Prediction of University Students at Academic Risk using Supervised Algorithms

Edson Nicks Lazaro-Camasca. en Ciencias de la Computación[1], Yuri Nuñez-Medrano MSc. en Ingeniería de Sistemas[2]
[1]Universidad Nacional de Ingeniería, Perú, *elazaroc@uni.pe,* [2] Universidad Nacional de Ingeniería, *ynunezm@uni.edu.pe*

Abstract– The purpose of this work was to create predictive models using Supervised Classification Algorithms, in order to make known that students were at academic risk and to be able to carry out a focused follow-up, this work is based on research such as [1], [2] and [3].

In this study, the CRISP-DM methodology was used to create predictive models, taking full advantage of the data obtained by the university itself, where these only contain academic qualifications. Some important findings obtained during the data analysis was the importance of the summer period, thanks to this cycle the number of students at risk decreases significantly.
Furthermore, the majority of at-risk students are focused on the first four semesters. Five classifiers are presented, Bayesian Classifier, Artificial Neural Network, Discriminant Quadratic Analysis, Support Vector Machine and Logistic Regression.
The choice of the best model is based on two Performance Measures, the ROC Curve and Sensitivity, then the two best models are presented according to the resources that the institution has, the Bayesian Classifier when there are enough resources and the Logistic Regression when resources are scarce.

Keywords— Supervised Algorithms, CRISP-DM, Academic Grades, Bayesian Classifier, Logistic Regression, Sensitivity.

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2021.1.1.363
ISBN: 978-958-52071-8-9 ISSN: 2414-6390

Predicción de Estudiantes Universitarios en Riesgo Académico usando Algoritmos Supervisados

Edson Nicks Lazaro-Camasca. en Ciencias de la Computación[1], Yuri Nuñez-Medrano MSc. en Ingeniería de Sistemas[2]
[1]Universidad Nacional de Ingeniería, Perú, *elazaroc@uni.pe,* [2] Universidad Nacional de Ingeniería, *ynunezm@uni.edu.pe*

Resumen– ***El presente trabajo, tuvo como propósito la creación de modelos predictivos usando Algoritmos Supervisados de Clasificación, con el fin de dar a conocer que alumnos llegaran a estar en riesgo académico y poder realizar un seguimiento focalizado, este trabajo está basado en investigaciones como [1], [2] y [3].***
En este estudio se usó la metodología CRISP-DM para la creación de los modelos predictivos, se aprovecho en su totalidad los datos obtenidos por la propia universidad, donde estos solo contienen calificaciones académicas. Algunos hallazgos importantes obtenidos durante el análisis de los datos fue la importancia del periodo de verano, gracias a este ciclo la cantidad de alumnos en riesgo disminuye significativamente. Además, que la mayoría de alumnos en riesgo se encuentra focalizado entre los cuatro primeros semestres. Se presentan cinco clasificadores, Clasificador Bayesiano, Red Neuronal Artificial, Análisis Cuadrático Discriminante, Maquina de Vectores de Soporte y Regresión Logística. La elección del mejor modelo está basada en dos Medidas de Rendimiento, la Curva ROC y la Sensibilidad, entonces se presentan a los dos mejores modelos de acuerdo a los recursos que posee la institución, el Clasificador Bayesiano cuando se tienen suficientes recursos y la Regresión Logística cuando los recursos son escasos.

Palabras claves— Algoritmos Supervisados, CRISP-DM, Calificaciones Académicas, Clasificador Bayesiano, Regresión Logística, Sensibilidad.

I. Introducción

En el Perú durante mucho tiempo los estudiantes universitarios podían culminar la carrera en un largo periodo de tiempo sobrepasando los 5 años, si bien esto no es un problema para las universidades particulares ya que estos obtienen ingresos por parte de los alumnos, si lo es para las universidades estatales ya que estos perciben ingresos por parte del estado y se estaban usando más recursos de los debidos, es por ello que se agregó el artículo 102 en la nueva Ley Universitaria donde señala que la desaprobación de una misma materia por tres veces da lugar a que el estudiante sea separado temporalmente por un año de la universidad. Al término de este plazo, el estudiante solo se podrá matricular en la materia que desaprobó anteriormente, para retornar de manera regular a sus estudios en el periodo siguiente. Si desaprueba por cuarta vez el mismo curso se procede a retirarlo de manera definitiva.

El problema que surge con el artículo 102, es en aquellas universidades donde la exigencia es muy alta como la UNI, sumando con el estrés de los alumnos en no querer desaprobar, genera que se retire a varios alumnos de manera parcial o definitiva, y por ende genera una escasez de profesional de alto nivel.

Si realizamos una aproximación de cuánto tiempo le toma a un estudiante poder ingresar a una universidad pública, sin haber tenido una preparación adecuada en el nivel secundario estaría rondando entre 2 o 3 años, sumado con un año de retiro parcial, el estudiante habrá desperdiciado de 3 a 4 años sin llegar a ser profesional, ese tiempo es equivalente a realizar una carrera técnica.

Este artículo, está desarrollado con el propósito de dar a conocer y prevenir a los estudiantes de la Facultad de Ciencias de la Universidad Nacional de Ingeniería que serán separados de manera parcial o definitiva, para cumplir con esto se desarrollaron los siguientes puntos: 1) Tratamiento de los datos; 2) Ingeniería de Características; 3) Desarrollo de modelos de clasificación; 4) Comparación y Selección de los mejores modelos.

II. OBJETIVOS

El objetivo de este trabajo es dar a conocer la utilidad que tienen los datos académicos para el mejoramiento de la calidad educativa desarrollando modelos de Aprendizaje Automático [4] capaces de clasificar a los estudiantes y predecir quienes se encontrar en Riesgo Académico.

Para el desarrollo de este trabajo se utilizó la plataforma Colab con un kernel de python3. Además, se usó la librería Scikit-learn [5] para el desarrollo de los modelos predictivos.

III. Tratamiento de los datos.

A. Descripción de la Colección de Datos

La colección de datos para los modelos fue proporcionada por el Departamento de Estadística de la Facultad de Ciencias de la Universidad Nacional de Ingeniería, esta consiste de calificaciones académicas de 1439 estudiantes desde que iniciaron su etapa universitaria hasta el semestre 2019-2, se tiene alrededor de 150 cursos en 5 diferentes especialidades.

En la tabla I, se describe a detalle la colección de datos, es importante notar que en esta tabla no se contemplan datos socioeconómicos por ser altamente sensibles y privados,

aunque estos datos proporcionan un mejor rendimiento a los modelos predictivos.

TABLA I
COLECCIÓN DE DATOS

Característica	Descripción
Código del Semestre	Identificador del semestre, consta del año seguido de semestre, se tiene 1 si es de marzo a julio, 2 si es de agosto a diciembre y 3 si es de enero a marzo, periodo de verano.
Especialidad	Escuela a la cual pertenece, esto va depender de la facultad por ejemplo en Ciencias se tiene Matemática, Física, Ingeniería Física, Química, Ciencia de la Computación
Código del curso	Identificador único del curso, consta de 2 caracteres seguido de 3 números
Créditos del curso	Peso de un curso basado en el número de horas.
Nota del curso	Calificación obtenida por el estudiante, esta tiene un rango de 0 a 20.
Condición	Condición de persistencia en el curso que puede ser N si es normal y R si es retirado

B. Extracción de Datos

En diversos proyectos en donde se quiere extraer conocimiento a partir de los datos estos se encuentran en formatos inadecuados [6], este problema se presentó en la investigación en donde la colección de datos estaba en un formato PDF, para poder explorar los datos se realizó la siguiente cadena de conversiones:

1. Se transformo de PDF a un archivo de texto, donde los datos no tenían ninguna estructura.
2. Se aplico expresiones regulares para encontrar patrones en los datos y extraerlos a un formato de tabla.
3. Los datos en formato tabla fueron almacenados como CSV, para ser reutilizados e instanciados como DataFrames.

C. Exploración de los datos

La primera característica a explorar fue el semestre, en específico una derivada de esta que es la cantidad de semestres cursados, ya que para que un estudiante se encuentre en riesgo debe haber cursado como mínimo dos semestres regulares.
A partir de esta característica se han definido 2 tipos.

- NSR: Numero de Semestres Regulares Cursados
- NSV: Numero de Semestres Verano Cursados

En la tabla II se puede ver la diferencia de condiciones que existen entre un periodo regular y verano en diversos factores.

TABLA II
COMPARACIÓN ENTRE PERIODO REGULAR Y VERANO

Factores	Periodo Regular	Periodo Verano
Tiempo	El periodo dura 4 meses	El periodo dura 2meses
Economía	Pago único de matrícula.	Pago por cada curso llevado.
Factor de Riesgo	Al desaprobar se agrega al récord histórico.	Al desaprobar no se agrega al récord histórico

El periodo de verano no genera riego para ser retirado de la universidad, ya que en caso de desaprobar un curso este no se tomará en cuenta en el histórico académico. Sin embargo, algunos cursos suelen ser costosos debido al uso de equipamiento especial en los laboratorios, a pesar de ello la mayoría de los alumnos optan por este semestre, esto se puede ver en la Fig.1, donde 448 alumnos han cursado por lo menos un periodo de verano, esto representa el 31% de la muestra del trabajo.

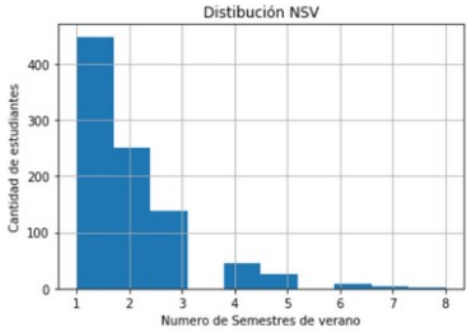


Fig.1 Distribucion de la cantidad de estudiantes por NSV.

Por otro lado, al analizar el NSR en la Fig. 2, se pudo notar que la mayoría tenia entre 1 a 12 semestres cursados de forma regular. Sin embargo, se encontró la existencia de estudiantes que han cursado mas de 20 semestres inclusive llegando a 60 semestres (aproximadamente 30 años). Estos estudiantes fueron considerados como outliers y fueron retirados de la muestra, esto por ser una cantidad escasa de estudiantes y ser casos especiales.

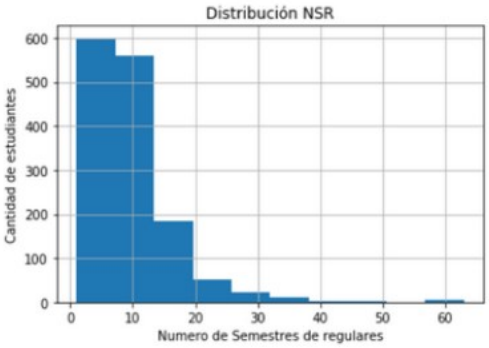
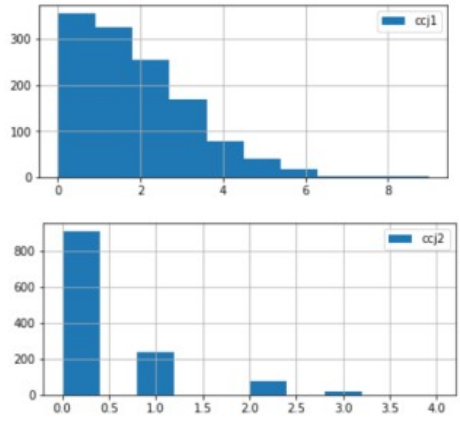


Fig.2 Distribucion de la cantidad de estudiantes por NSR.

Estableciendo Target de los modelos

El target o también llamado etiqueta objetivo va ser la variable de salida de los modelos [7], para ello era necesario transformar los datos y obtener la cantidad de cursos reprobados o "jalados" de cada estudiante, esta transformación se aplicó solo a los estudiantes que tienen como mínimo 3 semestres cursados.

En la Fig.3, se puede ver la cantidad de estudiantes por cursos jalados 1 vez (superior), 2 veces (medio) y 3 veces (inferior), en las gráficas el eje "x" es la cantidad de cursos y el eje "y" la cantidad de estudiantes, es importante destacar que la cantidad de cursos del eje x son todos enteros, en algunas barras estos cubren otros números como la primera barra del segundo gráfico, este solo marca a 0 cursos jalados 2 veces. Se encontró que la cantidad de estudiante en riesgo relativo aquellos que han desaprobado 2 veces un mismo curso son 317 estos son el 22% de la muestra. Por otro lado, la cantidad de estudiantes en riesgo definitivo aquellos que han desaprobado 3 veces un mismo curso son 156 que representa 10% de la muestra.



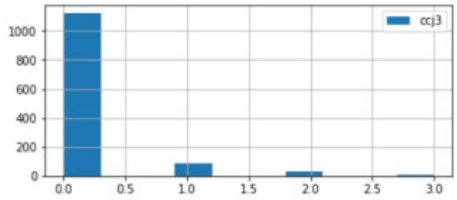


Fig 3. Distribucion de la cantidad de estudiantes por curso jalados 1, 2 y 3 veces

Se estableció el valor 1 al target de los 156 estudiantes en riesgo
definitivo y el valor de 0 para el resto de la muestra.

División de los datos para su procesamiento

Para dividir los datos hay que ver como los datos están estructurados y agrupados, en este trabajo se pudo observar que las calificaciones y demás características se encuentra agrupados por semestres. Para obtener el riesgo se debe conocer el último semestre, entonces surge la pregunta ¿A qué semestre se va predecir?, esta duda surge porque los datos se encuentran en diferentes semestres que son básicamente, los últimos semestres cursados por los estudiantes.

A continuación, se describe la división de datos para cada procedimiento, en cada una de estas se realiza la asignación del target tomando en cuenta los 3 últimos semestres.

- *Data de Validación:* Se estableció al periodo 2019-2 como el ultimo y se asignó el target usando 2 semestres anteriores, luego se eliminó todos los datos del 2019-2 ya que este periodo es el semestre final a predecir.

- *Data de Calibración – Prueba*: Se realizo un procedimiento similar al de validación estableciendo como último semestre al 2019-1.

Luego de dividir los datos por periodos se tenía como último periodo regular al 2018-2 y ultimo periodo de verano al 2018-3, estos datos fueron usados para la fase de Ingeniería de Características.

IV. Ingeniería de Características.

En este parte del trabajo se utilizaron diversas técnicas para crear, combinar, eliminar características con el objetivo de generar un mayor rendimiento en los modelos [8], para analizar a detalle cada característica era necesario conocer su contexto en la universidad y como esto afectaba al rendimiento académico de cada estudiante.

A. Cantidad de Cursos Reprobados

Esta característica ya se usó en la definición del target, pero en esta parte solo se considera 2 semestres, es decir la cantidad de cursos reprobados ("jalados") 1 o 2 veces. La abreviatura de estas este dado por CCJ1 y CCJ2 y la distribución que tienen estas son muy similares a la Fig. 3, pero la cantidad de estas es menor por la división de datos. En el análisis de CCJ2 se encontró 122 alumnos en riesgo parcial, es decir que si desaprueban una vez más serán retirados por un año.

B. Especialidad

El riesgo de desaprobar varia de acuerdo a la especialidad a la que se pertenece, si bien las especialidades que se trabajaron pertenecen a una misma Facultad y se complementan entre ellas, existen factores como la exigencia, la calidad de laboratorio, los profesores, la metodología que varía en cada especialidad.

En la Fig. 4 se puede ver como la cantidad de estudiantes (eje "x") se distribuye en cada especialidad (eje "y"), en naranja se representa los que se encuentran con riesgo y en azul sin riesgo. La cantidad de estudiantes en riesgo no varia mucho entre cada especialidad salvo en la especialidad 5.

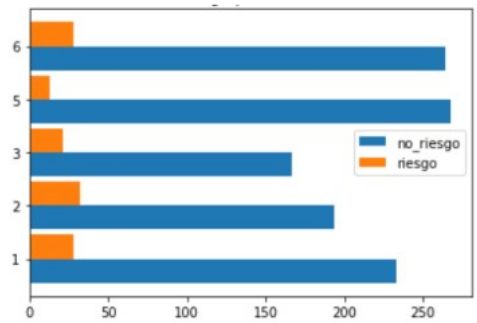


Fig 4. Distribucion de la cantidad de estudiantes por especialidad y riesgo.

En la tabla III se tiene la cantidad exacta de alumnos en riesgo por especialidad, donde Ingeniería Física tan solo tiene 13 alumnos en riesgo esto es un posible indicador de la calidad educativa que se tiene en esta especialidad.

TABLA III
RIESGO ACADÉMICO POR ESPECIALIDAD

Especialidad	Sin Riesgo	Con Riesgo
1_Física	233	28
2_Matemática	194	32
3_Química	167	21
4_Ingeniería Física	268	13
5_Ciencia de la Computacion	264	28

C. Ciclo Relativo

El ciclo relativo es el semestre en el que un estudiante se encuentra de acuerdo al éxito obtenido en sus cursos, para saber en qué ciclo relativo se encuentra un estudiante se define una característica previa llamada NCA, Numero de Créditos Aprobados, cada ciclo relativo se obtiene a partir de un rango de NCA.

Esta característica es muy importante ya que cada ciclo relativo tiene su propio contexto, los primeros ciclos son una fase de adaptación universitaria, donde se enseñan cursos básicos o generales de acuerdo a cada especialidad, a medida que se va avanzando en cada ciclo las prioridades van cambiando, los estudiantes van especializándose en áreas específicas llevando cursos electivos y complementarios, y por los últimos ciclos se tiene cursos más ligeros.

Sin embargo, se tiene que realizar otras actividades como practicas preprofesionales, actividades diversas, etc. Con esta breve descripción se puede ver que el enfoque en cada ciclo varia, saber en qué ciclo relativo existe una alta tasa de riesgo es muy importante, ya que con esto se puede focalizar los cursos, docentes, metodologías que pueden ser mejorados.

Se realiza un análisis de esta característica y se obtiene la distribución de la Fig. 5, en esta se tiene la cantidad de alumnos en riesgo por ciclo relativos (color naranja) y aquellos que no están en riesgo (color azul), se puede notar que a medida que el ciclo va creciendo la tasa de riesgo va decreciendo, donde la mayor cantidad de riesgo se concentra en los 4 primeros ciclos.

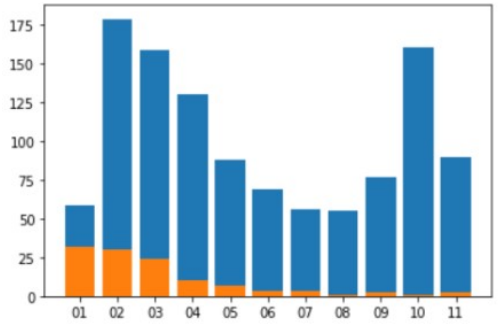


Fig 5. Distribucion de la cantidad de estudiantes por especialidad y riesgo.

D. Créditos

Esta característica ya viene por defecto en el conjunto de datos original especificada para cada curso, la característica del crédito es un valor numérico que se le asigna a un curso de acuerdo a la complejidad que esta tiene, ya sea por la cantidad de horas en que se dicta, el número de prácticas, el número de laboratorios, etc. Para esta parte realizamos una transformación de agrupación aplicada a los créditos, es decir se suma todos los créditos de acuerdo a cada semestre. En este caso se tienen 2 semestres y se define 2 nuevas características:

- NCAS: el Numero de Créditos del Antepenúltimo Semestre.
- NCUS: el Numero de Créditos del Último Semestre.

En la Fig.6, se tiene un gráfico de puntos de todos los estudiantes en el eje “x” se tiene NCAS y en el eje “y” se tiene NCUS, se puede notar que los estudiantes con riesgo se concentran en el centro, también se puede notar que aquellos alumnos que llevan más de 20 créditos en un periodo por lo general no se encuentran en riesgo.

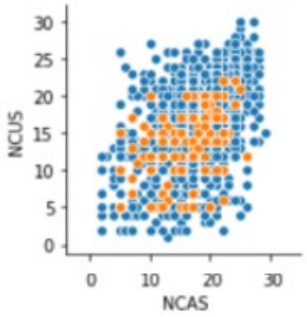


Fig 6. Distribucion de estudiantes por numero de creditos.

E. Score

En esta parte se define la característica más importante de para los modelos, ya que toma como base el rendimiento de los estudiantes y agrupa otras características que no se encuentran de forma explícita.

El score se define como la **sumatorio del promedio de cada curso por su crédito correspondiente.** Si bien esta definición planteada es similar a la del promedio ponderado por semestre, en esta no se divide entre el número total de créditos, al realizar una estandarización se pierde el rendimiento real de un estudiante, por ejemplo, si realizamos una comparación con dos estudiantes, el estudiante “a” tiene un promedio de 14 con 10 créditos y el estudiante “b” tiene un promedio 12 con 24 créditos, si bien el estudiante “a” tiene un promedio mayor que el estudiante “b”, el rendimiento del estudiante “b” fue mejor frente al estudiante “a", ya que se tiene una diferencia de 10 créditos a favor del estudiante “b” y por ende el estudiante “b” avanza con más rápido en la carrera.

El score es hallado para los 2 últimos semestres y se definió de la siguiente manera:

- SAS: Score del Antepenúltimo Semestre.
- SUS: Score del Último Semestre.

En la Fig.7, se tiene un gráfico de puntos de todos los estudiantes en el eje “x” se tiene SAS y en el eje “y” se tiene SUS, se puede notar que los estudiantes con riesgo se concentran de 0 a 200 de score, también se puede notar que aquellos alumnos que tienen un score superior a 200 en SAS y SUS por lo general no están en riesgo.

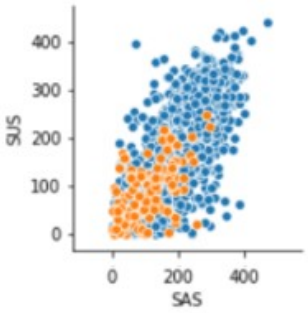


Fig 7. Distribucion de estudiantes por score.

F. Periodo de Verano

Las anteriores características han sido de semestres regulares, en esta parte se realizó un análisis del semestre de verano, en donde un estudiante que se encuentra en posible riesgo tiene la posibilidad de cursar la asignatura en riesgo con la ventaja de que, si desaprueba no va ser suspendido.

En esta parte se definió un score para este ciclo SSV de forma similar que el anterior, para aquellos alumnos que no cursaron un ciclo de verano se les asigna un valor negativo para que los modelos puedan discriminar entre aquello que si lo cursaron y aquellos que no. Sin lugar a dudas el periodo de verano es un ciclo donde los alumnos aprovechan para salir de riesgo o mejorar su rendimiento académico.

En la Tabla IV, se puede ver una comparación entre el periodo de verano y el riesgo, se tiene que 391 estudiantes cursaron en verano y de estos solo el 10.5% se encuentra en riesgo.

TABLA IV
RIESGO ACADÉMICO POR PERIODO DE VERANO

Verano	Sin Riesgo	Con Riesgo
SI	350	41
NO	837	82

G. Selección de Características

Para seleccionar las características es necesario encontrar la importancia que tiene cada una de estas [9], para ello se utilizó el algoritmo XGBoost.

En la Tabla V, se puede ver las 10 características finales y su grado de importancia, se obtuvo una media de importancia de 38, en base a esta media se depuro aquellas características que eran menores a este valor, las características que se eliminaron fueron CCJ1, ESP, NCUS.

TABLA V

CARACTERÍSTICAS FINALES

Abreviatura	Característica	Importancia
CCJ1	Cantidad de Cursos Jalados 1vez	30
CCJ2	Cantidad de Cursos Jalados 2 veces	42
ESP	Especialidad	13
NCA	Numero de Créditos Acumulados	51
CRel	Ciclo Relativo	39
SSV	Score del Semestre de Verano	69
NCAS	Número de Créditos Cursados en el Antepenúltimo Semestre	39
NCUS	Número de Créditos Cursados en el Último Semestre	20
SAS	Score del Antepenúltimo Semestre	44
SUS	Score del Último Semestre	38

V. DESARROLLO DE CLASIFICADORES.

En este parte del trabajo se implementó los modelos de clasificaron, pero antes de ello se realizo un preprocesamiento en las características, se estandarizo usando la librería StandarScaler de Sklearn para el mejor rendimiento de los modelos.

A. Partición de datos Calibración - Prueba

Aquí los datos están estandarizados y listos para ser aplicado hacia los modelos. Sin embargo, en este paso se dividió los datos usando un muestreo aleatorio simple donde los datos de calibración son el 80% y para la prueba el 20%, esta división en particular se realizó porque en la literatura de split_train_test sugiere en general usar una partición alrededor de 80/20 a 70/30, si el estudio contiene poca cantidad de datos como es en este caso se sugiere usar una partición 80/20, esto para que los modelos no experimenten overfitting o underfitting. En otras palabras, si la partición es menor como 60/40 el modelo va experimentar un underfitting es decir no tendrá la capacidad de poder memorizar los patrones de los datos y las predicciones serán erróneas, por otro lado, si la partición es mayor como 90/10 se experimenta un overfitting donde los modelos memorizan todos los patrones y cuando se les presentan nuevos datos estos no son capaces de predecir correctamente.

Para esta partición se usó la librería model_selection y el método train_test_split, se estableció al parámetro test_size el valor de 0.2 y a la semilla random_state un valor 360.

En el trabajo se podría haber usado algún método mas sofisticado como una validación cruzada. Sin embargo, no se optó por este método por la poca cantidad de datos.

B. Calibración de Clasificadores

La calibración de un modelo predicativo consiste en calcular los parámetros más óptimos de un modelo a través de diversas técnicas [10], una de ellas es una búsqueda de parámetros óptimos por iteraciones, donde en cada iteración se evalúan cada combinación de parámetros y se visualiza el desempeño que tiene.

En esta parte se realizó la calibración de 5 modelos, para ello se utilizó algunos algoritmos de la librería sklearn.

En la Tabla VI, se presentan los métodos y parámetros óptimos usados para cada modelo, si bien se pudo haber usados más modelos tales como Arboles de Decisiones, K-vecinos más cercanos, Bosques Aleatorios, etc., todos estos modelos no mencionados generaron un bajo rendimiento en la etapa de validación.

TABLA VI
CLASIFICADORES CON LOS PARÁMETROS ÓPTIMOS

Clasificador	Método Sklearn	Con Riesgo
Clasificador de Vectores de Soporte	SVC	Kernel = lineal, regularización C=0.065
Red Neuronal Artificial	MLPClassifier	Alfa=1, max_iteracion=50,000
Clasificador Bayesiano	GaussianNB	Parámetros por defecto
Análisis Cuadrático Discriminante	QuadraticDiscriminantAnalysis	Parámetros por defecto
Regresión Logística	LogisticRegesion	Regularización C=1e4

VI. RESULTADOS DE CLASIFICADORES

En esta sección se van a presentar los resultados de la calibración de cada modelo tanto para los datos de prueba como para el de validación. Sin embargo, como fase inicial es necesario mencionar cual es la medida de rendimiento, es decir se tiene que establecer las métricas de éxito, para poder decidir cuál es el mejor modelo entrenado.

A. Medida de Rendimiento

En este trabajo se estableció la Medida de Rendimiento como una métrica derivada de la matriz de confusión, para escoger alguna de las métricas se definió los valores de la matriz de confusión de acuerdo al problema que se aborda. El problema a solucionar es clasificar a un estudiante si está en riesgo o sin riesgo, cuando el modelo realiza una predicción de 1 significa que el estudiante está en riesgo, si se predice 0 el estudiante esta sin riesgo.

-	(real sin riesgo) $Y_i = 0$	(real en riesgo) $Y_i = 1$
(predicción sin riesgo) $Y_i = 0$	P_{11}	P_{12}
(predicción en riesgo) $Y_i = 1$	P_{21}	P_{22}

Fig 8. Matriz de Confusion de acuerdo al problema tratado.

En la Fig. 8, se tiene la Matriz de Confusión donde, a partir de esta se definen los valores de P.

- P11: Verdadero Negativo (TN), Aquellos estudiantes que han sido estimados en riesgo y en realidad están sin riesgo.
- P12: Falso Negativo (FN), Aquellos estudiantes que han sido estimados sin riesgo. Sin embargo, estos si se encuentran en riesgo. Dentro de la problemática estos alumnos representan un gasto para la entidad.
- P21: Falso Positivo (FP), Aquellos estudiantes que han sido estimados en riesgo, pero en realidad no lo están. En esta parte sucede todo lo contrario al Falso Negativo, aquí el error es más crítico ya que el modelo no es capaz de predecir que un estudiante va estar en riesgo.
- P22: Verdadero Positivo (TP), Aquellos estudiantes que han sido estimados en riesgo y en realidad si lo están

Para seleccionar una Medida de Rendimiento es importante mencionar las que se tomaron en este estudio.

- **Sensibilidad**: También conocido como Recall o Sensitivity, es la proporción de casos positivos que fueron correctamente identificadas por el algoritmo.

  La formula es la siguiente: ***TP / (TP + FN)***, en otras palabras, es los Verdaderos Positivos / Total de estudiantes en riesgo, en este estudio podríamos decir que *la sensibilidad es la capacidad de poder identificar correctamente el riesgo entre los estudiantes*.

- **Especificidad**: También conocido como Especificity, esta trata de los casos negativos que el algoritmo a clasificado correctamente.
  Se calcula como: ***TN / (TN + FP)***, en otras palabras, es los Verdaderos Negativos / Total de estudiantes sin riego, en este estudio podríamos decir que *la especificidad es la capacidad de poder identificar los casos de los estudiantes sin riesgo entre todos los estudiantes sin riesgo.*

La selección de alguna Medida de Rendimiento va depende mucho de la entidad, por ejemplo, si la entidad no acepta valores en los Falso Negativo se debe usar la **Sensibilidad**, en cambio si no se aceptar valores en los Falsos Positivos se debe priorizar por la **Especificidad**, y si solo le interesa los aciertos del modelo se optará por la Tasa de aciertos.

En este estudio la entidad es la Facultad de Ciencias de la Universidad Nacional de Ingeniería, esta no cuenta con muchos recursos para poder aceptar valores muy grandes en la sección de Falso Positivo aquellos que representan un gasto para la entidad. Es por ello que se optó por la métrica de Sensibilidad, donde se quiere tener la menor cantidad de fallos en la sección Falso Negativo, es decir es preferible realizar un seguimiento a un estudiante que supuestamente va estar en riesgo, a que no realizar el seguimiento y que este estudiante al finalizar el semestre este en riesgo.

Otra medida que va complementar a la Sensibilidad y que es más robusta es la Curva ROC, con esta medida podremos saber gráficamente que modelo genera un mejor desempeño.

B. Resultados de los Datos de Prueba

Los datos de test son el 20 % de los datos iniciales, como primer vistazo del mejor modelo se tiene la Fig. 9, en esta se gráfica las curvas ROC de cada modelo, se puede ver que los 2 mejores modelos son Naives Bayes y Neural Net ambos con un área bajo la curva de 0.94, seguidamente está el modelo QDA con un área de 0.88, luego esta Linear SVM con un área de 0.86 y finalmente esta Logic Regre con un área de 0.82.

Todos estos modelos antes mencionados tienen un buen rendimiento ya que superan un valor en el área mayor a 0.80. Entonces según la curva ROC se tiene un empate entre 2 modelos para saber cuál es el mejor.

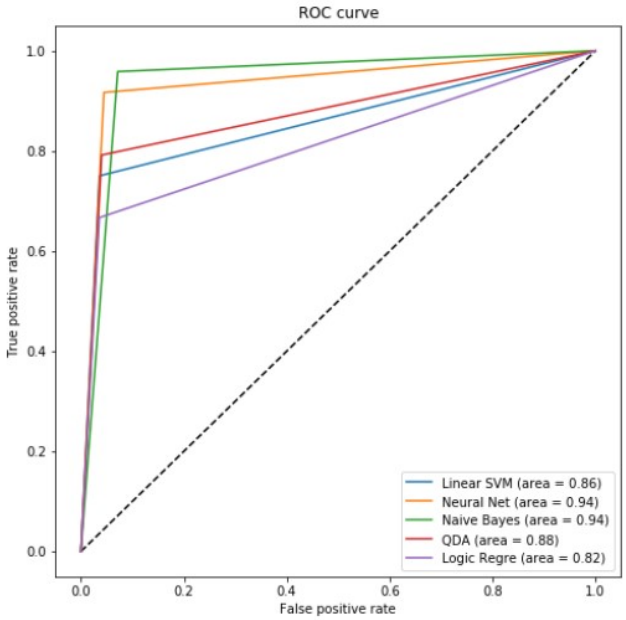


Fig 10. Curvas ROC usando datos de prueba.

En la Tabla VII, se encuentra a detalle los resultados esta se encuentra ordenada por la sensibilidad de mayor a menor, donde el primer registro es el modelo de Naives Bayes este modelo tiene la menor perdida en Falsos Negativos con un valor de 1 además tiene un valor en Falsos Positivos de 16.

Por otra parte, el segundo registro está el modelo Neural Net con un valor aceptable en los Falsos Negativos y con un mejor rendimiento en los Falsos Positivos con solo 10 estudiantes. Así que se podría decir que el mejor modelo es

Naives Bayes para una entidad con suficientes recursos, y si la entidad no tiene suficientes recursos el mejor modelo es Neural Net.

En esta parte se ha determinado los posibles candidatos para ser el mejor modelo, se dice posiblemente porque el mejor modelo se determina con los datos de validación que se podrán encontrar en la siguiente sección.

TABLA VII
RESULTADOS DE LOS DATOS DE PRUEBA

Clasificador	TN	TP	FN	FP	Sensibilidad
Naive Bayes	208	23	1	16	0.958
Neural Net	214	22	2	10	0.917
QDA	215	19	5	9	0.792
Linear SVM	216	18	6	8	0.750
Logic Regre	216	16	8	8	0.667

C. *Resultados de los Datos de Validación*

Es esta la última y definitiva fase para determinar cuál es el mejor modelo, en la Fig.10, se puede ver las gráficas de las curvas ROC de cada modelo usando la data de validación, se puede notar que la curva ROC de cada modelo se encuentra más abajo en comparación de la Fig. 9, el modelo que tuvo el menor cambio en área fue el modelo Naive Bayes con esto se demuestra la consistencia del modelo frente a los nuevos datos nunca antes visto en la fase de calibración. Los demás modelos tuvieron un cambio significativo, pero a pesar de eso poseen un buen rendimiento ya que todos tienen un área mayor o igual a 0.70.

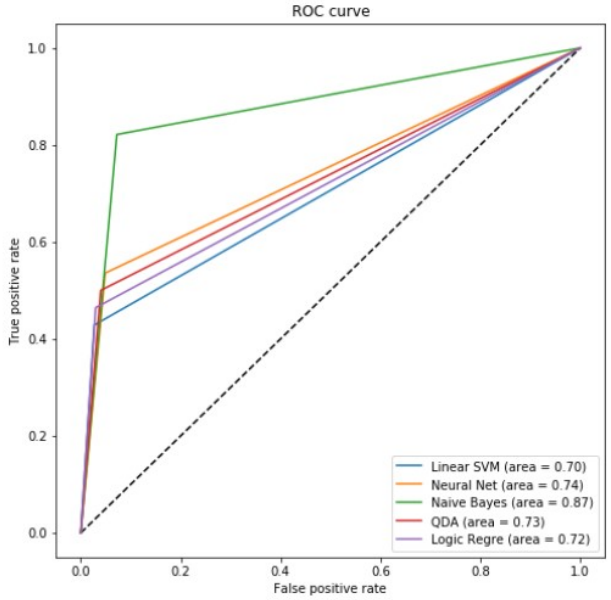


Fig 10. Curvas ROC usando datos de validación.

Se podría decir que el modelo Naive Bayes es el mejor modelo. Sin embargo, se debe tomar la métrica de la Sensibilidad.

En la Tabla VIII, se tiene en detalle los resultados de cada modelo en esta se puede apreciar que Naive Bayes tiene la mayor Sensibilidad y un valor en los Falsos Negativos de 5, también apreciar que tiene un valor muy elevado en los Falsos Positivos, esto significa que el modelo si bien es eficiente requiere que la entidad tenga mucha capacidad en recursos.

El modelo de Neural Net tiene un comportamiento similar ya que se va gastar recursos en 38 estudiantes. Un modelo aceptable para una entidad con recursos bajos es el modelo Logic Regression que a pesar de tener una baja Sensibilidad es muy útil ya que solo va gastar recursos en 21 estudiantes.

TABLA VIII
RESULTADOS DE LOS DATOS DE VALIDACIÓN

Clasificador	TN	TP	FN	FP	Sensibilidad
Naive Bayes	720	23	5	56	0.821
Neural Net	738	25	13	38	0.536
QDA	745	14	14	31	0.500
Linear SVM	753	13	15	23	0.464
Logic Regre	755	12	16	21	0.429

VII. CONCLUSIONES

Se demostró la utilidad del Aprendizaje Automático aplicado a la educación, en donde los datos académicos generan beneficios en favor de la comunidad universitaria proporcionando una mejora en la calidad educativa.

En el análisis de los datos se encontró valiosa información para trabajos futuros, tales como el periodo de verano es la característica que tiene mayor importancia frente a las otras para determinar si un estudiante va estar en riesgo. Se encontró que la mayor cantidad de alumnos en riesgo se focaliza en los cuatro primeros semestres.

Se logró determinar los mejores modelos en base a la Sensibilidad y Curva ROC como métricas de éxito. Al comparar los clasificadores se obtuvieron los 2 mejores en base a los recursos que posee la institución, si la entidad tiene suficientes recursos el mejor modelo fue el Clasificador Bayesiano que tiene un área de ROC de 0.87 y una Sensibilidad de 0.871. Por otro lado, si la entidad tiene recursos escasos el mejor modelo es Regresión Logística con un área de ROC de 0.72 y una Sensibilidad de 0.464.

AGRADECIMIENTOS

Se agradece al Vicerrectorado de Investigación de la Universidad Nacional de Ingeniería VRA-UNI y al Laboratorio 4 del Centro de Tecnologías de Información y Comunicaciones de la Universidad Nacional de Ingeniería CTIC-UNI, por el apoyo brindado con las instalaciones y

equipos para el desarrollo de esta investigación, así mismo se agradece a la Facultad de Ciencias de la Universidad Nacional de Ingeniería FC-UNI, por bridar y confiar los datos de los estudiantes en la realización del presente trabajo.

REFERENCIAS

[1] Zahyah Alharbi, James Cornford, Liam Dolder, and Beatriz De La Iglesia. Using data mining techniques to predict students at risk of poor performance. 2016.
[2] Johannes Berens, Kerstin Schneider, Simon Gortz, Simon Oster, and Julian Burghoff. Early detection of students at risk predicting student dropouts using administrative student data from German universities and machine learning methods. 2019.
[3] Miguel Gil, Norma Reyes, and Myriam Soria. Predicting early student with high risk to drop out of university using a netural netword-based approach. 2013.
[4] Field Cady. The data science handbook. USA: Springer, 2017.
[5] Aurélien Géron. Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow, volume Second. USA: O'Reilly, 2019.
[6] Jason Brownlee. Data Preparation for Machine Learning. USA: Machine Learning Mastery, 2019.
[7] Max Kuhn and Kjell Johnson. Applied Predictive Modeling. USA: Springer Science Business Media, 2013.
[8] Max Kuhn and Kjell Johnson. Feature Engineering and Selection: A Practical Approach for Predictive Models, volume 1st Edition. USA: CRC Press, 2019.
[9] C. Mihaescu and D. Burdescu. Testing attribute selection algorithms for classification performance on real data. 2006.
[10] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. He Elements of Statistical Learning: Data Mining, Inference, and Prediction. USA: Springer Series in Statistics, 2017.