico.

IV. ANÁLISIS PRELIMINAR DE LOS DATOS

A continuación, se muestra la gráfica de la condición de los estudiantes.

1º. Análisis por características.

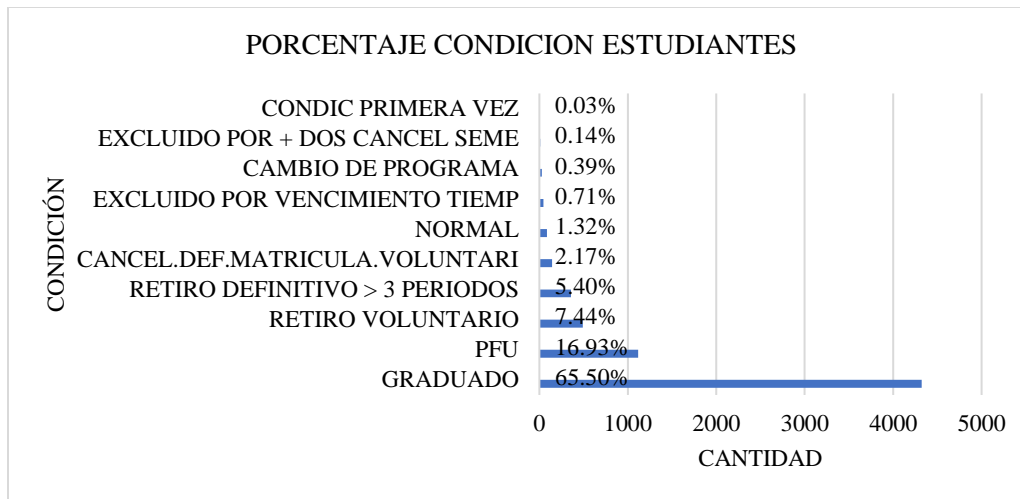


Figura 1

Se observa a partir de la gráfica que aproximadamente el 33.19 % de los estudiantes de la facultad se encuentra en una situación de riesgo a la deserción o ya ha desertado sus estudios de pregrado, este cálculo se halla agrupando los valores de estudiantes en la condición de PFU, retiro voluntario, retiro definitivo > 3 periodos, cancel.def.matricula.voluntaria, excluido por vencimiento tiempo, cambio de programa, excluido por + dos cancel seme y condic primera vez; esta cifra refleja una diferencia con el promedio de deserción de educación superior

universitaria en Colombia que es de 8.684 % entre 2015 y 2019 (SPADIES, 2020).

Admisiones especiales.

Según el Acuerdo 282 de 2017 del 7 de noviembre, la Universidad Industrial de Santander cuenta con el programa de ingreso a la Universidad de aspirantes por la modalidad de admisiones especiales. El número de estudiantes que ingresaron durante el periodo 2015 – por modalidad de ingreso especial fue de 81 distribuidos de la siguiente manera:

Tipo de ingreso	Cantidad
COMUNIDADES AFROCOLOMBIANAS	10
COMUNIDADES INDIGENAS	14
VICTIMAS DEL CONFLICTO ARMADO	31
ESTUDIANTES PROVENIENTES ZONAS DE DIFICIL ACCESO O PROBLEMAS ORDEN PUBLICO	26
Total	81

Tabla 1.

En la tabla 4 se muestra el estado de estos 81 estudiantes.

Condición	Cantidad	Porcentaje	Porcentaje de deserción
PFU	30	37.04%	93.83%
RETIRO VOLUNTARIO	27	33.33%	
RETIRO DEFINITIVO > 3 PERIODOS	16	19.75%	



GRADUADO	5	6.17%
CANCEL.DEF.MATRICULA.VOLUNTARI	3	3.70%
Total	81	100.00%

Tabla 2.

Como se puede observar el 93.83 % de los estudiantes que ingresan a la universidad por modalidad especial se encuentran en deserción estudiantil. Este fenómeno particularmente es preocupante puesto que el ingreso especial es un objetivo institucional, apegándose al inciso segundo del artículo 13 de la constitución política

de Colombia establece que “El estado promoverá las condiciones para que la igualdad sea real y efectiva y adoptará medidas en favor de grupos discriminados y marginados” (Consejo académico UIS, 2017).

Estudiantes PFU.

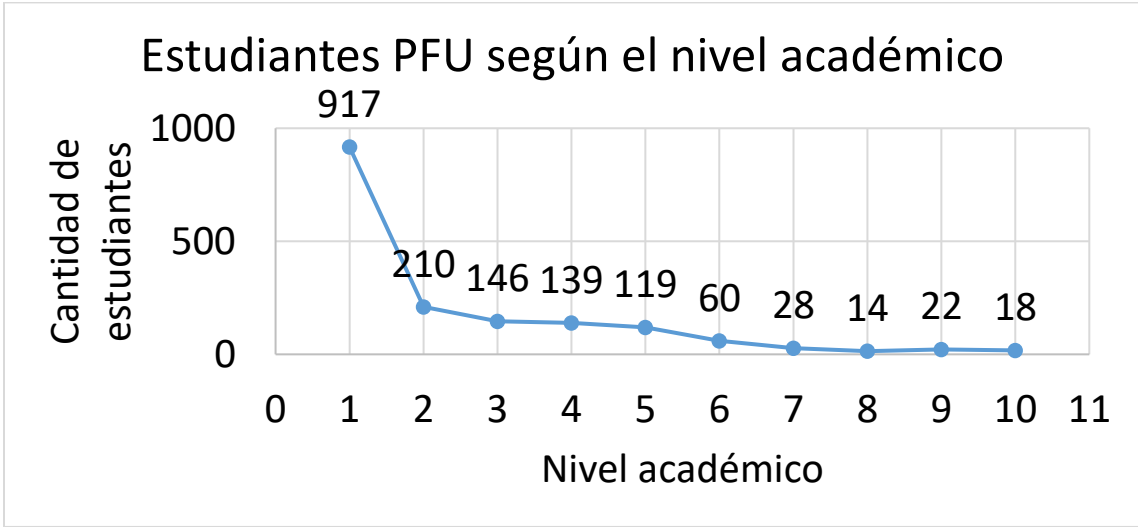


Figura 2.

De la figura anterior se obtiene que el 84.4 % de los estudiantes que quedan por fuera de la universidad (PFU) se encuentran dentro de los primeros cuatro semestres de su programa académico que para ingenierías sería el ciclo básico de la carrera.

A continuación, se muestran las tablas de contingencia de las variables condición, tipo de institución de grado de bachillerato (colegio) y tiempo ocioso en años que hay desde que se graduó un estudiante de bachillerato e ingresa a la universidad.

2°. Análisis multivariable.

Tablas de contingencia.

CONDICIÓN	PRIVADO	PÚBLICO
CAMBIO DE PROGRAMA	0.38%	0.40%
CANCEL.DEF.MATRICULA.VOLUNTARI	2.37%	2.09%
CONDIC PRIMERA VEZ	0.00%	0.02%



EXCLUIDO POR + DOS CANCEL SEME	0.11%	0.17%
EXCLUIDO POR VENCIMIENTO TIEMP	0.54%	0.78%
GRADUADO	62.57%	66.62%
NORMAL	0.70%	1.56%
PFU	15.89%	17.33%
RETIRO DEFINITIVO > 3 PERIODOS	6.62%	4.91%
RETIRO VOLUNTARIO	10.82%	6.11%

Tabla 3.

De la tabla de contingencia condición-colegio se deduce que de la cantidad total de estudiantes que ingresan a la universidad de colegios públicos hay 4.05 % más graduados de los que ingresan de colegios privados; con respecto del retiro voluntario, los estudiantes que son egresados de colegios privados tienen 4.71 % más deserciones por esta condición que los que ingresan de colegios públicos, puede estar sujeto a la facilidad que tienen los estudiantes de migrar a universidades privadas.

Al obtener el coeficiente V de Crammer para la tabla de contingencia de la tabla 5 se obtiene:

```

colegio
0      0.182477
1      0.114179
dtype: float64

```

Los resultados de los coeficientes de la condición-colegio indican que no existe una asociación entre la condición que tenga un estudiante y el tipo de colegio del que se haya graduado de bachillerato.

Siguiendo con el análisis multivariable, se presenta la tabla de contingencia de la condición de un estudiante y el tiempo ocioso (tiempo en años desde que se graduó del colegio e ingreso a la UIS).

CONDICION	0	1	2	3	4	5	6	7	8	9	10
CAMBIO DE PROGRAMA	0.00%	0.41%	0.39%	0.24%	0.78%	0.00%	0.99%	0.00%	0.00%	0.00%	0.00%
CANCEL.DEF MATRICULA VOLUNTARI	2.90%	2.56%	1.55%	1.67%	0.78%	1.28%	0.99%	1.85%	0.00%	0.00%	8.33%
CONDIC PRIMERA VEZ	0.00%	0.02%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
EXCLUIDO POR + DOS CANCEL SEME	1.45%	0.02%	0.23%	0.24%	0.39%	0.00%	0.99%	0.00%	0.00%	0.00%	8.33%
EXCLUIDO POR VENCI	0.00%	0.53%	0.93%	1.44%	0.39%	1.28%	1.98%	0.00%	2.63%	0.00%	0.00%

MIENTO TIEMP											
GRADUADO	55.07 %	64.82 %	71.27 %	64.35 %	63.81 %	69.23 %	52.48 %	59.26 %	63.16 %	44.44 %	8.33%
NORMAL	1.45%	0.94%	1.79%	1.67%	1.17%	3.85%	3.96%	5.56%	0.00%	0.00%	0.00%
PFU	20.29 %	17.71 %	14.13 %	18.42 %	17.51 %	12.82 %	21.78 %	9.26%	13.16 %	22.22 %	8.33%
RETIRO DEFINITIVO > 3 PERIODOS	7.25%	5.60%	4.58%	5.02%	3.11%	4.49%	5.94%	12.96 %	5.26%	7.41%	25.00 %
RETIRO VOLUNTARIO	11.59 %	7.38%	5.12%	6.94%	12.06 %	7.05%	10.89 %	11.11 %	15.79 %	25.93 %	41.67 %

Tabla 4.

La tabla 6 se puede ver completa en el Apéndice J; se observa que el 55.07 % de los estudiantes que se gradúan e ingresan a la UIS el siguiente periodo académico después de culminar su bachillerato logran graduarse de programa académico; el mejor porcentaje de graduados según el tiempo ocioso es el 71.27 % de estudiantes que ingresaron a la UIS 2 años después de haberse graduado del colegio; de todas las columnas de tiempo ocioso desde 0 años de tiempo ocioso a 8 años de tiempo ocioso los porcentajes de graduados varían entre un 0 % a 10 % exceptuando el tiempo ocioso de 2 años; para el año 9 se observa que el porcentaje de graduados es menor al 50 % de estudiantes y para el tiempo ocioso de 10 años el porcentaje de graduados no logra alcanzar sino un 8.33 %, siendo el tiempo ocioso de 10 años el que tiene peores resultados de deserción, el 91.66 % de los estudiantes que tienen un tiempo ocioso de 10 años deserta los estudios por distintas circunstancias pero la principal es el retiro voluntario con un 41.67 % seguido por el 25 % que no matriculan 3 periodos consecutivos.

Coeficientes V de Cramer para tabla de contingencia Condición-Tiempo ocioso:

```

ocio
0      1.041230
1      0.134357
2      0.240998
3      0.423042
4      0.539516
5      0.692483
6      0.860618
7      1.176994
8      1.403070
9      1.664521
10     2.496782
11     2.398830
12     4.993564
13     3.867998
14     4.993564
15     4.324553
16     4.993564
17     8.649106
19     8.649106
20     6.115842
23     8.649106
dtype: float64

```

Según la interpretación de los coeficientes se tiene que de 2 años a 4 años de tiempo ocioso que tiene un estudiante hay una asociación moderada entre la condición de un estudiante y esta cantidad de años de tiempo ocioso, cuando un estudiante tiene un tiempo ocioso de 5 o 6 años

si hay una asociación fuerte entre la condición y el tiempo ocioso.

Conglomerado 1: Naranja.

Conglomerado 2: Azul.

Teniendo en cuenta los dos clústeres y las gráficas anteriores podemos deducir que la

variable edad de grado y tiempo no agrupan a los individuos en grupos significativos, lo que se puede observar en las coordenadas de los centroides de cada clúster, las variables significativas que pueden dar grupos de individuos con características similares son el promedio y el nivel.

Análisis de Clústeres.

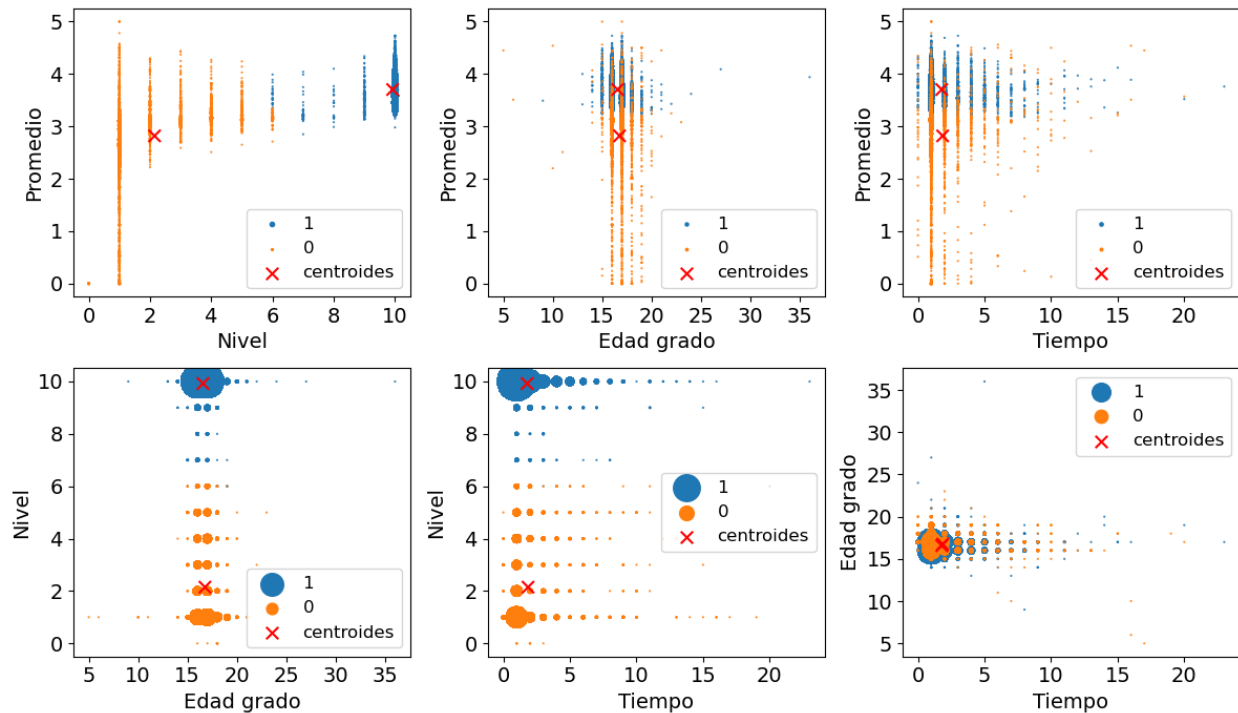


Figura 3.

Coordenadas de los centroides:

Centroide	Promedio	Nivel	Edad grado	Tiempo
1	2.818616	2.147557	16.759071	1.854378
2	3.702938	9.921024	16.561438	1.785352

Tabla 5.

En el conglomerado 1 la coordenada del centroide en la variable promedio es 2.82 y en la variable nivel es 2.15; esto se debe al bajo rendimiento académico reportado en los semestres iniciales, donde la mayoría de los

estudiantes desertores quedan PFU por su promedio ponderado, a partir del nivel 6 es poco probable que un estudiante quede PFU por su rendimiento académico, esto marca el conglomerado 2 donde en centroide de la

variable promedio es 3.70 y el centroide de la variable nivel es 9.9; los individuos del conglomerado 2 se encuentran es una normalidad académica marcada por su promedio y nivel dentro de la institución.

V. MODELADO

Los modelos se eligieron según la naturaleza de los datos y por la precisión mostrada en otras investigaciones:

- ✓ Árboles de decisión.
- ✓ Bosques aleatorios o Random Forest.
- ✓ Regresión Logística.

A continuación, se muestran las variables utilizadas para el modelado después de la transformación con One Hot Encoding.

Variables para utilizar en el modelado.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 5280 entries, 3158 to 2732
Data columns (total 19 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   prome                                5280 non-null   float64
1   nivel                                5280 non-null   int64
2   forma_ingreso                        5280 non-null   int64
3   institucion                          5280 non-null   int64
4   Edad bachillerato                   5280 non-null   int64
5   Tiempo_sabatico                     5280 non-null   int64
6   Tiempo_activo                       5280 non-null   int64
7   Edad de reporte                     5280 non-null   int64
8   programa_Civil                      5280 non-null   uint8
9   programa_Diseño                     5280 non-null   uint8
10  programa_Electrica                  5280 non-null   uint8
11  programa_Electronica                5280 non-null   uint8
12  programa_Industrial                 5280 non-null   uint8
13  programa_Mecanica                   5280 non-null   uint8
14  programa_Sistemas                   5280 non-null   uint8
15  municipio_Area                      5280 non-null   uint8
16  municipio_Desconocidos               5280 non-null   uint8
17  municipio_MunLejanos                 5280 non-null   uint8
18  municipio_MunSant                    5280 non-null   uint8
dtypes: float64(1), int64(7), uint8(11)
memory usage: 428.0 KB
```

1º. Modelo de Árbol de Decisión

Hiperparámetros del modelo árbol de decisión:

La cantidad máxima de características (max_features): son las variables independientes tomadas por el modelo árbol de decisión, que en esta investigación serán las 10 columnas de los datos transformadas en 19.

La profundidad máxima (max_depth): es el total de los niveles del árbol; este hiperparámetro se dejó abierto a que el algoritmo plasmara el máximo de niveles posibles y eligió 14 niveles de profundidad.

En el modelo se utilizaron el 80 % de los datos (5280) como entrenamiento y el 20 % restante (1320) como prueba de la precisión del árbol de

decisión. A continuación, se comparte el árbol de decisión obtenido.

En el primer nivel del árbol de decisión (adjunto en los apéndices) el modelo toma como decisión el nivel del estudiante que es una variable categórica, si el nivel del estudiante es mayor o igual a $7.5 \approx 8$ el estudiante estará en el estado normal, entre los datos de entrenamiento (80 %) el árbol de decisión clasificó 3527 estudiantes en estado normal y 1753 en riesgo; la segunda característica del árbol para definir la situación del estudiante es el promedio, donde si un estudiante tiene un promedio menor o igual a $3.185 \approx 3.2$ el estudiante se encontrará en riesgo.

La tabla de Feature Importance contiene la información de la importancia de las variables independientes sobre el modelo predictivo.

#	feat	importance
1	nivel	0.958210825
0	prome	0.015733969
6	Tiempo activo	0.005214758

7	Edad de reporte	0.004319351
4	Edad bachillerato	0.00282612
5	Tiempo_sabatico	0.002721737
18	municipio_MunSant	0.002159486
3	institucion	0.001853828
8	programa_Civil	0.001302495
14	programa_Sistemas	0.00100334
12	programa_Industrial	0.000916679
17	municipio_MunLejanos	0.00075077
13	programa_Mecanica	0.000747881
9	programa_Diseño	0.000732308
15	municipio_Area	0.000489814
2	forma_ingreso	0.000426989
11	programa_Electronica	0.000426989
10	programa_Electrica	0.000162662
16	municipio_Desconocidos	0

Tabla 6.

De la tabla 8 se deduce que la única variable que tiene influencia en la condición de un estudiante es la variable categórica del nivel; esto concuerda con los hallazgos realizados en el análisis multivariable de los clústeres.

Matriz de confusión del modelo árbol de decisión.

- Matriz de confusión

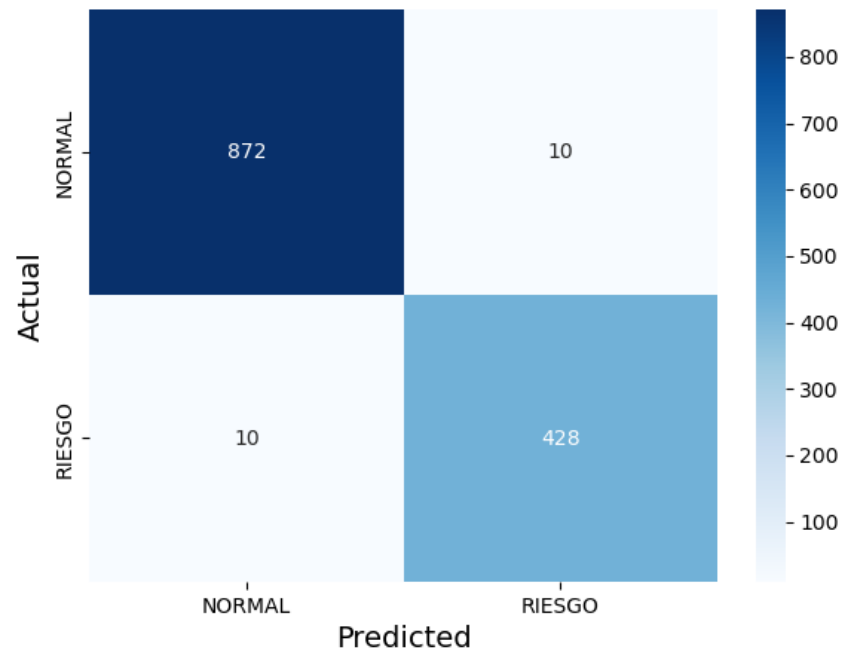


Figura 4.

Según la matriz de confusión, el modelo de árbol de decisión categorizó correctamente 872 estudiantes de los 882 que eran estudiantes con normalidad académica y 428 de los 438 que estaban en riesgo según sus características. Según lo obtenido anteriormente se halla que la precisión global del árbol de decisión es 0.986363 (98.63 %).

2º. Modelo de Bosques Aleatorios

Hiperparámetros del modelo Random Forest:

Para el modelado con bosques aleatorios o Random Forest también se utilizaron el 80 % de los datos para entrenamiento y el 20 % para prueba del modelo.

Números de estimadores (n_estimators): 250, que son el número de árboles en el bosque aleatorio.

Máximas características (max_features): son las variables independientes tomadas por el modelo árbol de decisión, que en esta investigación serán las 10 columnas de los datos transformadas en 19 columnas.

Mínimo de muestras necesarias (min_samples_leaf): es el número de ejemplo mínimo para que un nodo sea considerado hoja, para este caso se eligió 16.

A continuación, se comparte la tabla de Feature Importance del modelo de bosques aleatorios.

#	feat	importance
1	nivel	0.420042177
6	Tiempo activo	0.245420473
0	prome	0.187216902
7	Edad de reporte	0.12529818
12	programa_Industrial	0.00500875
15	municipio_Area	0.002838556
14	programa_Sistemas	0.002668224

5	Tiempo_sabatico	0.002577936
4	Edad bachillerato	0.002068453
2	forma_ingreso	0.001833049
11	programa_Electronica	0.001797051
8	programa_Civil	0.001245939
18	municipio_MunSant	0.000984129
3	institucion	0.000261835
9	programa_Diseño	0.000220717
10	programa_Electrica	0.000205826
13	programa_Mecanica	0.000193911
17	municipio_MunLejanos	0.000083677
16	municipio_Desconocidos	0.000034216

Tabla 7.

Observando la importancia de las variables dentro del modelo de Random Forest y al hacer la comparación con el modelo de árbol aleatorio se deduce que, en el primer modelo toma sólo una variable (nivel del estudiante, importancia de 95.82%) para clasificar los estudiantes en riesgo o en normalidad académica, mientras que el

modelo de Random Forest tiene cuatro variables que tienen un importancia significativa dentro de la clasificación de los estudiantes (Nivel del estudiante 42 %, Tiempo activo en la universidad 24.54 %, promedio ponderado acumulado 18.72 % y edad de reporte 12.53 %).

Según la matriz de confusión, el modelo de Random Forest categorizo correctamente 879 estudiantes de los 882 que eran estudiantes con normalidad académica y 428 de los 438 que estaban en riesgo según sus características. Según lo obtenido anteriormente se halla que la precisión global del árbol de decisión es 0.9902 (99.02 %). A comparación del primer modelo de árbol aleatorio tuvo una mejora de 0.4 % aproximadamente en la clasificación correcta de los estudiantes en los dos grupos respuesta.

Matriz de confusión bosques aleatorios.

- Matriz de confusión

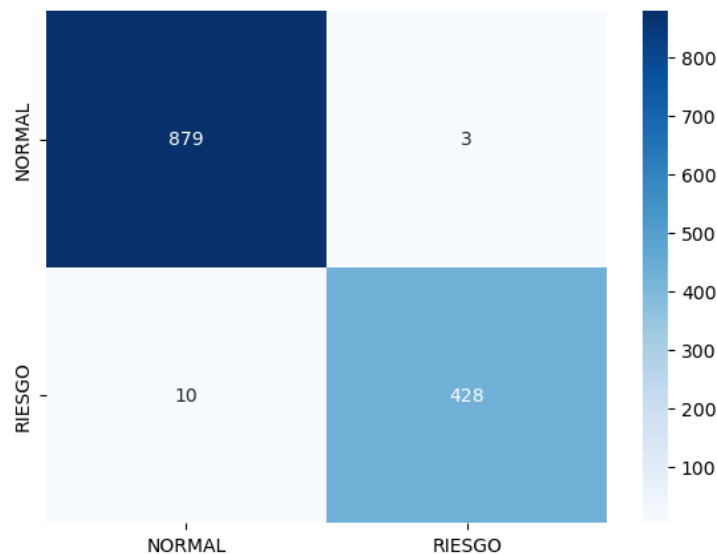


Figura 5.

3º. Modelo de Regresión Logística

Hiperparámetros del modelo de Regresión Logística:

Máximas iteraciones (max_iter): Número máximo de iteraciones necesarias para que los solucionadores converjan, que en el caso específico de la investigación es de 5000 iteraciones.

Matriz de confusión Regresión Logística.

- Matriz de confusión

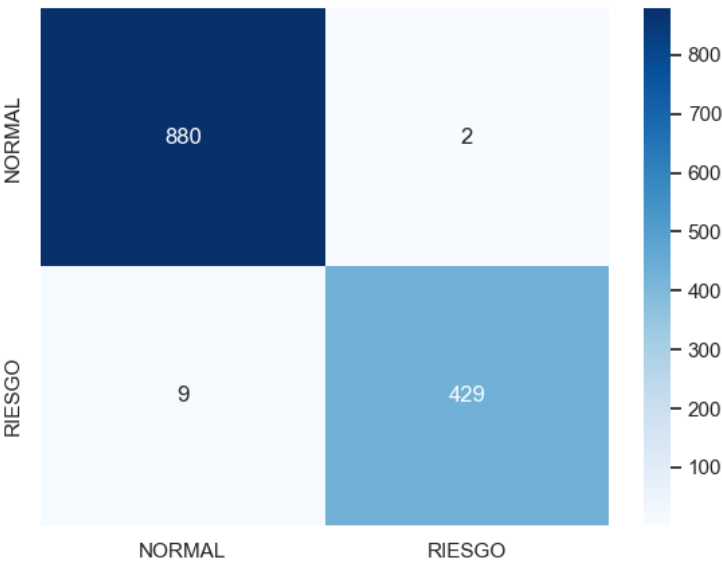


Figura 6.

En la matriz de confusión de la Regresión Logística clasifíco correctamente 880 estudiantes de 882 como estudiantes con normalidad académica y 429 de 438 en riesgo de deserción. Comparado con los dos modelos anteriores, el modelo de Regresión Logística tiene una precisión global de 0.99166 99 (99.17 %); contra

el primer modelo presenta un aumento en la precisión de 0.54 % y contra el segundo modelo, presento un aumento en la precisión global de 0.15 %. Lo que infiere que el modelo de Regresión Logística presenta mejor rendimiento en su clasificación correcta de los estudiantes.

Coefficientes de variables Regresión Logística.

#	Feat	p-values	Coef	Odds
1	nivel	0.00E+00	-8.09449	0.00031
0	prome	1.40E-70	-2.09052	0.12362
4	Edad bachillerato	5.78E-02	-0.23579	0.78995
5	Tiempo_sabatico	1.64E-03	-0.0726	0.92997

14	municipio_Desconocidos	1.49E-02	-0.04272	0.95818
15	municipio_MunLejanos	1.09E-02	-0.02933	0.9711
3	institucion	4.33E-02	-0.02319	0.97708
10	programa_Electronica	1.72E-26	-0.00355	0.99646
16	municipio_MunSant	4.75E-23	0	1
2	forma_ingreso	4.78E-31	0	1
9	programa_Electrica	5.93E-04	0.0189	1.01908
12	programa_Mecanica	3.64E-01	0.0309	1.03138
11	programa_Industrial	8.18E-38	0.06646	1.06872
7	Edad de reporte	1.63E-263	0.09988	1.10504
8	programa_Diseño	5.18E-10	0.25771	1.29396
13	programa_Sistemas	1.48E-35	0.33251	1.39446
6	Tiempo activo	0.00E+00	0.51057	1.66624

Tabla 8.

De la tabla de coeficientes se deduce que las variables más significativas para el modelo predictivo y que influyen mayormente en la variable respuesta son dos variables de programa académico (programa de Diseño Industrial y Programa de Ingeniería de Sistemas) y una variable del tiempo activo que lleva un estudiante en la universidad.

Después de la normalización de las variables, el nivel de referencia del programa académico que utiliza el modelo es Ingeniería Civil, el nivel de referencia se ajustan los valores medidos en escalas comparadas a una escala en común que para esta investigación es la carrera de Ingeniería Civil contrastada con los demás programas académicos; se interpreta que un estudiante que pertenezca al programa de Diseño Industrial tiene 29.4 % más de probabilidad de estar en riesgo de deserción que un estudiante de Ingeniería Civil. Igualmente, un estudiante de ingeniería de Sistemas tiene un 39.4 % más de probabilidad de estar en riesgo que un estudiante de Ingeniería Civil.

Por último, la variable de tiempo activo del estudiante indica que los años activos máximos

de un pregrado asociado a la Facultad de Ingenierías Fisicomecánicas son 5, sin embargo, hay estudiantes que llegan a tener un tiempo activo mayor a 20 años. Según la tabla de coeficientes del modelo después del quinto año activo de un estudiante, por cada año más que un estudiante está activo disminuye la probabilidad de estar en riesgo de deserción un 66 %. Al contrastar la información obtenida en la tabla de coeficientes del modelo con lo encontrado en el análisis multivariable hecho antes del modelado, se encuentra coherencia, puesto que los primeros semestres (primero y segundo) son los que cuentan con mayor tasa de estudiantes en riesgo.

VI. OPTIMIZACIÓN Y VALIDACIÓN DE LOS MODELOS

Para la evaluación de los modelos se utilizó validación cruzada (Cross Validation). La validación cruzada busca reducir el sobreajuste (overfit) de los modelos encontrando los mejores hiperparámetros donde los modelos muestran los mejores resultados. La personalización de hiperparámetros se hace utilizando una grilla de

hiperparámetros que usan una misma métrica para cada modelo en particular.

En la validación cruzada de esta investigación se tomó un N (número de grupos de prueba) que en este caso para los tres modelos fue de cinco grupos.

Árbol de decisión:

Para el Cross Validation del modelo de árbol de decisión se variaron sus hiperparámetros para encontrar los que más optimizan la clasificación de los estudiantes en el modelo.

En el hiperparámetro de max_depth [8, 9, 10, 11, 12, 13, 14], cuando el árbol de decisión tiene 11 niveles de profundidad es cuando se optimiza la clasificación en el Cross Validation.

Comparando la matriz de confusión de la Cross Validation del árbol de decisión con la matriz de confusión del árbol de decisión tenemos, que disminuyó su eficiencia de clasificar los estudiantes en un 0.23 %.

Matriz de confusión Cross Validation del modelo árbol de decisión.

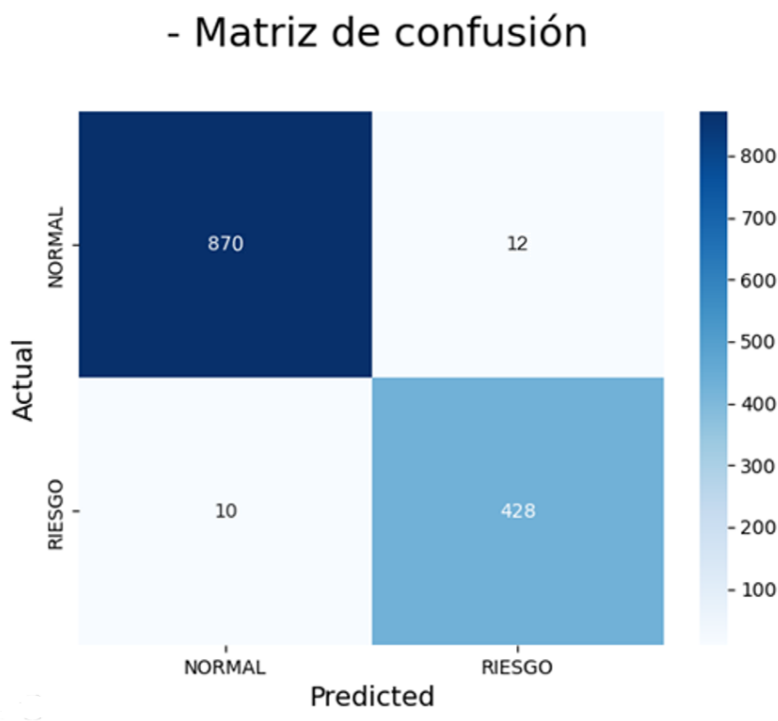


Figura 7.

Random Forest:

Para el Cross Validation del modelo Random Forest se probaron los siguientes hiperparámetros:

N_estimators: [50, 100, 200, 250, 500, 1000].

Min_samples_leaf: [2, 3, 4, 5, 7, 8, 16, 32, 64, 124].

De los hiperparámetros seleccionados, tenemos que los hiperparámetros que optimizan la clasificación de los estudiantes son n_estimators = 100 y min_samples_leaf = 7. De estos dos

hiperparámetros surge la siguiente matriz de confusión.

Comparando la matriz resultante de la validación cruzada del modelo Random Forest con la matriz

del modelado sin optimizar los hiperparámetros, se tiene que disminuyo su eficiencia en la clasificación de los estudiantes en un 0.12 %.

Matriz de confusión Cross Validation del modelo Random Forest.

- Matriz de confusión

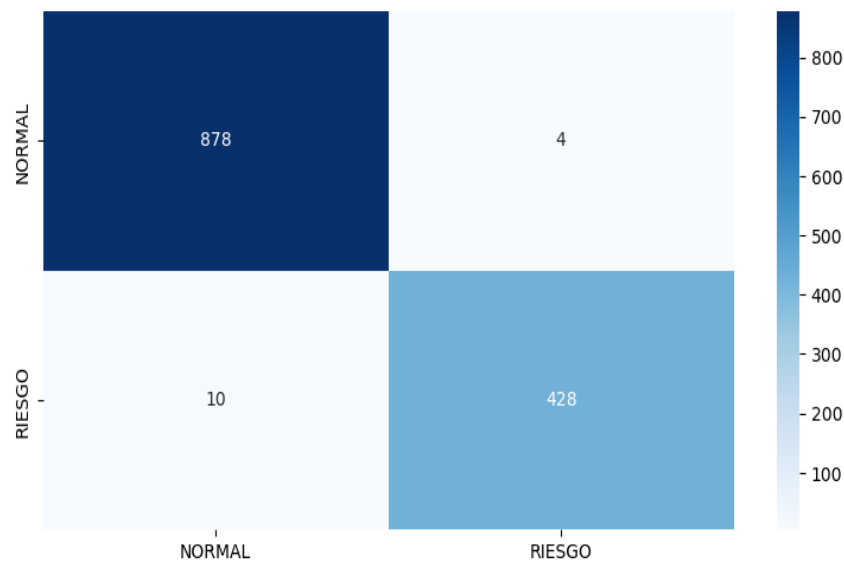


Figura 8.

El solucionador en el Cross Validation del modelo de Regresión Logística es Saga porque es el único que permite hacer Elastic-Net.

Overfitting o sobreajuste es un comportamiento de aprendizaje automático no deseado que se produce cuando el modelo de aprendizaje automático proporciona predicciones precisas para los datos de entrenamiento, pero no para los datos nuevos (Horrillo I. y Barrena M. 2008).

En el caso de la Cross Validation del modelo, este uso las propiedades de la regularización de Lasso (L1).

Al comparar la matriz de confusión de la Cross Validation de la Regresión Logística con la matriz de confusión de la Regresión Logística se tiene que, la matriz obtuvo el mismo porcentaje de precisión al clasificar los estudiantes.

Matriz de confusión Cross Validation de Regresión Logística.

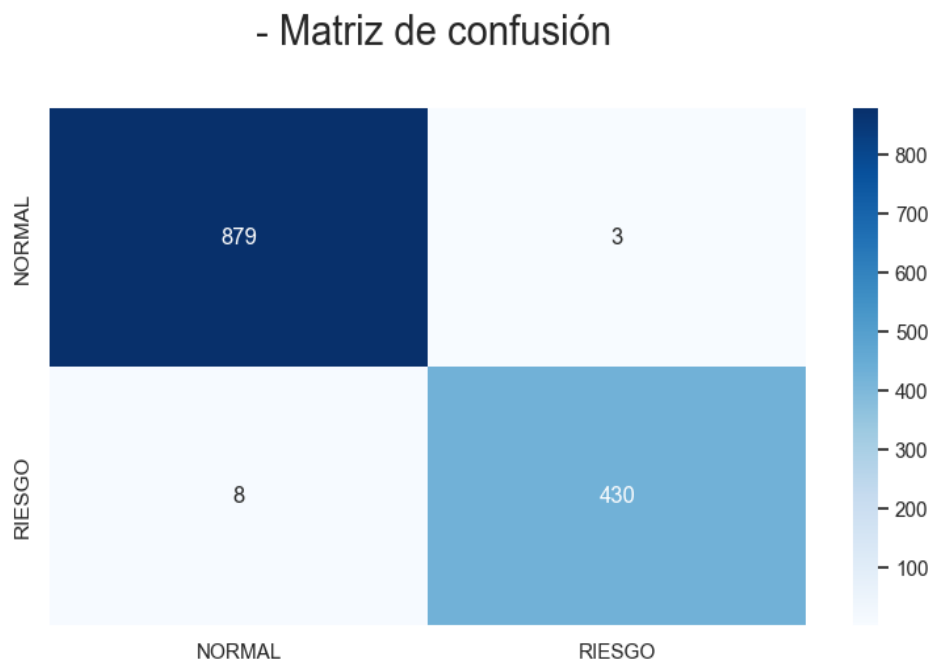


Figura 9.

VII. ELECCIÓN DEL MODELO PREDICTIVO

Lo que hace el Cross Validation es tomar N grupos de datos que fueron 5, toma el primer grupo (20 %) de datos y utilizarlos para entrenamiento y el restante de datos (80 %) los utiliza para probar el modelo, por eso en los resultados obtenidos en los modelos de árbol de decisión y de Random Forest se evidencia una disminución en la efectividad del modelo al clasificar los estudiantes. Esto puede señalar problemas de overfit de los dos primeros modelos sin aplicar la Cross Validation, después de aplicarla se eliminan estos problemas y generalizan mejor los modelos.

El modelo de Regresión Logística fue el que mantuvo su nivel de precisión al hacer la comprobación con Cross Validation, por ello es el modelo que mostró mejores resultados fue el de Regresión Logística.

VIII. CONCLUSIONES

El interés de encontrar soluciones que contribuyan a la disminución de la deserción estudiantil es algo que comparten todas las naciones alrededor del mundo, con el constante avance de la tecnología en la predicción de estudiantes en riesgo de deserción y los estudios compartidos por países de Latinoamérica, hace que haya un precedente histórico de información.

Aunque la Universidad Industrial de Santander cuenta con programas y áreas enfocadas en disminuir la deserción estudiantil, se evidencia según esta investigación que la tasa de deserción es alta viendo los datos de la Facultad de Ingeniería Fisicomecánicas.

Según los datos presentados en la investigación se observa que la deserción está directamente relacionada con la inequidad y afecta a personas

que viven en entornos socioeconómicos vulnerables; aunque la universidad tenga programas de acceso a la educación para personas que pertenecen a estos entornos o grupos vulnerados, no logra retener estos estudiantes lo que conlleva a que realmente el programa no cumpla el objetivo de que la igualdad sea real y efectiva (Acuerdo No. 282 de 2017, Universidad Industrial de Santander).

Teniendo en cuenta lo hallado con respecto a los programas académicos adjuntos a la Facultad de Ingeniería Fisicomecánicas, los programas de ingeniería electrónica, ingeniería eléctrica, ingeniería mecánica e ingeniería de sistemas presentan altas tasas de deserción y de cancelaciones de matrícula. Lo que concuerda con la hipótesis 3 (H3) presentada en el plan.

Así mismo las características que evidencian relación con el riesgo de un estudiante a desertar sus estudios dentro de los modelos analizados son principalmente el promedio, el nivel y el tiempo activo del estudiante; si el promedio ponderado de un estudiante está entre 2.7 y 3.2 lo pone en riesgo de condicionalidad, si es menor a 2.7 inmediatamente queda PFU (por fuera de la universidad) y se convierte en desertor; con respecto al nivel, dentro del análisis de variables se encontró que el 84.4 % de los estudiantes que quedan PFU se encuentran dentro de los primeros cuatro niveles del programa académico que para los programas de ingenierías estaría comprendido por el ciclo básico, además que entre mayor sea el nivel alcanzado por un estudiante menos es la probabilidad de que este en riesgo de desertar sus estudios, esto se relaciona con el tiempo activo del estudiante.

Finalmente se concluye que el modelo que mejor predice la probabilidad que tiene un estudiante de estar riesgo de desertar sus estudios luego de la

optimización por medio de Cross Validation es el modelo de Regresión Logística dado que refleja mejores resultados a la hora del tuneo de sus hiperparámetros y presenta menos sobre ajuste manteniendo la efectividad como predictor.

IX. RECOMENDACIONES

Para futuras investigaciones, existen oportunidades de extender la presente investigación, utilizando datos de todos los estudiantes de pregrado de la Universidad Industrial de Santander, solicitando características adicionales que puedan estar relacionadas al riesgo de desertar los estudios.

Incentivar a los estudiantes a investigar sobre la deserción estudiantil, que la universidad y la Dirección de Admisiones y Registro Académico conceda acceso a la información no sensible de los estudiantes para un estudio más profundo sobre la relación de las características de un estudiante con la posible deserción de sus estudios.

Considerar la posibilidad de tener un sistema de acceso abierto a los datos de los estudiantes de pregrado para que en futuras investigaciones se puedan utilizar variables adicionales como el género del estudiante, nivel económico de los padres, nivel educativo máximo alcanzado por los padres, promedio del último grado del colegio, entre otras que pudieran tener una relación con la deserción.

X. BIBLIOGRAFÍA

Acuerdo No. 282 de 2017. Por el cual se dictan disposiciones sobre el ingreso a la Universidad de aspirantes por la modalidad de Admisiones Especiales. 7 de noviembre de 2017. Consejo académico de la Universidad Industrial de Santander.

- Alban, M.; Mauricio, D. (2019). Neural Networks to Predict Dropout at the Universities. *International Journal of Machine Learning and Computing*. 9 (2), 149–153.
<http://www.ijmlc.org/index.php?m=content&c=index&a=show&catid=84&id=905>
- Amat, J. (2020). *Regularización Ridge, Lasso y Elastic Net con Python*. Ciencia de datos.
<https://cienciadedatos.net/documentos/py14-ridge-lasso-elastic-net-python#:~:text=La%20regularizaci%C3%B3n%20Ridge%20penaliza%20la,que%20estos%20lleguen%20a%20cero.>
- Barbosa-Camargo, MI; García-Sánchez, A.; Ridao-Carlí, ML. (2021). *Desigualdad y Deserción en la Educación Superior en Colombia. Un análisis multinivel de las diferencias regionales, las instituciones y el campo de estudio*.
<https://doi.org/10.3390/math9243280>
- Cabrera, E.; Díaz, E. (2021). *Manual de uso de Jupyter Notebook para aplicaciones docentes*. Universidad Complutense de Madrid.
- Cabrera, L.; Bethencourt, J.; Álvarez, P.; González, M. (2006). El problema del abandono de los estudios universitarios. [The dropout problem in university study]. *Revista Electrónica de Investigación y Evaluación Educativa*. (12), 171–203.
<https://www.redalyc.org/pdf/916/91612201.pdf>
- Camacho M., Montalvo A., Galezo P. (2019). Determinantes de la Deserción estudiantil en estudiantes universitarios. *Panorama Económico*. 27 (1), 134-162.
- Eckert, K.; Suenaga, R. (2014). Análisis de Deserción-Permanencia de Estudiantes Universitarios Utilizando Técnica de Clasificación en Minería de Datos. *Formación Universitaria*. 8 (5), 3–12.
- Fischer, E. (2012). *Modelo Para la Automatización del Proceso de Determinación de Riesgo de Deserción en Alumnos Universitarios*. (Tesis de maestría), Universidad de Chile, Santiago, Chile.
- Heredia, D.; Amaya, Y.; Barrientos, E. (2015). Student Dropout Predictive Model Using Data Mining Techniques. *IEEE Lat. Am. Trans.* 13, 3127–3134.
<https://ieeexplore.ieee.org/document/7350068>
- Kumar, S.; Bharadwaj, B.; Pal, S. (2012). Mining Education Data to Predict Student's Retention: A comparative Study). *International Journal of Computer Science and Information Security*. 10, 113–117
- Kumar, B.; Pal, S. (2011). Data Mining: A prediction of performer or underperformer using classification. *International Journal of Computer Science and Information Security*. 2, 686–690.
- Ojeda, V.; Fernández, J. Isea, R. Gutiérrez, A. Salazar, V. (2018). *Coefficiente V de Cramer (V)*. Facultad de humanidades y educación. Universidad Central de Venezuela.
- Pascarella, E. T. & Terenzini, P. (1977). Patterns of Student-Faculty Informal Interaction Beyond the Classroom and Voluntary Freshman Attrition. *The Journal of Higher Education*, 48 (5), 540-562.
- Pascarella, E. T. & Terenzini, P. (1991) *How College Affects Students: Findings and Insights from Twenty Years of Research*; Jossey-Bass Publishers: San Francisco, CA, USA

- Siri, A. (2015) Predicting Students' Dropout at University Using Artificial Neural Networks. *Italian Journal of Sociology of Education*. 7, 225–247.
- Tinto, V. (1975). *Dropout from Higher Education: A Theoretical Synthesis of Recent Research*. 45(1), 89–125.
<https://doi.org/10.3102/00346543045001089>
- Valero, S.; Salvador, A.; García, M. (2005). Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. *Ene*. 779(73), 33.
<https://www.academia.edu/download/34203825/e1.pdf>
- Villanueva, J. A. (2019, 2 abril). *Los mejores sistemas nacionales de educación superior de 2019 – Oficina de Acreditación y Calidad*. Oficina de Acreditación y Calidad, Universidad Nacional de Ingeniería. <https://acreditacion-fiis.com/los-mejores-sistemas-nacionales-de-educacion-superior-de-2019/>
- Zahra, Ghazanfar, M. A., Khalid, A., Azam, M. A., Naeem, U., & Prugel-Bennett, A. (2015). Novel centroid selection approaches for KMeans-clustering based recommender systems. *Information Sciences*, 320, 156–189.
<https://doi.org/10.1016/j.ins.2015.03.062>
- Zhang, G. (2000). Neural networks for classification: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*. 30 (4), 451–462.
<https://doi.org/10.1109/5326.897072>