Prediction of the opening of university schedule groups using data mining techniques

Predicción de la apertura de grupos de horarios universitarios utilizando técnicas de minería de datos

Aradiel Castañeda Hilario, Doctor[1], Acosta de la Cruz Pedro Raul, Msc[1], Mas Azahuanche Guillermo Antonio, Doctor[2],
Universidad Nacional de Ingeniería, Peru, haradiel@uni.edu.pe, pacosta@uni.edu.pe
Universidad Nacional del Callao, Peru, gamasa@unac.edu.pe

Resumen*—Las universidades tienen un papel importante en la prestación de servicios educativos, especialmente en la elaboración de la programación académica del semestre. Esta programación se realiza a base de prueba y error, es decir, se programan cursos y no se abren porque hay pocos o ningún alumno matriculado, afectando gravemente los servicios educativos. La minería de datos es la mejor solución para encontrar patrones ocultos y ofrecer sugerencias para mejorar la programación académica del semestre. En este artículo se presenta un modelo basado en algoritmos predictivos supervisados para predecir con precisión el número de grupos de programación académica. En este modelo predictivo se utilizó la clasificación de árboles de decisión, además de los expedientes académicos de los estudiantes. El registro académico del estudiante consta de 8 campos: identificación del estudiante, semestre de estudio, identificación del curso, grupo de horario, número de créditos del curso, estado de aprobación del curso y lugar de estudio. Se recogieron un total de 44737 registros. La precisión del modelo fue del 89.5% con un error de 10.5%. La biblioteca Python sklearn se utilizó para construir este modelo. Finalmente, los resultados muestran en los indicadores de la cantidad de alumnos por estado (aprobado, desaprobado y NSP) y el indicador porcentaje de grupos que nos muestra por cada ciclo los cursos programados.

Palabras clave: minería de datos, arboles de decisión, grupo horario.

I. INTRODUCCIÓN

Recientemente, el creciente volumen de datos y su uso para mejorar los procesos de los servicios académicos es uno de los principales retos de las universidades. Las universidades generalmente están interesadas en brindar un servicio académico de calidad a los estudiantes durante sus estudios [1] [2]. Las universidades tienen un gran conjunto de información sobre los estudiantes almacenada en sus bases de datos. Sin embargo, el almacenamiento no es un problema. Manejar los datos, extraer patrones y descubrir el conocimiento almacenado en la base de datos masiva es tremendamente difícil. En consecuencia, la minería de datos puede considerarse una herramienta prometedora para lograr estos objetivos [3].

La minería de datos se utiliza para detectar y extraer información relevante y valiosa de un gran volumen de datos. Este proceso ha ganado mucha atención y preocupación por parte de la industria de la información y la sociedad. Esta técnica está recibiendo una atención significativa en el análisis de datos y ha sido reconocida como una herramienta emergente para el análisis [4].

En particular, las herramientas de minería de datos se han vuelto muy populares entre los investigadores y usuarios debido a su facilidad de uso y disponibilidad, como las siguientes herramientas (la mayoría de ellas están disponibles en línea): Microsoft Excel, SPSS, Python y R. Varias de estas herramientas están disponibles gratuitamente para las universidades con las que pueden beneficiarse con su conocimiento de aplicación existente. Accesibilidad, disponibilidad, facilidad de uso y comprensibilidad son las principales razones para incluir estas herramientas en la investigación. La biblioteca Python Sklearn [5] ha sido elegida como una herramienta de análisis de minería de datos para respaldar la conclusión debido a sus resultados altamente legibles y comprensibles.

Nuestro objetivo es analizar la información colectiva de los estudiantes a través de los expedientes académicos que se encuentran en la base de datos de la universidad, y clasificar los datos para predecir la apertura de horarios académicos de los grupos. El trabajo busca encontrar la posibilidad de apoyarse en los resultados de los algoritmos de Random Forest para apoyar las decisiones académicas relacionadas con la apertura de grupos horarios y dar una hoja de ruta adecuada para estudiantes y docentes. Este conocimiento recién descubierto puede ayudar a las universidades a llevar a cabo mejoras en los servicios académicos y ayudar a los estudiantes a desempeñarse mejor académicamente.

II. ANTECEDENTES

Heli León [6] pretende comprender las causas de la deserción en un proceso psicoterapéutico para poder predecir, desde el primer contacto entre el paciente y la institución, la permanencia del paciente. Para ello se propone el desarrollo de un prototipo funcional para predecir la permanencia de los pacientes utilizando algoritmos de árboles de decisión para la predicción. Para la elaboración del prototipo funcional y el cumplimiento de los objetivos, se hizo uso de la herramienta Weka, la cual permitió el análisis y selección del algoritmo a utilizar para implementar el prototipo. Las clases desequilibradas dificultaron el proceso de análisis algorítmico; por lo tanto, la minería de datos

Se aplicaron métodos para analizar los conjuntos de datos desequilibrados. El lenguaje de programación utilizado fue Java y los algoritmos que permitieron la predicción fueron

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2022.1.1.264
ISBN: 978-628-95207-0-5 **ISSN:** 2414-6390

incorporados de las bibliotecas API de Weka. Los resultados obtenidos fueron satisfactorios, en base a los datos que se extrajeron de la base de datos institucional.

El objetivo de Emir Piscoya [7] es proponer una herramienta mediante técnicas de minería de datos, que permita al usuario tener acceso a información veraz donde se realicen predicciones sobre los estudiantes que ingresarán en los próximos años, obteniendo resultados a corto plazo, que asegurará su confiabilidad, sirviendo de apoyo a la institución para futuras decisiones que pueda tomar. Dentro de las técnicas predictivas, se determinó utilizar los algoritmos de ETS y Redes Neuronales; al realizar el análisis se descartaron algunas técnicas adicionales por no contar con los criterios necesarios para su implementación en el modelo a desarrollar.

El principal objetivo de Dante Mayorca [8] es encontrar los atributos del servicio de telefonía móvil que afectarían significativamente la satisfacción de sus usuarios en las zonas urbanas. Para ello se utilizó información de la encuesta "Estudio sobre el nivel de satisfacción de los usuarios de telecomunicaciones y el nivel de conocimiento de los derechos y obligaciones de los usuarios de los servicios públicos de telecomunicaciones", la cual es representativa a nivel nacional y fue realizada por OSIPTEL. Para llevar a cabo este análisis se realizó un análisis factorial sobre las variables de calificación de los diferentes atributos del servicio de telefonía móvil. Posteriormente, se estimó el efecto que tienen sobre la satisfacción general con el servicio mediante un modelo Logit Ordenado.

A. Minería de datos

La minería de datos busca obtener conocimiento de diferentes fuentes y para diferentes sectores; por ejemplo, se pueden utilizar para mejorar la productividad, mejorar los productos, aumentar las ventas y otros [9].

La disponibilidad de grandes volúmenes de información y el uso generalizado de herramientas informáticas ha transformado el análisis de datos orientado hacia determinadas técnicas especializadas englobadas bajo el nombre de minería de datos o Data Mining.

Las técnicas de minería de datos persiguen el descubrimiento automático de conocimiento en la información almacenada de forma ordenada en grandes bases de datos. Estas técnicas tienen como objetivo descubrir patrones, perfiles y tendencias a través del análisis de datos utilizando tecnología de reconocimiento de patrones, redes neuronales, lógica difusa, algoritmos genéticos y otras técnicas avanzadas de análisis de datos [10].

B. Bosque aleatorio

Los bosques aleatorios [11] son un algoritmo de aprendizaje de conjunto que se puede usar para la clasificación, es decir, para predecir una variable de respuesta categórica, y también para la regresión, que implica predecir una variable de respuesta continua. Los modelos de clasificación y regresión de bosque aleatorio ajustan un conjunto de modelos de árboles de decisión a un conjunto de datos. Para cada árbol, los datos se dividen recursivamente en unidades más homogéneas, que comúnmente se denominan nodos, para mejorar la predictibilidad de la variable de respuesta. Los puntos de división se basan en los valores de las variables predictoras. Así, las variables utilizadas para dividir los datos se consideran importantes variables explicativas. Los bosques aleatorios ajustan árboles de decisión separados a un número predefinido de conjuntos de datos de arranque. El valor predicho de una respuesta categórica es la moda de las clases de todos los árboles de decisión ajustados individuales, y el valor predicho de una respuesta continua es la respuesta ajustada media de todos los árboles individuales que resultaron de cada muestra autocargada. Los modelos de clasificación y regresión de bosques aleatorios se construyeron utilizando el paquete "randomForest" [12] en el software estadístico gratuito R (R Core Team 2014). Se construyeron modelos de regresión de bosque aleatorio utilizando 500 árboles derivados de 500 conjuntos de datos de arranque. Los puntos de división se eligieron de un subconjunto aleatorio de todas las variables predictoras disponibles [11]. De forma predeterminada, el tamaño del subconjunto aleatorio del paquete randomForest es la raíz cuadrada del número de predictores para los modelos de clasificación y un tercio de todas las variables predictoras disponibles para los modelos de regresión [12]. También por defecto, cada nodo está restringido a un tamaño mínimo de uno para clasificación o cinco para regresión. Los tamaños de nodo más grandes dan como resultado que se cultiven árboles más pequeños, lo que reduce el tiempo de cómputo [12]. En este estudio, los valores predeterminados del algoritmo se usaron ya que las pruebas preliminares mostraron poca o ninguna mejora si se modificaban los valores predeterminados. Esto tenía la ventaja añadida de mantener las entradas constantes para las tres fechas previstas.

El algoritmo de bosque aleatorio puede clasificar la importancia relativa de cada variable predictora. La importancia de la variable se basa en el error de predicción de regresión de la porción de los datos fuera de bolsa, también llamada OOB [11] [12]. Aproximadamente el 30% de los datos son OOB y no se usan para construir el árbol [13]. Para los modelos de clasificación, el error de predicción se calcula como la tasa de error de clasificación, mientras que, para la regresión, se calcula el error cuadrático medio. En el paquete randomForest edad, la importancia de la variable predictora se informa como una disminución porcentual media en la tasa de clasificación para el modelo de clasificación o un aumento medio en el error cuadrático medio para el modelo de regresión si esa variable se eliminó del análisis.

III. METODOLOGÍA

A. Preprocesamiento de datos

1) Recolección de datos: En esta investigación se tomó como caso de estudio los expedientes de la Escuela Profesional de Ingeniería en Sistemas de la Universidad Nacional del Callao (UNAC). Se consideró para este estudio el ciclo verano, donde se matriculan los alumnos que desean adelantar sus cursos. Los registros obtenidos de la base de datos del centro de cómputo de la UNAC son 44737 con datos de código de estudiante,

grupo horario, número de semestre, estatus (aprobado/reprobado) y lugar de estudio.

Se realizó una descripción de los datos utilizados en los registros. La descripción se muestra en la tabla I.

TABLE I
DESCRIPCION DEL REGISTRO DE LA DATA

Data	Descripción
Identificación del Estudiante	Se identifica a través de un código único y está compuesto por 10 dígitos.
Número de semestre	Representa la progresión curricular en el plan de estudios.
Identificación del curso	Expresado como un código alfanumérico.
Sección	Indica el grupo de programación
Créditos	El peso de cada curso
Condición	Condición lógica (aprobado/fallido)
Semestre	Codificado por semestre

Los datos fueron acondicionados para asegurar resultados válidos. Esto implicó la eliminación de valores de atributos incorrectos producidos por errores humanos, errores computacionales y datos erróneos ingresados debido a campos de entrada obligatorios.

Se eliminaron los registros con datos que pudieran sesgar los resultados de la minería. Los métodos gráficos y numéricos basados en medianas y "cuartiles" se utilizaron principalmente para buscar datos irregulares. Estos permitieron detectar la presencia de conglomerados y valores atípicos. Para esta tarea se utilizó el lenguaje de base de datos T-SQL

2.- Transformación de los datos

La principal transformación se realizó con los valores de la variable Tipo de código de grupo horario. Esto se debió a que no están codificados de manera estándar, y la clasificación utilizada es excesivamente extensa. Se usó el Sistema de Clasificación Los valores de la variable grupo horario fueron transformados a una escala de Numérica

3.- Descubrimiento de información

Con la participación de expertos del dominio, a partir de los agrupamientos encontrados, se identificaron los patrones. Los principales resultados obtenidos de la primera etapa del trabajo se presentan en la siguiente sección.

4.- Consideraciones en el Diseño

Consideraciones para pronosticar con la base limpia

a) Se consideraron los alumnos de la sede CALLAO Y LA ANTIGUA MALLA (AM) Y NUEVA MALLA (NM).

b) Se elaboró una tabla aparte donde se muestran los cursos de la Antigua y Nueva Malla, a los cursos de la nueva malla se les designo un código ya que estos venían sin código numérico. (101 – 168)

TABLA II
PRE REQUISITOS NUEVA MALLA

CICLO	CÓDIGO	REQUISITO	CANTIDAD-ALUMNO-SECCION*
I	101	Ninguno	377
	102	Ninguno	
	103	Ninguno	
	104	Ninguno	
	105	Ninguno	
	106	Ninguno	
II	107	Ninguno	323
	108	105	
	109	102	
	110	101	
	111	103	
	112	106	
III	113	107, 101	351
	114	110	
	115	108	
	116	108	
	117	110	
	118	112	
IV	119	105	511
	120	102	
	121	113	
	122	104	
	123	115	
	124	118	
V	125	114	400
	126	119	
	127	117	
	128	121	
	129	116	
	130	124	
VI	131	116	688
	132	123,129	
	133	128	
	134	122	
	135	125	
	136	119,123	

(*) la cantidad de alumnos por sección significa que el alumno aparece con todos los grupos horarios

TABLA III
PRE REQUISITOS ANTIGUA MALLA

CICLO	CODIGO		PRE-REQ	CANTIDAD ALUMNO -SECCION
VII	36	PCO71	31,32	94
	37	PCO72	21	
	38	BEC73	33	
	39	PSI74	34	
	40	PGE75	30	
	59	EIN77	56	
	60	ETC78	57	
	CODIGO			
VIII	41	PCO81	36,37	337
	42	PSI82	28,33	
	43	PCO83	37	
	44	PGE84	29	
	45	PGE85	18	
	61	EIN86	39	
	62	ETC87	27	
	CODIGO			
IX	46	PCO91	41	444
	47	PSI92	24,42	
	48	PCO93	43	
	49	PSI94	39	
	50	PSI95	44	
	63	EIN96	NINGUNO	
	64	ETC97	62	
	CODIGO			
X	51	BHU01	46	252
	52	PSI02	42	
	53	PSI03	49	
	54	PSI04	50	
	55	PGE05	45	
	65	EIN06	39	
	66	ETC07	64	

c) Se pasó a dato numérico el campo de Aprobado / Desaprobado

TABLA IV
CODIFICACION DEAPROBADO-DESAPROBADO

Desaprobado	0
Aprobado	1
No se presento	2

TABLA V
EJEMPLO CODIFICACION

Aprobado / Desaprobado	COD_APP
APROBADO	1
APROBADO	1
NO SE PRESENTO	2
APROBADO	1
NO SE PRESENTO	2
APROBADO	1
DESAPROBADO	0

d) Se añadió el campo TERMINAL que es una condición para determinar si el curso tiene o no prerrequisito.

TABLA VI
CODIFICACION DE TERMINAL

No es prerrequisito	TERMINAL	0
Es prerrequisito	NO TERMINAL	1

TABLA VII
EJEMPLO CODIFICACION DE TERMINAL

COD_CURSO	Créditos	Aprobado / Desaprobado	COD_APP	Semestre	Sede	TERMINAL
136	3	APROBADO	1	2018B	CALLAO	0
115	4	APROBADO	1	2018A	CALLAO	1
136	3	NO SE PRESENTO	2	2018B	CALLAO	0
105	4	APROBADO	1	2017A	CALLAO	1

e) Se añadió el campo FUTURO que es una condición para determinar si el alumno TERMINA, CONTINÚA O REPITE el curso.

TABLA VIII
CODIFICACION DE FUTURO

Aprobado	No terminal	Termino
Aprobado	Terminal	Continua
No se presento	No terminal	Repite
Desaprobado	NO terminal	Repite

TABLA IX
EJEMPLO DE LA CODIFICACION FUTURO

Aprobado / Desaprobado	COD_APP	Semestre	Sede	TERMINAL	FUTURO
APROBADO	1	2018B	CALLAO	0	TERMINO
APROBADO	1	2018A	CALLAO	1	CONTINUA
NO SE PRESENTO	2	2018B	CALLAO	0	REPITE
APROBADO	1	2017A	CALLAO	1	CONTINUA
NO SE PRESENTO	2	2017B	CALLAO	0	REPITE
APROBADO	1	2019S	CALLAO	0	TERMINO
DESAPROBADO	0	2018B	CALLAO	0	REPITE

f) Como la variable es compleja, se tuvo que realizar un árbol predictivo para cada ciclo.

g) para la validez de cada árbol predictivo, se midió por el error o riesgo, el riesgo es la magnitud de que el árbol fracase al momento de predecir

TABLA X
ERROR

estimación	Error estándar
0.105	0.10

Método de crecimiento: CRT

Variable dependiente: Cod_curso

Precisión: 89.5% entre más grande sea la precisión mayor es la predicción.

El 10.5% es el error o riesgo es el % de fracaso de la técnica predictivo del árbol, a veces acierta y a veces no, eso es el riesgo

III. RESULTADOS

ANÁLISIS DESCRIPTIVO

En la presente investigación se utilizó la minería de datos aplicando árboles de decisión para medir los indicadores porcentaje de grupos horarios y la cantidad de alumnos por estado: de repitencia, NSP y aprobados, los cuales nos permitieron evaluar los resultados y cómo ha influido la minería de datos en la predicción de los grupos horarios en Escuela Profesional de Ingeniería de Sistemas de la Universidad Nacional del Callao.

Indicador: Porcentaje de Grupos horarios por curso

Los resultados descriptivos del porcentaje de grupos horarios en la predicción de los grupos horarios. Del primero al sexto ciclo de la nueva currícula y del séptimo al décimo ciclo antigua currícula, se evidencian en la siguiente tabla. IX

TABLA XI
CANTIDAD DE ALUMNOS POR GRUPO HORARIO-CICLO 1

Observed	Predicted						
	101	102	103	104	105	106	Percent Correct
101	39	0	0	31	0	0	55.7%
102	14	0	0	40	0	0	0.0%
103	28	0	0	34	0	0	0.0%
104	19	0	0	51	0	0	72.9%
105	23	0	0	31	0	0	0.0%
106	18	0	0	49	0	0	0.0%
Overall Percentage	37.4%	0.0%	0.0%	62.6%	0.0%	0.0%	23.9%

Significa que para el grupo horario 101 se matricularan 39 alumnos que han repetido el curso, que representa un 55.7% y 31 alumnos que aprobaron el curso el curso y pasan al siguiente ciclo que representa el 44.3%. Para grupo horario 102 se matricularán 14 alumnos que han repetido el curso y 40 alumnos que aprobaron el curso y pasan al siguiente ciclo. Para el grupo horario 103 se matricularán 28 alumnos que han repetido el curso, y 34 alumnos que aprobaron el curso y pasan al siguiente ciclo. Para grupo horario 104, se matricularán 19 alumnos que han repetido el curso, que representa el 72.9% y 51 alumnos que aprobaron el curso y pasan al siguiente ciclo. Para el grupo horario 105, se matricularán 23 alumnos que repitieron el curso y 31 alumnos que aprobaron el curso que pasan al siguiente ciclo. Para el grupo horario 106, se matricularán 18 alumnos que repitieron el curso y 49 alumnos que aprobaron el curso y pasan al siguiente ciclo. El total alumnos que se matricularán en el primer ciclo son 141 y 236 alumnos pasan al ciclo 2.

TABLA XII

CANTIDAD DE ALUMNOS POR GRUPO HORARIO-CICLO 2

Observed	Predicted						
	107	108	109	110	111	112	Percent Correct
107	15	0	7	0	0	31	28.3%
108	3	0	2	0	0	46	0.0%
109	1	0	17	0	0	53	23.9%
110	11	0	6	0	0	34	0.0%
111	0	0	0	0	40	0	100.0%
112	2	0	2	0	0	53	93.0%
Overall Percentage	9.9%	0.0%	10.5%	0.0%	12.4%	67.2%	38.7%

Significa que para el grupo horario 107 se matricularan 15 alumnos que desaprobaron el curso, que representa un 28.3%, 7 alumnos que nunca se presentaron al curso y 31 alumnos que aprobaron pasan al siguiente ciclo. Para grupo horario 108 se matricularán 3 alumnos que desaprobaron el curso, 2 alumnos que nunca se presentaron y 46 alumnos que aprobaron el curso y pasan al siguiente ciclo. Para grupo horario 109 se matricularán 1 alumno, 17 alumnos que nunca se presentaron y representa el 23.9%, y 53 alumnos que aprobaron y pasan al siguiente ciclo. Para grupo horario 110, se matricularán 11 alumnos que desaprobaron, 6 alumnos que nunca se presentaron y 34 alumnos que aprobaron y pasan al siguiente ciclo. Para el grupo horario 111, hay 40 alumnos que aprobaron y terminan sus prerrequisitos. Y en el grupo horario 112, se matricularán 2 alumnos desaprobados, 2 alumnos que nunca se presentaron y 53 alumnos que aprobaron y pasan al siguiente ciclo que representa el 93.0%. El total alumnos que se matricularan en el segundo ciclo son 106 y 217 pasan al siguiente ciclo.

TABLA XIII
CANTIDAD DE ALUMNOS POR GRUPO HORARIO CICLO 3

Classification

Obs.	Predicted						
	113	114	115	116	117	118	Percent Correct
113	17	28	0	5	0	0	34.0%
114	5	56	0	6	0	0	83.6%
115	7	53	0	0	0	0	0.0%
116	3	50	0	8	0	0	13.1%
117	0	0	0	0	54	0	100.0%
118	9	50	0	0	0	0	0.0%
Overall Percen	11.7%	67.5%	0.0%	5.4%	15.4%	0.0%	38.5%

Significa que para el grupo horario 113 se matricularan 17 alumnos desaprobados, que representa un 34.0%, 5 alumnos que nunca se presentaron y 28 alumnos que aprobaron y pasan al siguiente ciclo. Para grupo horario 114 se matricularán 5 alumnos desaprobados, 6 alumnos que nunca se presentaron y 56 alumnos aprobados y pasan al siguiente ciclo que representa 83.6%. Para grupo horario 115, se matricularán 7 alumnos desaprobados y 53 alumnos aprobados que pasan al siguiente ciclo. Para grupo horario 116, se matricularán 3 alumnos desaprobados, 8 alumnos que nunca se presentaron y representa el 13.1% y 50 alumnos que aprobaron pasan al siguiente ciclo. Para el grupo horario 117, hay 54 alumnos aprobados y terminan sus prerrequisitos y el grupo horario 118, se matricularon 9 alumnos desaprobados y 50 alumnos que aprobaron y pasan al siguiente ciclo. El total alumnos que se matricularán en el tercer ciclo son 114 y 237 pasan al siguiente ciclo

TABLA XIV
CANTIDAD DE ALUMNOS POR GRUPO HORARIO CICLO 4

Classification

Observed	Predicted						
	119	120	121	122	123	124	Percent Correct
119	22	0	0	0	0	53	29.3%
120	0	96	0	0	0	0	100.0%
121	8	0	0	0	0	59	0.0%
122	2	0	0	0	0	79	0.0%
123	9	0	0	0	0	69	0.0%
124	4	0	0	0	0	110	96.5%
Overall Percentage	8.8%	18.8 %	0.0%	0.0%	0.0%	72.4 %	44.6%

Significa que para el grupo horario 119 se matricularán 22 alumnos desaprobados y alumnos que nunca se presentaron que representa un 29.3% y 53 alumnos que aprobaron y pasan al siguiente ciclo. Para el grupo horario 120 aprobaron 96 alumnos y terminan sus prerrequisitos, que representa 100.0%. Para el grupo horario 121, se matricularán 8 alumnos desaprobados y alumnos que nunca se presentaron y 59 alumnos que aprobaron y pasan al siguiente ciclo. Para grupo horario 122, se matricularán 2 alumnos desaprobados y 79 alumnos aprobados y pasan al siguiente ciclo. Para el grupo horario 123, se matricularán 9 alumnos desaprobados y 69 alumnos aprobados y pasan al siguiente ciclo y el grupo horario 124, se matricularán 4 alumnos desaprobados y 110 alumnos aprobados que pasan al siguiente ciclo, que representa el 96.5%, el total alumnos que se matricularan en el cuarto ciclo son 45 y 370 pasan al siguiente ciclo.

TABLA XV
CANTIDAD DE ALUMNOS POR GRUPO HORARIO CICLO 5

Classification

Observed	Predicted						
	125	126	127	128	129	130	Percent Correct
125	0	0	0	0	66	0	0.0%
126	0	66	0	0	0	0	100.0%
127	0	0	0	0	51	0	0.0%
128	0	0	0	0	75	0	0.0%
129	0	0	0	0	80	0	100.0%
130	0	62	0	0	0	0	0.0%
Overall Percentage	0.0%	32.0%	0.0%	0.0%	68.0%	0.0%	36.5%

Significa que para el grupo horario 125 se matricularán 3 alumnos desaprobados y 63 alumnos aprobados que pasan al siguiente ciclo. Para el grupo horario 126 se aprobaron 66 alumnos y terminan sus prerrequisitos. Para grupo horario 127, se matricularán 13 alumnos y 38 alumnos que aprobaron y pasan al siguiente ciclo. Para grupo horario 128, se matricularán 12 alumnos desaprobados y 63 alumnos aprobados que pasan al siguiente ciclo. Para el grupo horario 129, se matricularán 2 alumnos desaprobados y 78 alumnos que aprobaron y pasan al siguiente ciclo y en el grupo horario 130, aprobaron 62 alumnos que terminan sus prerrequisitos. El total alumnos que se matricularán en el quinto ciclo son 30, 242 alumnos que aprobaron pasan al siguiente ciclo y 128 alumnos que ya no continúan porque terminan sus prerrequisitos.

TABLA XVI
CANTIDAD DE ALUMNOS POR GRUPO HORARIO CICLO 6

Classification

Observed	Predicted						
	131	132	133	134	135	136	Percent Correct
131	0	0	0	97	0	0	0.0%
132	0	0	7	98	0	0	0.0%
133	0	0	33	69	0	0	32.4%
134	0	0	9	148	0	0	94.3%
135	0	0	0	0	45	69	39.5%
136	0	0	0	0	9	104	92.0%
Overall Percentage	0.0%	0.0%	7.1%	59.9 %	7.8%	25.1 %	48.0%

Significa que para el grupo horario 131 no existe alumnos desaprobados, pero hay 97 alumnos que aprobaron y pasan al siguiente ciclo. Para grupo horario 132 se matricularán 7 alumnos desaprobados y 98 alumnos que aprueban y pasan al siguiente ciclo. Para grupo horario 133, se matricularán 33 alumnos que repiten el curso representa el 32.4% y 69 que aprueban y pasan al siguiente ciclo. Para el grupo horario 134, se matricularán 9 que repiten el curso y 148 alumnos que aprobaron y pasan al siguiente ciclo, representa el 94.3%. Para el grupo horario 135, se matricularán 45 alumnos que están desaprobados, que representa el 39.5% y 69 alumnos que están aprobados y pasan al siguiente curso. El grupo horario 136, se matricularán 9 alumnos y 104 alumnos que aprobaron y pasan al siguiente ciclo, que representa el 92.0%. El total alumnos que se matricularan en el sexto ciclo son 284 y 404 que aprobaron y pasan al siguiente ciclo

TABLA XVII
CANTIDAD DE ALUMNOS POR GRUPO HORARIO CICLO 7

Classification

Observed	Predicted							
	36	37	38	39	40	59	60	Percent Correct
36	36	0	0	0	0	0	0	100.0%
37	9	2	0	0	0	0	0	18.2%
38	0	0	0	0	0	4	0	0.0%
39	14	0	0	0	0	0	0	0.0%
40	0	0	0	0	0	8	0	0.0%
59	0	0	0	0	0	12	0	100.0%
60	0	0	0	0	0	9	0	0.0%
Overall Percentage	62.8%	2.1%	0.0 %	0.0 %	0.0%	35.1 %	0.0 %	53.2%

Significa que para el grupo horario 36, hay 26 alumnos aprobados y pasan al siguiente ciclo. Para el grupo horario 37, hay 2 alumnos desaprobados, que representan el 18.2% y 9

alumnos aprobados y pasan al siguiente ciclo. Para el grupo horario 38 hay 4 alumnos aprobados y terminan sus prerrequisitos. Para el grupo horario 39, hay 14 alumnos aprobados y pasan al siguiente ciclo. Para el grupo horario 40 hay 8 alumnos aprobados y pasan al siguiente ciclo. Para el grupo horario 59 hay 12 alumnos aprobados y pasan al siguiente ciclo y para el grupo horario 60 hay 9 alumnos aprobados y pasan al siguiente ciclo. Hay 19 alumnos aprobados y pasan siguiente ciclo y 33 alumnos que terminan sus prerrequisitos.

TABLA XVIII
CANTIDAD DE ALUMNOS POR GRUPO HORARIO CICLO 8

Classification

Orbs	Predicted							
	41	42	43	44	45	61	62	Percent Correct
41	48	0	0	0	0	13	0	78.7%
42	43	0	0	0	0	13	0	0.0%
43	37	0	0	0	0	26	0	0.0%
44	46	0	0	0	0	10	0	0.0%
45	35	0	0	0	0	7	0	0.0%
61	0	0	0	0	0	41	0	100.0%
62	14	0	0	0	0	4	0	0.0%
Overall Percent	66.2%	0.0%	0.0%	0.0%	0.0%	33.8%	0.0%	26.4%

Significa que para el grupo horario 41 se matricularan 13 alumnos, y aprobaron 48 que representa un 78.7%, para grupo horario 42 se matricularan 13 alumnos y 43 alumnos aprobaron para grupo horario 43, se matricularan 26 alumnos y aprobaron 37 alumnos. Para grupo horario 44, se matricularán 10 alumnos y aprobaron 46 alumnos que pasan al siguiente ciclo. Para el grupo horario 45, se matricularán 7 alumnos y aprobaron 35 alumnos que pasan al siguiente ciclo. Para el grupo horario 61 aprobaron 41 alumnos y terminan su prerrequisito y el grupo horario 62 se matricularán 4 alumnos y 14 alumnos aprobaron y terminan sus prerrequisitos. El total alumnos que se matricularan en el octavo ciclo son 114 y 223 pasan al siguiente ciclo.

TABLA XIX
CANTIDAD DE ALUMNOS POR GRUPO HORARIO CICLO 9

Classification

Bos	Predicted							
	46	47	48	49	50	63	64	Percent Correct
46	0	0	0	61	9	0	0	0.0%
47	0	41	36	0	0	0	0	53.2%
48	0	14	48	0	0	0	0	77.4%
49	0	0	0	87	3	0	0	96.7%
50	0	0	0	66	22	0	0	25.0%
63	0	4	13	0	0	0	0	0.0%
64	0	0	0	27	13	0	0	0.0%
Overall Percen	0.0%	13.3%	21.8%	54.3%	10.6%	0.0%	0.0%	44.6%

Significa que para el grupo horario 46 se matricularán 9 desaprobados y 61 alumnos que aprobaron y pasan al siguiente curso; para grupo horario 47 se matricularán 41 alumnos desaprobados, que representa 53.2% y 36 alumnos aprobados que pasan al siguiente ciclo; para grupo horario 48 se matricularán 14 alumnos desaprobados y 48 alumnos aprobados que pasan al siguiente ciclo y que representa el 77.4%, para grupo horario 49, se matricularan 3 alumnos desaprobados y 87 alumnos aprobados que pasan al siguiente ciclo, para el grupo horario 50, se matricularan 22 alumnos desaprobados , que representa el 25.0% y 66 alumnos aprobados que pasan al siguiente ciclo, para el grupo horario 63 se matricularan 17 alumnos que representa el 3.8% y el grupo horario 64 se matricularan 13 alumnos y 27 alumnos aprobados que pasan al siguiente ciclo, el total alumnos que se matricularan en el noveno ciclo son 101 y 343 pasan al siguiente ciclo

TABLA XX
CANTIDAD DE ALUMNOS PRO GRUPO HORARIO CICLO 10

Classification

Observed	Predicted						
	51	52	53	54	55	65	Percent Correct
51	0	36	13	0	0	0	0.0%
52	0	40	6	0	0	0	87.0%
53	0	38	16	0	0	0	29.6%
54	0	36	14	0	0	0	0.0%
55	0	39	0	0	0	0	0.0%
65	0	13	1	0	0	0	0.0%
Overall Percentage	0.0%	80.2%	19.8%	0.0%	0.0%	0.0%	22.2%

Significa que para el grupo horario 51 se matricularan 13 alumnos desaprobados y 36 alumnos aprobados. Para grupo horario 52 se matricularán 6 alumnos desaprobados y 40 alumnos aprobados, que representa 87.0%. Para grupo horario 53 se matricularan 16 alumnos desaprobados que representa el 29.6% y 38 alumnos aprobados. Para grupo horario 54, se matricularán 14 alumnos desaprobados y 36 alumnos aprobados. Para el grupo horario 55 aprobaron 39 alumnos, y el grupo horario 65 se matricularán 1 alumno y 13 alumnos aprobados, el total alumnos que se matricularán en el décimo ciclo son 50

Indicador: cantidad de alumnos por estado: de tipo repiten, Nsp, y aprobados por grupos horarios

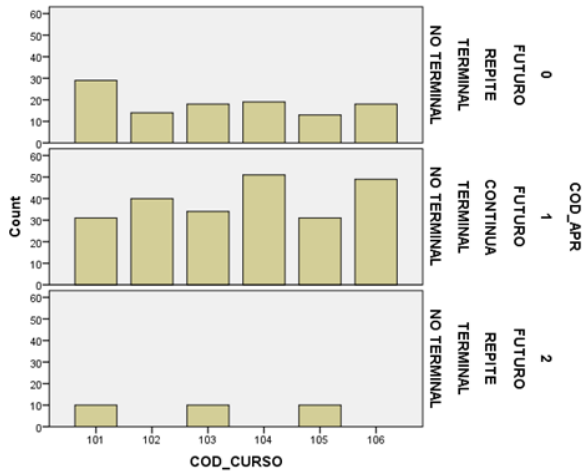

FIG.1. COMPOSICION GRUPOS HORARIOS: ALUMNOS: REPITEN-NSP-APROBADOS CICLO I

TABLA XXI
CANTIDAD DE ALUMNOS POR ESTADO CICLO I

GRUPO HORARIO	NSP	REPITE	APROBADOS
101	10	29	31
102	-	14	40
103	10	18	34
104	-	19	51
105	10	13	31
106	-	8	49
TOTAL	30	111	236

Significa que hay 30 alumnos que nunca se presentaron y llevaran el curso nuevamente, hay 111 alumnos que desaprobaron el curso que lo llevaran nuevamente y 236 que aprobaron y continúan el siguiente curso.

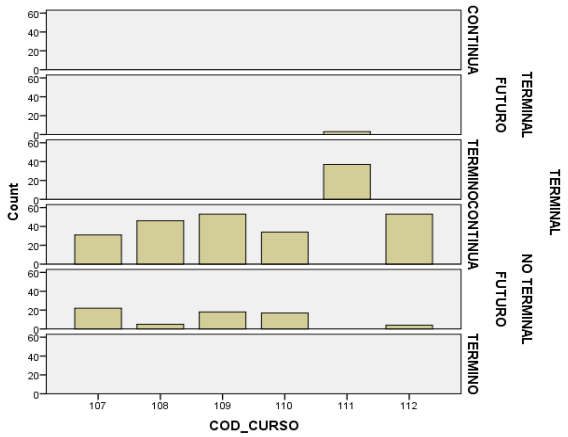

Fig. 2 Composición grupos horarios: alumnos: repiten-nsp-aprobados ciclo 2

TABLA XXII
CANTIDAD DE ALUMNOS POR ESTADO CICLO II

GRUPO HORARIO	NSP	REPITE	APROBADOS
107	7	15	31
108	2	3	46
109	17	1	53
110	6	11	54
111		-	-
112	2	2	53
TOTAL	32	34	217

Significa que hay 32 alumnos que nunca se presentaron y 34 llevaran el curso nuevamente, hay 217 alumnos que aprobaron y continúan el siguiente curso.

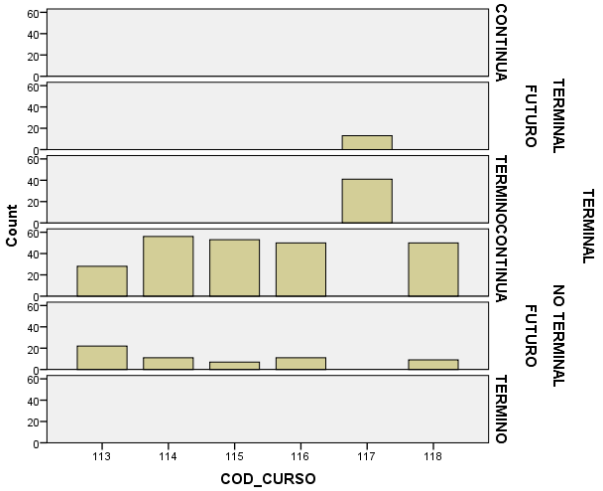

Fig. 3 Composición grupos horarios: alumnos: repiten-nsp-aprobados ciclo 3

TABLA XXIII
CANTIDAD DE ALUMNOS POR ESTADO CICLO III

GRUPO HORARIO	NSP	REPITE	APROBADOS
113	5	17	28
114	6	5	56
115	-	7	53
116	8	3	50
117	-	-	-
118	-	9	50
TOTAL	19	41	237

Significa que hay 19 alumnos que nunca se presentaron y 41 llevaran el curso nuevamente, hay 237 alumnos que aprobaron y continúan el siguiente curso.

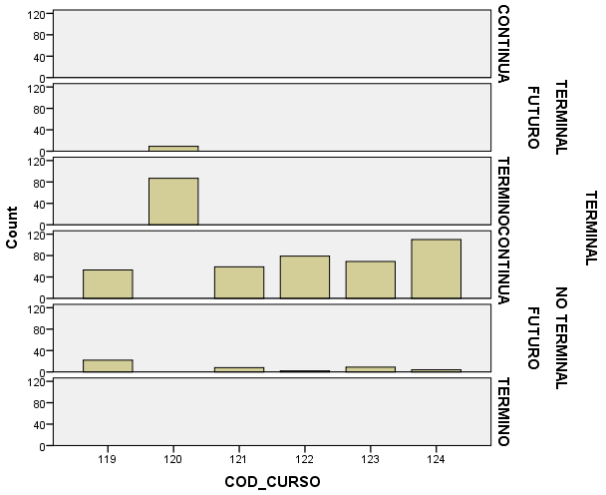

Fig. 4 Composición grupos horarios: alumnos: repiten-nsp-aprobados ciclo 4

TABLA XXIV
CANTIDAD DE ALUMNOS POR ESTADO CICLO IV

GRUPO HORARIO	NSP	REPITE	APROBADOS
119	10	12	53
120	-	-	-
121	6	2	59
122	-	2	79
123	5	4	69
124	1	3	110
TOTAL	22	23	370

Significa que hay 22 alumnos que nunca se presentaron y 23 llevaran el curso nuevamente, hay 370 alumnos que aprobaron y continúan el siguiente curso.

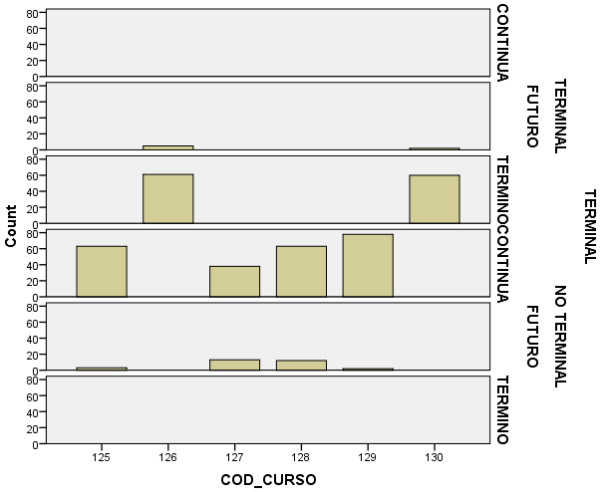

Fig. 5 Composición grupos horarios: alumnos: repiten-nsp-aprobados ciclo 5

TABLA XXV
CANTIDAD DE ALUMNOS POR ESTADO CICLO V

GRUPO HORARIO	NSP	REPITE	APROBADOS
125	2	1	63
126	-	-	-
127	-	13	38
128	8	4	63
129	-	2	78
130	-	-	-
TOTAL	10	20	242

Significa que hay 10 alumnos que nunca se presentaron y 20 llevaran el curso nuevamente, hay 242 alumnos que aprobaron y continúan el siguiente curso

IV. CONCLUSIONES

La investigación realizada determina como la aplicación de la minería de datos mejora la predicción del porcentaje de grupos horarios en la apertura de grupos horarios del ciclo verano de la Escuela de Sistemas de la Universidad Nacional del Callao.

2. Referente al primer indicador porcentaje de grupos horarios se observar se pudo predecir la cantidad de alumnos por cada grupo horario, permitiendo una mejor planificación de

3. Los cursos que tienen menor de 25 alumnos no se aperturan porque no cumplen con el punto de equilibrio, debido a que cumplen con gastos son mayores que los ingresos

4. Referente al segundo indicador se pudo determinar con más precisión el estado de los alumnos; es decir, desaprobados, nunca se presentaron, y aprobados. means “for example.” An excellent style manual for science writers is [8].

AGRADECIMIENTOS

Agradecer a la Universidad Nacional de Ingeniería por su apoyo en el Desarrollo de la investigación e investigadores por su apoyo en dicha investigación.

REFERENCIAS

[1] A. M. ABAIDULLAH, N. AHMED, AND E. ALI, “IDENTIFYING HIDDEN PATTERNS IN STUDENTS” FEEDBACK THROUGH CLUSTER ANALYSIS,” INTERNATIONAL JOURNAL OF COMPUTER THEORY AND ENGINEERING, VOL. 7, NO. 1, P. 16, 2015.

[2] M. GOYAL AND R. VOHRA, “APPLICATIONS OF DATA MINING IN HIGHER EDUCA- TION,” INTERNATIONAL JOURNAL OF COMPUTER SCIENCE ISSUES (IJCSI), VOL. 9, NO. 2, P. 113, 2012.

[3] M. A. YEHUALA, “APPLICATION OF DATA MINING TECHNIQUES FOR STUDENT SUCCESS AND FAILURE PREDICTION (THE CASE OF DEBRE MARKOS UNIVERSITY),” INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH, VOL. 4, NO. 4, PP. 91–94, 2015.

[4] K.-M. OSEI-BRYSON, "TOWARDS SUPPORTING EXPERT EVALUATION OF CLUSTERING RESULTS USING A DATA MINING PROCESS MODEL," INFORMATION SCIENCES, VOL. 180, NO. 3, PP. 414–431, 2010.

[5] F. PEDREGOSA, G. VAROQUAUX, A. GRAMFORT, V. MICHEL, B. THIRION,

O. GRISEL, M. BLONDEL, P. PRETTENHOFER, R. WEISS, V. DUBOURG ET AL., "SCIKIT-LEARN: MACHINE LEARNING IN PYTHON," THE JOURNAL OF MACHINE LEARNING RESEARCH, VOL. 12, PP. 2825–2830, 2011.

[6] H. E. LEON ATIQUIPA, "DESARROLLO DE UN MODELO ALGOR´ITMICO BASADO EN A´RBOLES DE DECISIO´N PARA LA PREDICCIO´N DE LA PERMANENCIA DE UN PACIENTE EN UN PROCESO PSICOTERAPE´UTICO," 2018.

[7] L. E. PISCOYA ORDOÑEZ, "APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA PREDECIR LA DESERCIO´N ESTUDIANTIL EN LA EDUCACIO´N BA´SICA REGULAR EN LA REGIO´N DE LAMBAYEQUE," 2017.

[8] D. J. P. MAYORCA PÉREZ, "CARACTERIZACIÓN DE LA SATISFACCIÓN DE LOS USUARIOS DEL SERVICIO DE TELEFONÍA MÓVIL EN ÁREAS URBANAS DEL PERÚ´," 2016.

[9] M. INFANTE, Y. ABREU, M. DELGADO, AND O. INFANTE, "MINER´IA TECNOLÓGICA PARA EL ANÁLISIS DE OPORTUNIDADES DE PUBLICACIONES EN LA UNIVERSIDAD," REVISTA CENIC. CIENCIAS BIOLÓGICAS, VOL. 41, 2010.

[10] C. P. LÓPEZ AND D. S. GONZÁLEZ, MINER´IA DE DATOS: TE´CNICAS Y HERRAMIENTAS. THOMSON, 2007.

[11] L. BREIMAN, "RANDOM FORESTS," MACHINE LEARNING, VOL. 45, NO. 1, PP. 5–32, 2001.

[12] A. LIAW, M. WIENER ET AL., "CLASSIFICATION AND REGRESSION BY RANDOMFOREST," R NEWS, VOL. 2, NO. 3, PP. 18–22, 2002.

[13] E. M. ABDEL-RAHMAN, F. B. AHMED, AND R. ISMAIL, "RANDOM FOREST REGRESSION AND SPECTRAL BAND SELECTION FOR ESTIMATING SUGARCANE LEAF NITROGEN CONCENTRATION USING EO-1 HYPERION HYPERSPECTRAL DATA," INTER- NATIONAL JOURNAL OF REMOTE SENSING, VOL. 34, NO. 2, PP. 712–728, 2013.