UNS
ESCUELA DE POSGRADO

MODELO PREDICTIVO BASADO EN MACHINE LEARNING COMO SOPORTE PARA EL SEGUIMIENTO ACADÉMICO DEL ESTUDIANTE UNIVERSITARIO

Tesis para optar el grado de
Doctor en Ingeniería de Sistemas e Informática

Autor:

Mag. Hugo Esteban Caselli Gismondi

Asesor:

Dr. Luis Vladimir Urrelo Huiman

NUEVO CHIMBOTE - PERÚ
2021

UNS
ESCUELA DE
POSGRADO

CONSTANCIA DE ASESORAMIENTO DE LA TESIS DOCTORAL

Yo, **Luis Vladimir Urrelo Huiman**, mediante la presente certifico mi asesoramiento de la Tesis Doctoral titulada: **MODELO PREDICTIVO BASADO EN MACHINE LEARNING COMO SOPORTE PARA EL SEGUIMIENTO ACADÉMICO DEL ESTUDIANTE UNIVERSITARIO**, elaborada por el magister **Hugo Esteban Caselli Gismondi** para obtener el Grado Académico de Doctor en **Ingeniería de Sistemas e Informática** en la Escuela de Posgrado de la Universidad Nacional del Santa.

Nuevo Chimbote, octubre del 2021

Dr. Luis Vladimir Urrelo Huiman

ASESOR

UNS
ESCUELA DE POSGRADO

CONFORMIDAD DEL JURADO EVALUADOR

MODELO PREDICTIVO BASADO EN MACHINE LEARNING COMO SOPORTE PARA EL SEGUIMIENTO ACADÉMICO DEL ESTUDIANTE UNIVERSITARIO

TESIS PARA OPTAR EL GRADO DE DOCTOR en

Ingeniería de Sistemas e Informática

Revisado y Aprobado por el Jurado Evaluador:

Dr. Juan Pablo Sánchez Chávez

PRESIDENTE

Dr. Carlos Eugenio Vega Moreno

SECRETARIO

Dr. Luis Vladimir Urrelo Huiman

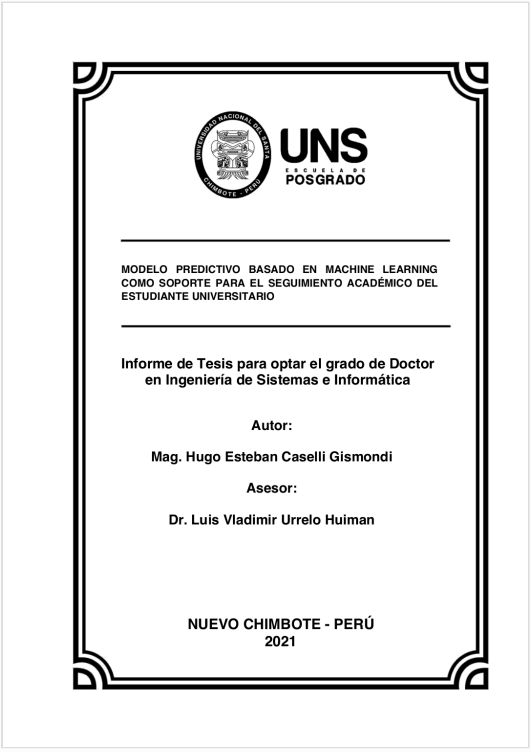
VOCAL

Recibo digital

Este recibo confirma quesu trabajo ha sido recibido por Turnitin. A continuación podrá ver la información del recibo con respecto a su entrega.

La primera página de tus entregas se muestra abajo.

Autor de la entrega:	Hugo Esteban Caselli Gismondi
Título del ejercicio:	Informes
Título de la entrega:	Informe Tesis Doctoral H. Caselli G.
Nombre del archivo:	Informe_Tesis_Doctoral_CASELLI_08.oct.2021.docx
Tamaño del archivo:	3.97M
Total páginas:	104
Total de palabras:	18,650
Total de caracteres:	106,943
Fecha de entrega:	27-oct.-2021 09:26p. m. (UTC-0500)
Identificador de la entre...	1686157238



UNS
ESCUELA DE POSGRADO

MODELO PREDICTIVO BASADO EN MACHINE LEARNING COMO SOPORTE PARA EL SEGUIMIENTO ACADÉMICO DEL ESTUDIANTE UNIVERSITARIO

Informe de Tesis para optar el grado de Doctor en Ingeniería de Sistemas e Informática

Autor:

Mag. Hugo Esteban Caselli Gismondi

Asesor:

Dr. Luis Vladimir Urrelo Huiman

NUEVO CHIMBOTE - PERÚ
2021

Dedicatoria

A mis padres que desde el cielo celebran
porque soy el fruto del esfuerzo de ambos.

A mi esposa y a mis 3 hijas
por su apoyo y comprensión
y porque nunca es tarde para
conseguir los sueños.

Agradecimientos

A mi asesor el Dr. Luis Urrelo Huiman
por conducirme por la senda de la ciencia de los datos.

A mis jurados
por sus acertadas y necesarias apreciaciones

INDICE

INDICE DE TABLAS

INDICE DE FIGURAS

RESUMEN

La educación universitaria con tanta antigüedad, aun en estos tiempos tiene el problema de gestionar el desempeño de los estudiantes de cara a obtener mejores resultados en cuanto a egresar, graduarse y/u obtener el título profesional, o incurrir en abandono de la carrera sin lograrlo, está tesis quiere contribuir en la búsqueda de una solución a través de la inteligencia artificial, machine learning y deep learning, con las limitaciones de la calidad y la cantidad de la data colectada, es por ello que se inició seleccionando los atributos más relevantes para proponer un modelo de predicción de aprendizaje profundo, se implementó un modelo inicial de red neuronal de 2 capas y se compararon con modelos alternos de 3, 4, 5, 6 y 7 capas con cantidades variables de neuronas entre ellos, los cuales fueron evaluados a través del ratio de precisión del conjunto de entrenamiento y de prueba, consiguiéndose un modelo capaz de tener una precisión de predicción de 98.97%, lo cual coadyuvará en el seguimiento eficiente a los estudiantes y poder de manera temprana orientar a los estudiantes con perfil de riesgo de abandono temporal o permanente de la carrera a conseguir sus metas, teniendo en cuenta que la variable que mayor incidencia tuvo fue el número de semestres cursado por el estudiante.

Palabras claves

Desempeño estudiantil, Aprendizaje profundo, Multiclase, Universidad

ABSTRACT

University education with such antiquity, even in these times has the problem of managing the performance of students in order to obtain better results in terms of complete studies, graduating and / or obtaining a professional degree, or incurring abandonment of the career without achieving it , this thesis wants to contribute in the search for a solution in this chain of artificial intelligence, machine learning and deep learning, with the limitations of the quality and quantity of the data collected, that is why it began by selecting the most relevant attributes To propose a deep learning prediction model, an initial 2-layer neural network model was implemented and compared with alternate models of 3, 4, 5, 6 and 7 layers with variable amounts of neurons between them, which were evaluated through the precision ratio of the training and test set, achieving a model capable of having a prediction precision of 98.97%, which will help in the efficient follow-up of students and to be able to guide students with a risk profile of temporary or permanent abandonment of the career to achieve their goals, taking into account that the variable that had the greatest incidence was the number of semesters attended by the student.

Keywords

Student performance, Deep Learning, Multiclass, University

INTRODUCCION

El bajo rendimiento del estudiante universitario, que lo conlleva en muchos casos al abandono de la carrera profesional universitaria, no consiguiendo egresar, graduarse o titularse, tratándose de universidades públicas, el gasto inmerso del estado realiza por formar profesionales se diluye, por lo que contextualmente es un tema latente, que esta tesis trata de abordar para buscar una solución de inteligencia artificial que se sume a los preceptos pedagógicos de tal manera que se pueda mejorar la gestión académica en estos aspectos.

La presente tesis, parte del contexto de la realidad genérica para llegar a una realidad específica, sobre la cual poder intervenir, es por ello que el caso de estudio específico se centra en 04 carreras profesionales de ingeniería de la Universidad Nacional del Santa. Se ha hecho la revisión del estado del arte de lo que ocurre con la gestión del desempeño y del abandono de los estudiantes universitarios, así como de las alternativas de solución que con inteligencia artificial se han dado hasta el momento, para poder presentar una alternativa novedosa de solución. Se planteó el objetivo de mejorar el seguimiento académico y nuestra propuesta es la predicción multiclase, para determinar si el alumno precisamente concluye sus estudios (egresa), se gradúa, titula o abandona. Para poder plantear un modelo de aprendizaje profundo que contemple lo anterior, se realizó la ingesta de la data que se coleccionó y con ella se alimentó el modelo inicial de red neuronal de 2 capas, se modelaron escenarios alternativos de 3, 4, 5, 6 y 7 capas con cantidad variable de neuronas, lo que permitió obtener un modelo con una precisión del 98.97% en el conjunto de entrenamiento, lo cual es satisfactorio en relación a los modelos alternativos que se trabajaron en 15 experimentos distintos y los cuales fueron contrastados con los algoritmos de clasificación obtenidos con la herramienta de experimentación AutoAI de Watson Studio de IBM quien nos devolvió el Clasificador XGB como el mejor predictor con 87,1% por debajo de lo obtenido por la red neuronal de 6 capas.

CAPÍTULO I

PROBLEMA DE INVESTIGACIÓN

1.1. Planteamiento y fundamentación del problema de investigación

1.1.1. El problema en la actualidad a nivel internacional

Sobre la base del título del presente proyecto de investigación, en primer lugar, vamos a determinar la brecha de investigación para poder conseguir una propuesta innovadora e inédita; Para la búsqueda se marcó que la población son las Universidades; La intervención que se hace sobre ellas con métodos y técnicas de predicción de acuerdo con el Machine Learning o Aprendizaje automático, enmarcado en el descubrimiento del conocimiento; Los resultados obtenidos en tanto la precisión de la predicción, éxitos de las técnicas de predicción; Y en el contexto cuál es el alcance de las predicciones en esas instituciones universitarias.

El seguimiento académico del estudiante universitario está íntimamente ligado a medir el desempeño académico a través de los índices de promoción (por asignatura hasta conseguir egresar), la desaprobación de alguna asignatura (lo que implica el tener que repetir la asignatura) y/o el nivel de deserción de los mismos (aquellos que dejan de matricularse) (Salvador Blanco & Garcia-Valcarcel Muñoz-Repiso, 1989), el Banco Mundial nos dice que la tasa de graduación de México y Perú es comparable con la de Estados Unidos que es el 65%, como se puede ver en la Figura 1.

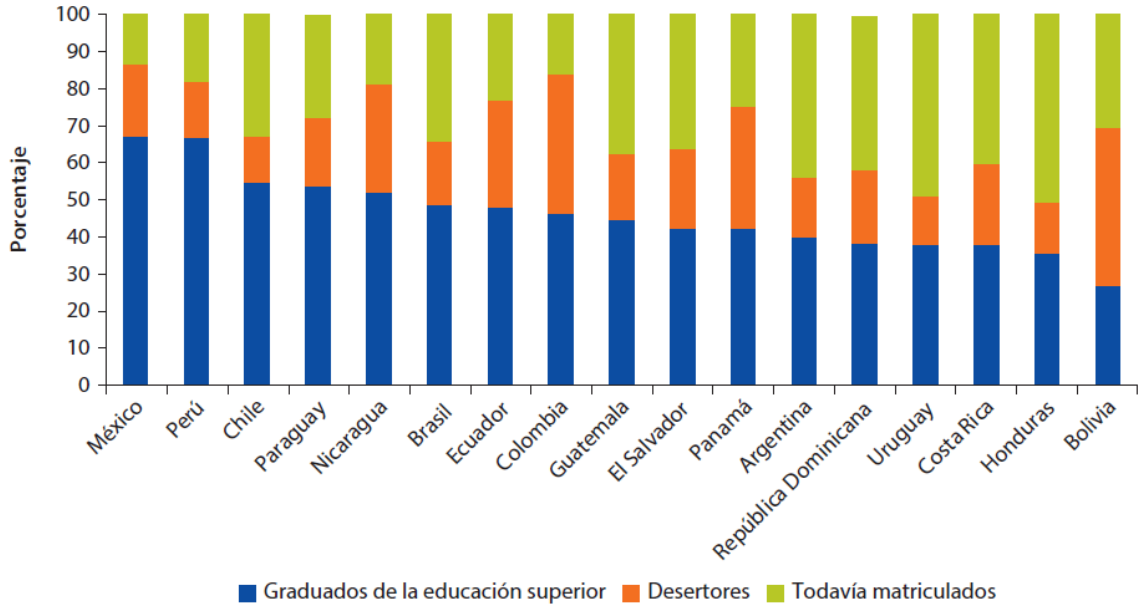


Figura 1 Tasa de graduación, jóvenes de edades 25-29 años, América Latina y el Caribe, circa 2013. Fuente: (Ferreyra, Avitabile, Álvarez, Paz, & Urzúa, 2017)

Tenemos ejemplos particulares del seguimiento de estudiantes y titulados como en la Universidad Diego Portales de Santiago de Chile (Valenzuela & Pérez, 2012), quienes han diseñado e implementado un sistema de seguimiento en 5 etapas, a través de encuestas, que se aplican en una primera etapa: en el primer año donde recogen las características de los estudiantes al ingresar a la universidad; en una segunda etapa: a mitad de carrera para obtener información de su experiencia universitaria y además incluye un registro de deserción temprana; la tercera etapa: es la encuesta de fin de carrera, donde se consigue su percepción de la evaluación de cursos, profesores, compañeros servicios y otros; Más adelante, como cuarta etapa: luego de una año de haber obtenido el título profesional, se recaba información sobre su situación laboral actual, habilidades requeridas en su trabajo, sus estudios de posgrado y expectativas de demandas de servicios de su alma mater; la quinta etapa: ocurre tres años después de haber obtenido su título profesional con una encuesta muy similar a la cuarta etapa.

Por otro lado, la (Universidad Estatal de Sonora, 2019) de México (UES), tiene un programa de apoyo y seguimiento académico el cual

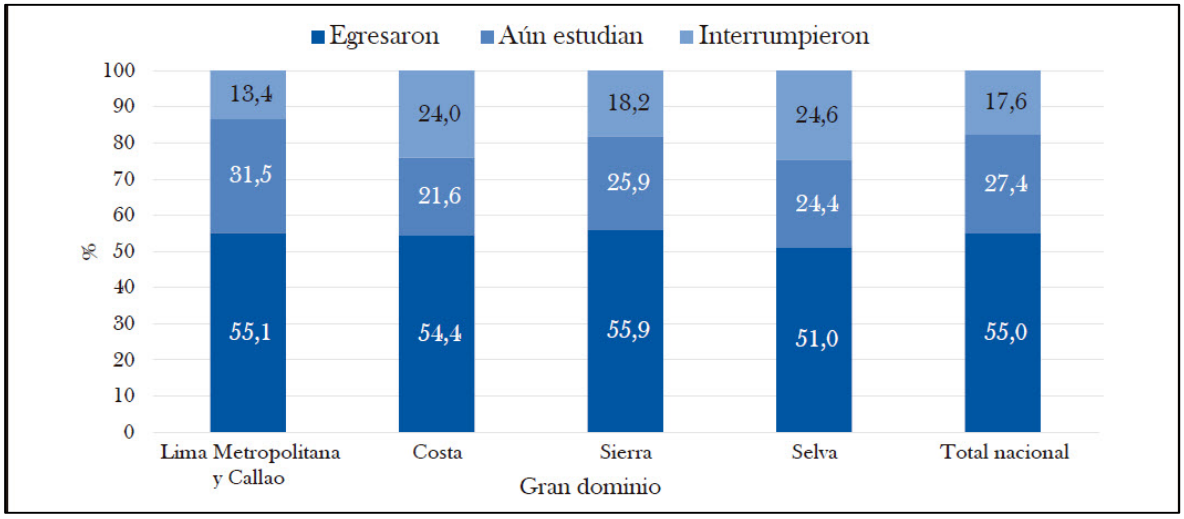
se denomina PASA, que maneja tres indicadores principales: 1° retención del primer y segundo año, 2° tasas de aprobación y 3° eficiencia terminal. Este programa tiene como objetivo dar atención y orientación académica complementaria a los alumnos de la UES, para ello el responsable del programa, debe dar seguimiento para detectar a estudiantes irregulares y aplicar las estrategias necesaria para mejorar los tres indicadores antes descritos, esto se soporta con la guía de docentes y alumnos asesores.

Por otro lado, en lo que se refiere a la consejería y tutoría a los estudiantes (Dresel & Rindermann, 2011) sostienen que la consejería desarrollada después de las evaluaciones permiten asegurar y aumentar la calidad de la enseñanza en la educación superior, que se colige con (Arco-Tirado, Fernández-Martín, & Fernández-Balboa, 2011) en tanto ellos de su estudio recomiendan que la tutoría, en particular la realizada por compañeros sea reevaluada en la educación superior, (Colvin, 2015) resumen dentro de los beneficios de la mentoría por parte de compañeros que hay más retención, mejora el desempeño académico y se consigue un nuevo amigo que es el lado afectivo del trabajo no personalizado que a lo mejor puede ocurrir en el aula.

1.1.2. El problema en la actualidad a nivel nacional

A través del "*II Informe bienal sobre la realidad universitaria en el Perú*" elaborado por la (SUNEDU II, 2020) introducen como tema nuevo, el seguimiento a los estudiantes, número y factores que se asocian a la interrupción o abandono de los estudios universitarios, entre otros datos adicionales.

Figura 2 Porcentaje de jóvenes que egresaron, interrumpieron o aun estudian

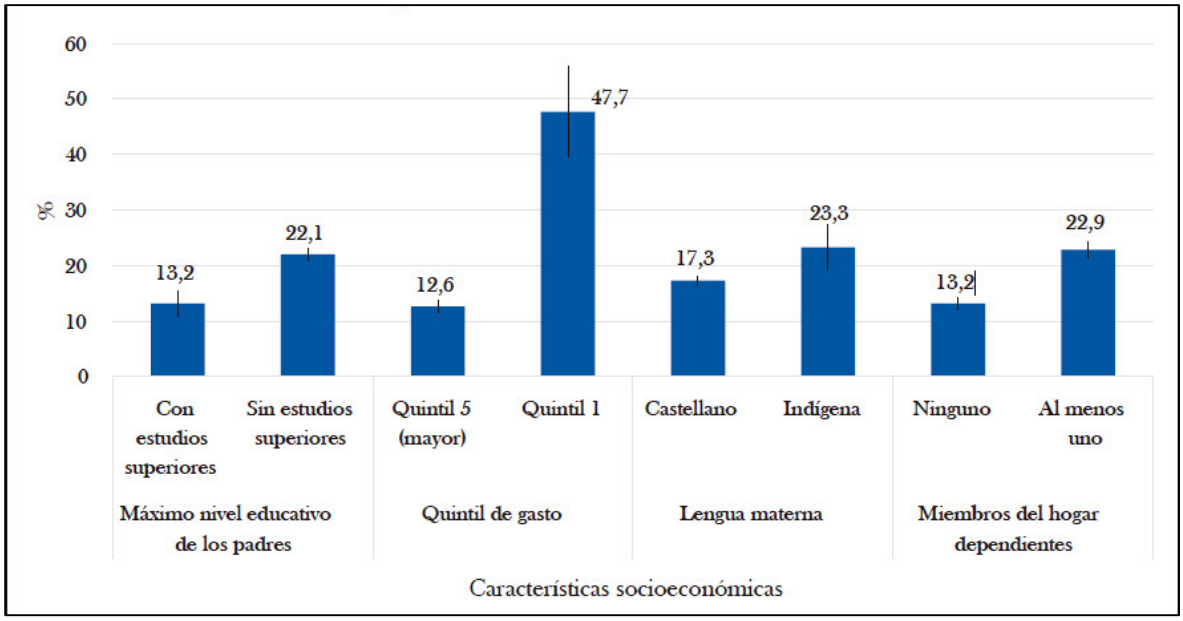


Fuente (SUNEDU II, 2020)

En la Figura 2 observamos que, a nivel nacional, el 55% de jóvenes entre 25-29 años en el año 2018 lograron egresar, el 27,4% aún estudia y el 17,6% interrumpieron sus estudios.

Por otro lado, SUNEDU también obtiene el porcentaje de jóvenes quienes interrumpieron sus estudios a través de 4 características significativas.

Figura 3 Porcentaje de jóvenes que interrumpieron estudios



Fuente (SUNEDU II, 2020)

En la Figura 3, notamos las 4 características: máximo nivel educativo de los padres, quintil de gasto, lengua materna y miembro del hogar dependientes, de las cuales la característica Quintil 1 de gasto, hogares con el nivel económico más bajo, el 47,7% de los jóvenes interrumpieron sus estudios. También es significativo que el hecho de tener padres sin estudios superiores el 22,1% hace que los jóvenes interrumpan sus estudios.

También encontramos algunos intentos de revisar el tema principalmente de las universidades privadas, por el lado de la deserción universitaria tenemos a a (Mori Sánchez, 2012) de UP de Iquitos, (Jara Tuesta, 2017) UCV Lima Norte, (Ruiz Palacios, 2018) U. señor de Sipan de Lambayeque y revisando las variables de desempeño académico de los estudiantes desde el punto de vista de una universidad pública está (Ocaña Fernández, 2011) UNSMSM-Lima.

Dato adicional el que proporciona (Mauricio Salas, 2018) quién cita que INSAN Consultores (2017) dice que la deserción universitaria en el Perú alcanza el 30% y que son causadas principalmente por no darse una buena orientación vocacional y subyacente están los problemas económicos.

1.1.3. El problema en la realidad específica local

El presente proyecto de investigación, surge de lo que acontece con respecto al seguimiento del desempeño de los estudiantes de la Universidad Nacional del Santa, el cual no está normado, y esto se refleja por la calidad de notas que se observa en los diversos boletines estadísticos que se publican al interno de nuestra universidad y que no se ve intervención ni mejoría con el paso del tiempo. Egresados nuestros ven recortadas sus posibilidades de si quiera postular a una beca de estudios de posgrado. Debido a sus calificaciones bajas como se ve en las Tabla N° 1, teniendo en cuenta que las Becas de Estudios, como por ejemplo Fulbright, entre otros solicitan “Excelentes

antecedentes académicos. Certificado de tercio superior emitido por la universidad", en el caso de Pronabec en una convocatoria de Beca de Posgrado para la República de Corea es un requisito: "Registro de notas de pregrado evidenciando un mínimo de 13 de promedio"

Durante los periodos 2011-2017, el promedio global de las asignaturas de todas las carreras universitarias de nuestra universidad estuvo alrededor de 12.72 en la escala vigesimal.

Tabla 1.

Promedios de notas de asignaturas por facultades

Año / Facultad	2011	2012	2013	2014	2015	2016	2017	Promedio de promedios
Ingeniería	11.76	11.86	12.19	12.07	12.18	12.10	12.70	12.12
Ciencias	12.31	12.32	12.67	12.76	12.94	12.63	12.96	12.65
Educación	12.86	13.27	13.54	13.46	13.64	13.51	13.45	13.39
						Promedio:		**12.72**

Aquí podemos notar que la Facultad de Ingeniería es la que posee el promedio general de notas de las asignaturas más bajo: 12.12.

Tabla 2

Promedios de notas de asignaturas escuelas-facultades

Escuela	2011	2012	2013	2014	2015	2016	2017	Promedios por Escuela
Ingeniería en Energía	11.00	11.00	12.00	11.00	12.00	11.90	12.00	**11.56**
Ingeniería Civil	12.03	12.08	12.16	11.86	11.79	11.88	12.15	**11.99**
Ingeniería de Sistemas e Informática	11.61	11.65	11.95	12.53	12.47	12.03	12.94	**12.17**
Ingeniería Agroindustrial	12.00	12.10	12.40	12.60	12.50	12.20	12.60	**12.34**
						Promedio Facultad Ingeniería:		**12.02**

En la Tabla 2, de todas las escuelas profesionales que conforman la Facultad de Ingeniería, la Escuela de Ingeniería en Energía es quien presenta el menor promedio global de asignaturas por año, entre los periodos 2011-2017 con 11.56, la que presenta el más alto promedio es la Escuela de Ingeniería Agrónoma con 12.51, que no es muy

cercano a los estándares de promedios de notas requeridos por los entes que otorgan Becas de Estudio de posgrado.

Ahora dentro de la Escuela de Ingeniería en Energía, vamos a bajar de nivel y mostraremos las asignaturas que durante el periodo de muestra tuvieron los más bajos promedios de manera constante.

Tabla 3.

Asignaturas Escuela Profesional Ingeniería en Energía con Promedios más Bajos

Curso \ Año	2011	2012	2013	2014	2015	2016	PROM
Termodinámica I	7.44	7.44	7.50	8.48	6.92	9.43	7.87
Mecánica de Fluidos	9.90	9.47	11.31	8.78	7.47	8.49	9.24
Matemática I	8.87	10.10	10.79	9.55	10.84	9.54	9.95
Control Automático	9.74	10.30	10.89	9.15	11.17	9.33	10.10
Máquinas Térmicas I	11.02	10.90	10.60	9.79	9.21	9.80	10.22

Como podemos observar en la Tabla 3 estas asignaturas, tienen un promedio global de aprobación muy por debajo de la nota 11 (en escala vigesimal), utilizada para promocionar a los alumnos.

En la Tabla 4 acompañamos las asignaturas de la Escuela Profesional de Ingeniería de Sistemas con los promedios más bajos entre los periodos 2011-2017, aun cuando tiene promedios por encima de la Escuela de Ingeniería en energía, por igual está bordeando el promedio 10 sobre la base de la escala vigesimal como se pude observar en la Tabla siguiente:

Tabla 4.

Asignaturas Escuela Profesional Ingeniería de Sistemas e Informática con Promedios más Bajos

Curso \ Año	2011	2012	2013	2014	2015	2016	2017	PROM
Física III	10.76	10.11	10.45	11.00	10.20	11.18	10.85	**10.65**
Investigación de Operaciones I	11.39	9.66	9.92	11.80	10.50	9.29	12.18	**10.68**
Física II	10.60	11.80	11.49	10.91	10.44	10.48	10.85	**10.94**
Dinamica de Sistemas I	11.35	10.70	10.85	10.65	10.90	11.39	12.3	**11.16**

A esto le sumamos los niveles de abandono que ocurren al final luego de haber culminado los estudios, en lo que respecta la obtención del Grado académico y el Título profesional. Si se tiene en cuenta que desde que inició el funcionamiento de la carrera de Ingeniería de Sistemas e Informática de la UNS, hasta el 2018 ha habido 1522 ingresantes, de los cuales solo el 32% ha conseguido obtener el grado de bachiller y el 23% el Título Profesional, lo cual se puede ver en la Tabla 5

Tabla 5

Ingresantes, Graduados y Titulados en la EPISI UNS

Año	Ingresantes	Graduados	Titulados	ACUM Ingresantes	ACUM Graduados	ACUM Titulados	Diferencia Graduados-Titulados
1991	63			63			
1992	50			113			
1993	50			163			
1994	53			216			
1995	55			271			
1996	55			326			
1997	45			371			
1998	55	3		426	3		3
1999	55	8	3	481	11	3	8
2000	55	13	3	536	24	6	18
2001	55	11	5	591	35	11	24
2002	63	17	4	654	52	15	37
2003	60	9	30	714	61	45	16
2004	50	33	20	764	94	65	29
2005	52	23	35	816	117	100	17
2006	52	23	40	868	140	140	0
2007	52	12	5	920	152	145	7
2008	52	16	15	972	168	160	8
2009	55	30	9	1027	198	169	29
2010	55	43	26	1082	241	195	46
2011	55	46	23	1137	287	218	69
2012	55	24	19	1192	311	237	74
2013	55	23	26	1247	334	263	71
2014	55	33	15	1302	367	278	89
2015	55	29	11	1357	396	289	107
2016	55	28	20	1412	424	309	115
2017	55	28	17	1467	452	326	126
2018	55	29	28	1522	481	354	127
TOTAL	**1522**	**481**	**354**				
PORCENTAJE:		**32%**	**23%**				

En la Figura 4, se nota claramente la gran brecha entre ingresantes y Graduados de la EPISI-UNS.

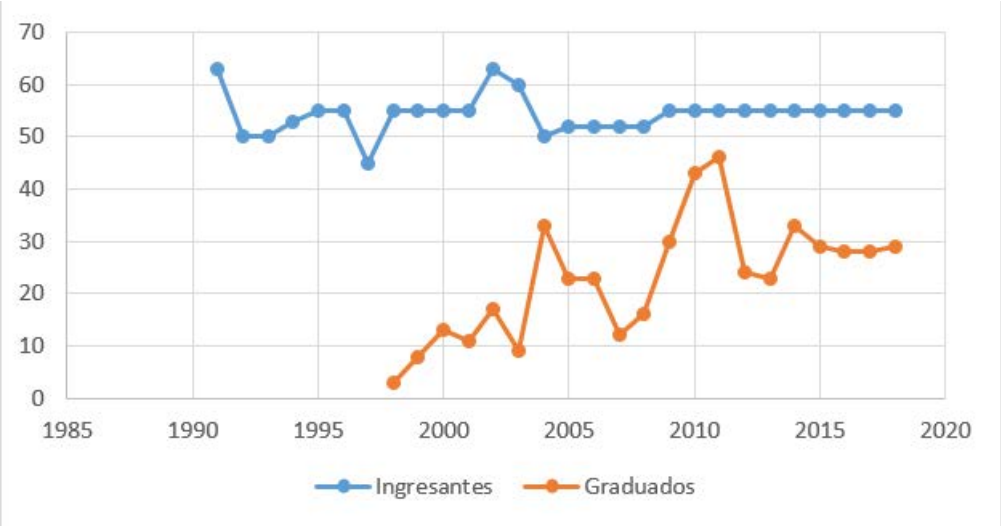


Figura 4 Ingresante vs. Graduados en la EPISI UNS
Fuente: Compendios Estadísticos UNS 1987-2018

Y en la Figura 5, en adición mostramos la brecha entre Ingresantes y Titulados de la EPISI-UNS desde que inició la carrera hasta el 2018.

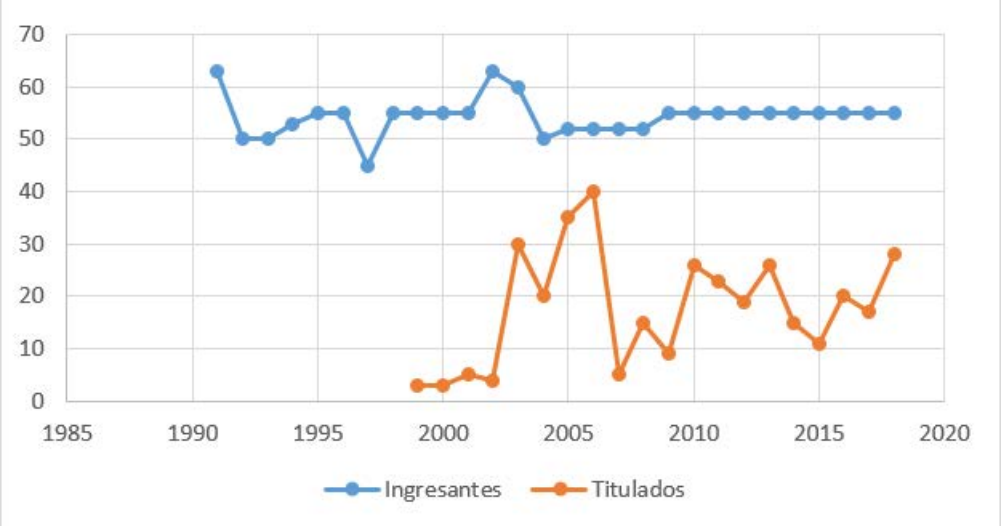


Figura 5 Ingresante vs. Titulados en la EPISI UNS
Fuente: Compendios Estadísticos UNS 1987-2018

De igual manera, podemos observar el comportamiento de las Escuelas Profesionales de: Ingeniería en Energía, Figuras 6 y 7,

Ingeniería Agroindustrial, Figuras 8 y 9, así como Ingeniería Civil, Figuras 10 y 11

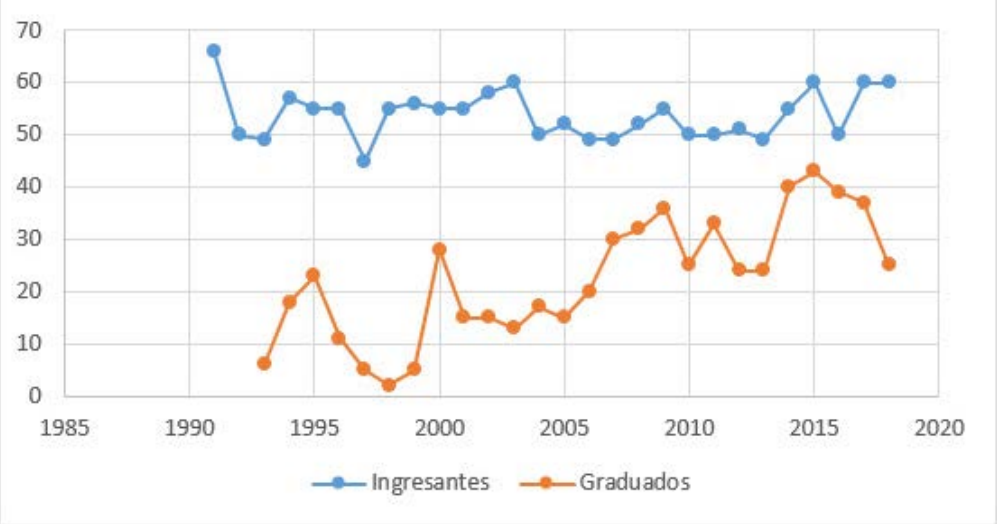


Figura 6 Ingresante vs. Graduados en la EPIE UNS
Fuente: Compendios Estadísticos UNS 1987-2018

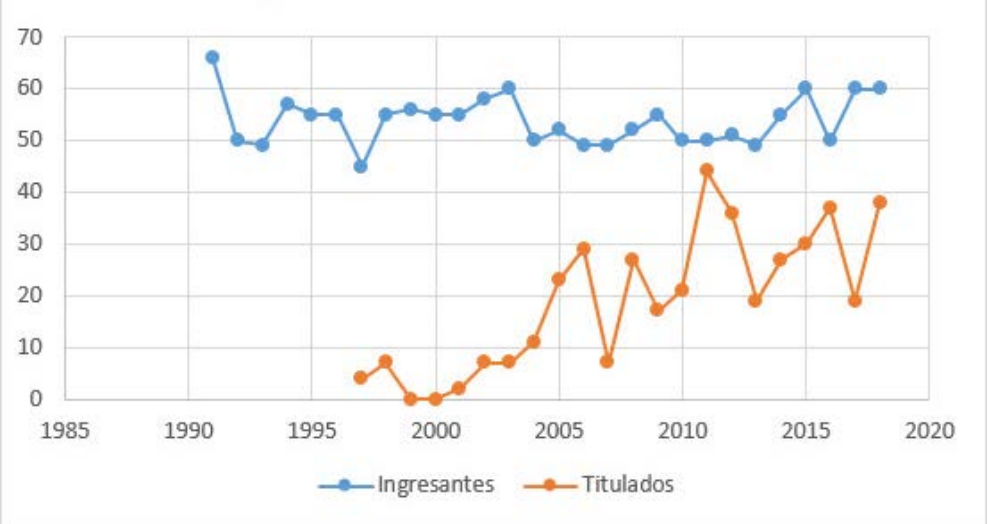


Figura 7 Ingresante vs. Titulados en la EPIE UNS
Fuente: Compendios Estadísticos UNS 1987-2018

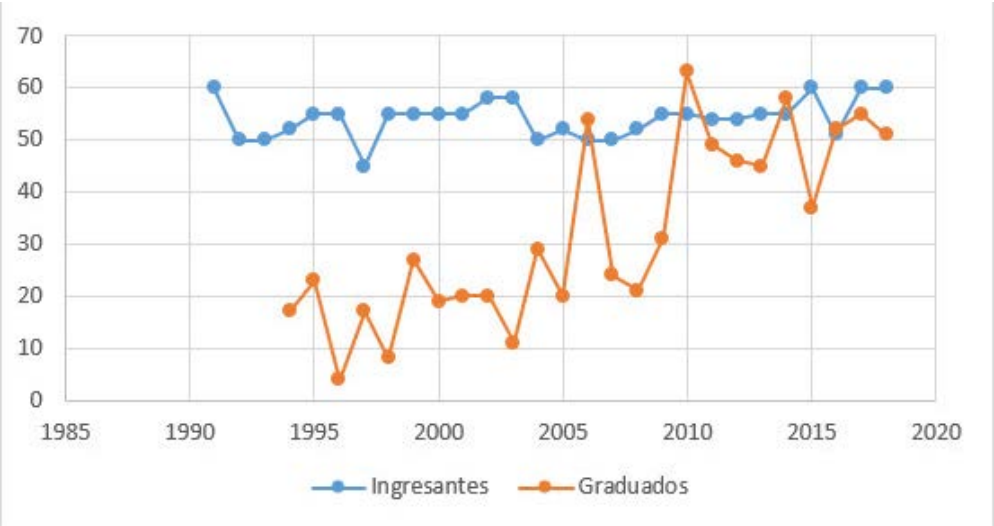


Figura 8 Ingresante vs. Graduados en la EPAI UNS
Fuente: Compendios Estadísticos UNS 1987-2018

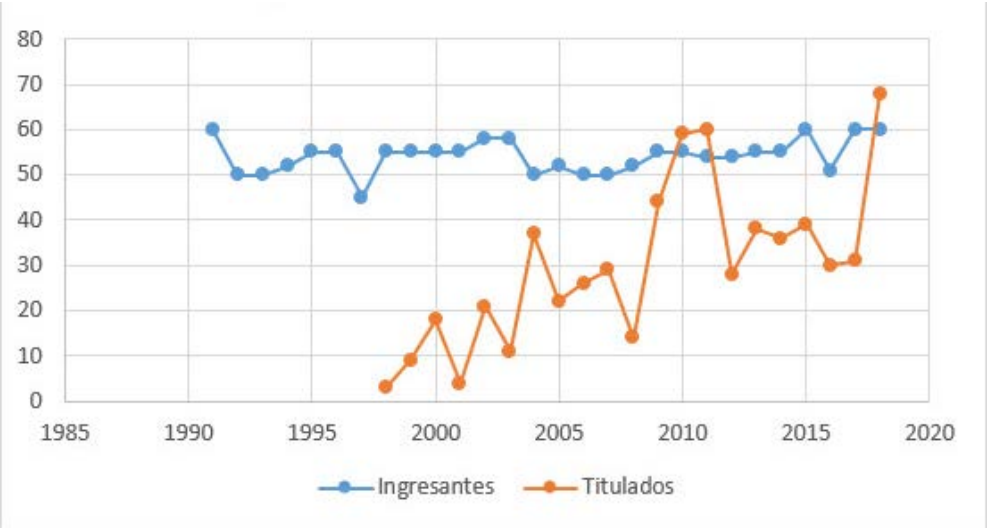


Figura 9 Ingresante vs. Titulados en la EPAI UNS
Fuente: Compendios Estadísticos UNS 1987-2018

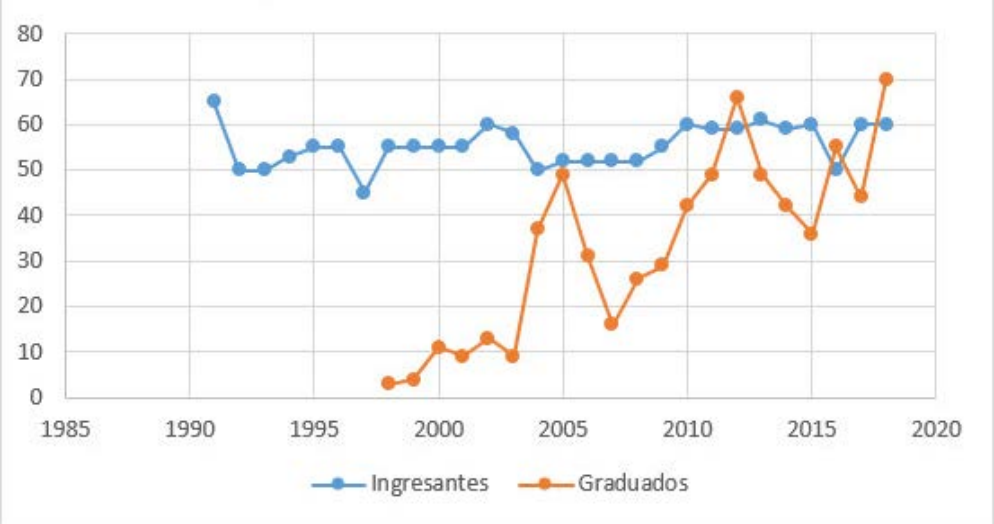


Figura 10 Ingresante vs. Graduados en la EPIC UNS
Fuente: Compendios Estadísticos UNS 1987-2018

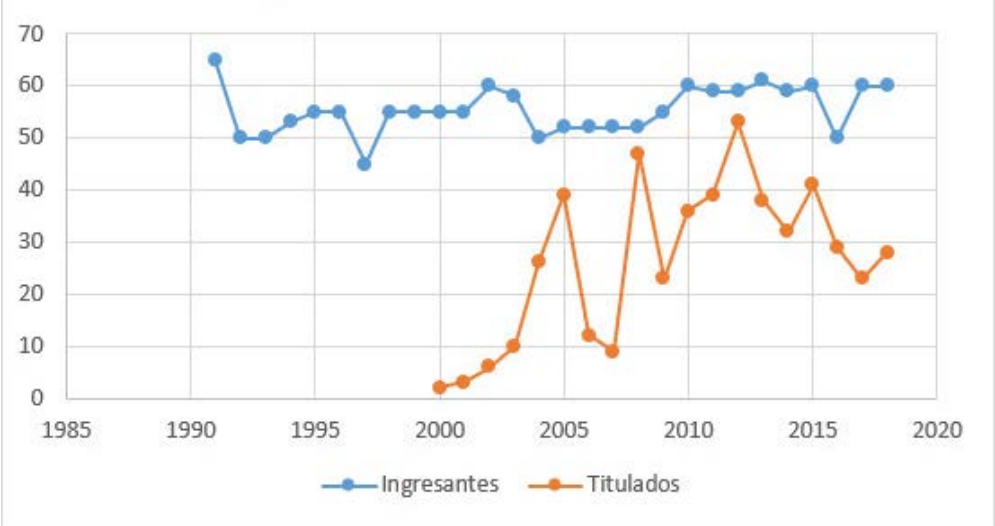


Figura 11 Ingresante vs. Titulados en la EPIA UNS
Fuente: Compendios Estadísticos UNS 1987-2018

Tabla 6 Porcentaje de Graduados y Titulados vs. Ingresantes

Escuela	Ingresantes (I)	Graduados (G)	% (G/I)	Titulados (T)	% (T/I)
Sistemas	1522	481	31.60%	354	23.26%
Energía	1508	581	38.53%	412	27.32%
Agroindustria	1516	801	52.84%	627	41.36%
Civil	1552	690	44.46%	496	31.96%

Como podemos observar en la Tabla 6, en estás cuatro Escuelas de Ingeniería, en graduados vs el número de ingresantes están por debajo del 53%, y en titulados vs el número de ingresantes por debajo del 42%, y en ambos casos la Escuela Profesional de Sistemas e Informática es la que menor proporción tiene.

Tabla 7 Porcentaje Graduados Titulados Facultades

	FACULTAD												
	Ingeniería					Ciencias			Educación				
Escuelas	Energía	Agroindustria	Civil	Sistemas	Prom. Acum. (1)	Enfermería	Acuicultura	Prom. Acum. (2)	Inicial	Primaria	Secundaria	Comunicación	Prom. Acum. (3)
Ingresantes	1710	1723	1552	1522		1460	1316		1195	1075	2966	1342	
Bachilleres	581	801	690	481		902	432		557	624	1241	490	
% Bach	33.98%	46.49%	44.46%	31.60%	39.13%	61.78%	32.83%	47.30%	46.61%	58.05%	41.84%	36.51%	45.75%
Titulados	418	630	496	354		855	305		432	531	918	210	
% Titu	24.44%	36.56%	31.96%	23.26%	29.06%	58.56%	23.18%	40.87%	36.15%	49.40%	30.95%	15.65%	33.04%

En la Tabla 7, se evidencia que la Facultad de Ingeniería, es la que tiene los promedios acumulados más bajos de la proporción de bachilleres vs ingresantes, así como titulados vs ingresantes con 39.13% y 29.06% respectivamente, con respecto a las otras 2 Facultades.

A partir de lo mostrado podemos plantear muchas interrogantes, ¿Los alumnos no responden a las exigencias de la asignatura y/o de la carrera profesional? o ¿Los docentes son demasiado exigentes para el nivel de preparación de los alumnos?, preguntas aisladas que no

permiten visualizar el contexto del porque realmente está ocurriendo estos escenarios del desempeño de los alumnos (¿o de los docentes?).

Es también pertinente indicar que, de acuerdo con el Reglamento General del Estudiante, la Universidad Nacional del Santa da el seguimiento al alumno en su desempeño académico a través de la tutoría, consejería y asesoría, siendo los docentes quienes prestan este apoyo de manera permanente (art° 201), de acuerdo con el art° 209 inciso b que a la letra dice: "*Orientar y ayudar al estudiante en el trabajo académico de enseñanza-aprendizaje e investigación para el logro de sus objetivos curricular y competencias de egreso"*.

1.2. Antecedentes de la investigación

1.2.1. Internacionales

Desde la mirada de la variable independiente de nuestra investigación, encontramos la disertación de la tesis doctoral intitulada "*Using Data Science and Predictive Analytics to Understand 4-Year University Student Churn*" de (Whitlock, 2018) quien realizó un estudio preliminar para descubrir los factores asociados al éxito de los estudiantes al conseguir graduarse, para de esta manera a través de un modelo predictivo poder identificar a nuevos estudiantes que tendrían el mismo éxito, para ello trabajo con cuatro factores de los estudiantes: características institucionales, académicas, aptitud estudiantil y comunidad de estudiantes, que aplicados a los métodos predictivos: regresión logística, arboles de decisión, bosque aleatorio, redes neuronales artificiales y máquina de vectores, siendo este último y la regresión logística quienes tenían más poder predictivo 59%. Hecho contra el cual podremos contrastar nuestros resultados.

(Musso, Hernández, & Cascallar, 2020) en su artículo "*Predicting key educational outcomes in academic trajectories: a machine-learning approach*" abordan la predicción y comprensión de los resultados en la trayectoria académica de un estudiante como la finalización del título, empleando aprendizaje automático con redes neuronales artificiales de perceptron multicapa, con un algoritmo de retropropagación, para clasificar entre otros los resultados de finalización de títulos en una muestra de 655 estudiantes de una universidad privada. Las estrategias de afrontamiento fueron los mejores predictores para la finalización de un título, consiguiendo ellos 80.7% de precisión en el conjunto de entrenamiento, y la información de antecedentes tuvo mayor peso predictivo para la identificación de los estudiantes que abandonarán o no los programas universitarios, con un 60.5% de precisión también en el conjunto de entrenamiento.

Mirando las consecuencias del bajo desempeño de los alumnos que conducen al abandono (Vijayalakshmi & Vengatachalapathy, 2019) en su artículo intitulado "*Deep Neural Network for Multi-Class Prediction of Student Performance in Educational Data*" aplica la técnica de aprendizaje profundo con un modelo de red neuronal de dos capas ocultas usando la función de activación ReLu y una capa de salida usando la función de activación softmax, después del procesamiento sobre la base de tres clases de predicción de salida del abandono como bajo, medio y alto, obtuvieron una precisión del 85% en el conjunto de entrenamiento, esto nos será útil en la medida que los estudiantes que no logran titularse u obtener el grado de bachiller están incurriendo en abandono de la carrera.

Del trabajo comparativo intitulado "*Comparison of Predicting Student's Performance using Machine Learning Algorithms*" de (Vijayalakshmi & Venkatachalapathy, 2019) que entrenaron modelo empleando diferentes algoritmos como Decision Tree (C5.0), Naïve Bayes, Random Forest, Support Vector Machine, K-Nehest Neighbor y Deep neural network en R Programming, para predecir el desempeño de los estudiantes consiguiendo la precisión más alta en el conjunto de entrenamiento de 84% con la implementación de Deep neural network, lo cual nos será de utilidad en nuestra discusión, aun cuando ellos manejan tres clases de predicción en la salida, para determinar si el rendimiento será bajo, medio o alto.

Según (Ojha, Heileman, Martinez-Ramon, & Slim, 2017) es su artículo intitulado "*Prediction of graduation delay based on student performance*" numerosos factores pueden afectar la capacidad de un estudiante para tener éxito y, en última instancia, graduarse, la preparación preuniversitaria, los servicios de apoyo al estudiante proporcionados por una universidad. Ellos analizaron el impacto de dichos factores en las tasas de graduación de una universidad utilizando tres modelos predictivos: Máquinas de Vector de Soporte

(SVMs), Procesos Gaussianos (GPs) y Máquinas de Boltzmann Profundo (DBMs). Luego de entrenar sus modelos utilizando datos reales de los estudiantes, sus resultados muestran que los DBM superan a los SVM y GP, siendo la diferencia con respecto a las precisiones obtenidas insignificantes, alcanzando los DBM 86% de precisión para decir que hay cero demoras para la graduación, demora de un año o demora de 2 o más años. Aun cuando la cualificación está dada por el tiempo, nos servirá para poder establecer un comparativo de modelo de predicción multiclase.

1.2.2. Nacionales

En relación a la variable independiente de nuestra investigación, tenemos la tesis doctoral intitulada "*Potencia Predictiva de Variables Académicas en el Rendimiento Académico de Estudiantes Universitarios del Primer Ciclo-2015-1. Caso de la Universidad Privada del Norte-Cajamarca*" de (Calua Torres, 2016) quién ha investigado si las variables académicas: rendimiento previo de la secundaria, satisfacción académica, programa de tutoría y aptitud académica tiene potencia predictiva del rendimiento académico de los estudiante universitarios de primer ciclo, el cual fue probado a través del modelo regresión múltiple, concluyendo que la variable con mayor potencia predictiva fue rendimiento previo de la secundaria, aun cuando no alcanzó los niveles de otros estudios, y las otras variables que conforman el estudio también son susceptibles de conformar un modelo predictivo con la salvedad de hacerlas más específicas, para que su contribución a la predicción sea mayor. Tendremos en cuenta la recomendación para nuestro modelo a realizar.

Asociada a la variable dependiente, tenemos la tesis doctoral intitulada "*Factores asociados a la deserción de estudiantes universitarios*" de (Castañeda Castañeda, 2013) quien trata de determinar qué factores influyen en la deserción de estudiantes universitarios de pregrado, concluyendo que los factores económicos

generan deserción temporal está en el orden del 57.4%, el mismo factor genera deserción parcial el 41.4% de las veces y el 22.9% de las ocurrencias generan deserción definitiva. Datos de referencia útiles para un contraste posterior con el modelo que propondremos.

1.2.3. Locales

Se realizó una búsqueda exhaustiva en el repositorio Alicia-CONCYTEC y no se encontró ningún antecedente local.

1.3. Formulación del problema de investigación

Teniendo en cuenta lo anterior planteamos el siguiente problema: ¿En qué medida un modelo predictivo basado en Machine Learning mejorará la gestión del seguimiento académico del estudiante universitario?

1.4. Delimitación del estudio

El alcance de esta tesis doctoral es proponer un modelo predictivo basado en machine learning para el seguimiento académico del estudiante universitario en tanto poder predecir el logro de obtener el grado de bachiller, el título profesional, simplemente egresar o mantener el estado de estudiante por abandono temporal (muchas veces permanente), el cual se validará con la data proveniente de 04 Escuelas Profesionales de Ingeniería de la Universidad Nacional del Santa: Sistemas e Informática (EPISI), Energía y Física (EPIE), Agroindustrial (EPIA) y Civil (EPIC).

1.5. Justificación e importancia de la investigación

Dada la complejidad del seguimiento académico de los estudiantes universitarios por los diversos factores que influyen en él, siendo necesario poder determinar el nivel de permanencia, promoción y/o abandono, es por ello que se ha desarrollado un modelo de predicción basado en machine learning, que sea ajustable a diversas realidades, teniendo en cuenta que no todas las instituciones educativas tienen u obtienen la data que pudiera considerarse necesaria para tal análisis. En adición nuestro estudio se enfoca de manera distinta a reconocer quienes tienen mayor probabilidad de egresar,

graduarse, titularse o en su defecto abandonar la carrera y no conseguir el diploma correspondiente.

1.6. Objetivos de la investigación

1.6.1. General

Obtener un modelo predictivo basado en Machine Learning para mejorar la gestión del seguimiento académico del estudiante universitario.

1.6.2. Específicos

1. Diseñar modelo predictivo empleando machine learning mediante algoritmos de clasificación.
2. Seleccionar las características o atributos que servirán de entradas en el modelo predictivo.
3. Implementar propuesta de modelo predictivo aplicando técnicas de machine learning para identificar a los estudiantes con alto riesgo de abandono y que no lograran culminar y obtener el diploma correspondiente.
4. Determinar variable que mayor influencia tiene en la predicción.
5. Evaluación del modelo predictivo basado en Machine Learning como soporte para el seguimiento académico del estudiante universitario en 04 Escuelas Profesionales de la Facultad de Ingeniería de la Universidad Nacional del Santa

CAPÍTULO II

MARCO TEORICO

2.1. Fundamentos teóricos de la investigación

a) Seguimiento académico

Significa medir el desempeño académico a través de los índices de promoción como se mencionó anteriormente de acuerdo con (Salvador Blanco & Garcia-Valcarcel Muñoz-Repiso, 1989) y el Banco Mundial sobre la tasa de graduación de Perú alrededor del 65% tan igual como Estados Unidos

Ante ellos las universidades, tanto europeas como americanas, vienen realizando sendos estudios en estos temas, dada la complejidad del mismo, es así que en Latinoamérica a través del CLABES, se hace incidencia en el estudio del abandono en la educación superior, el cual realizó su VIII Congreso en el año 2018 en donde se abarca de igual manera la calidad de la enseñanza en las universidades (CLABES, 2019).

b) Retención del estudiante

Daremos una mirada al seguimiento académico del estudiante universitario desde el punto de vista positivo, es decir pensando en la retención del estudiante, ya (Rossmann & Kirk, 1970) nos alcanzaban un estudio de la persistencia entre estudiantes universitarios utilizando 15 factores del Manual Omnibus Personality Inventory de la Universidad de California de 1962, en donde también como parte del seguimiento de manera paralela se monitoreaba también los retiros. Encontramos la misma preocupación en la retención del estudiante universitario en Australia (Scott, Shah, Grebennikov, & Singh, 2008) ellos identifican las medidas que contribuyeron a la retención, como los programas de orientación en el primer año, los programas de apoyo por pares, grupos de estudio y tutoría por pares, entre otros. En Chile (Donoso Díaz & Arias R., 2010) proponen estrategias diversas de soporte académico y financiero en al menos los dos primeros años; (Puchi, Moraga, & Villagran, 2016) nos dicen que son

elementos clave identificar las condiciones de entrada de los estudiantes, para de esta manera darles apoyo académico para la nivelación y adaptación y realizar el seguimiento a través de las calificaciones parciales. En Colombia (Torres Guevara, 2012) refieren que la retención estudiantil universitaria se realiza principalmente a través de programas de apoyo: financiero, académico, psicológico y gestión universitaria. En México (Hernández Herrera, 2016) propone un plan que integre 03 factores principales: el docente, la evaluación y la motivación de los estudiantes. En Paraguay (Fernández, 2018) concluye que las calificaciones no debe ser el único indicador para medir desempeño académico, y se debe poner atención en los docentes y que estos se tornen innovadores para que de esta manera los estudiantes se motiven, y mejoren su rendimiento, teniendo responsabilidad compartida del desempeño académico, tanto docente como estudiante.

c) Del abandono y/o deserciones

La otra mirada, es la mirada no positiva, es decir realizar el seguimiento del desempeño académico del estudiante para evitar los abandonos o deserciones y/o la prolongación del estudio de la carrera en más tiempo debido a las desaprobaciones. Con (Tinto, 1975) ya se trataba de entender la naturaleza del proceso del abandono, teniendo marcado el contrasentido, de que si se conocen y evitan los abandonos, pues habrá una mayor tasa de retención de los estudiantes, y distinguía el abandono principalmente el que ocurría como despido académico por bajo desempeño, y el retiro voluntario que puede ocurrir por la influencia del clima intelectual de la institución o el sistema social de los pares, el cual se verá en menor o mayor medida afectado por los objetivos y compromisos institucionales. (Abarca Rodríguez & Sánchez Vindas, 2005) de Costa Rica nos refieren diversas formas de clasificar los abandonos, por ejemplo, teniendo en cuenta el momento que ocurre puede ser durante el semestre (intra-semestral) o también entre semestres (inter-semestral). Otra sería: el abandono temporal con alto índice de retorno (parcial) o el abandono definitivo sin retorno (total). Se suman también la deserción institucional el

cual sería el abandono permanente de una universidad y la deserción del sistema que es el retiro de todo el sistema universitario.

d) Categorías o variables, métodos y temas de predicción más empleados, que inciden en la retención y/o abandono universitario

Luego de entender que es retención y abandono, notamos que ellos han sido configurados en diversos modelos de estudio que explican el comportamiento de grupos determinados de estudiantes que están inmersos en el sistema universitario y estos modelos incluyen dentro de sí, las categorías o variables que pueden explicar la ocurrencia de cada caso.

Tal es el caso del modelo de (Spady, 1970) quien basa su modelo en la teoría del suicidio de Durkheim señalando los atributos sociales y académicos que los estudiantes presentan en el proceso de abandono de la carrera. Siguiendo la línea psicológica (Fishbein & Ajzen, 1975) se enfoca en ambos casos de retención y abandono académico los cuales se explican a través de las creencias y actitudes del estudiante que denotan la intención para inclinarse a un determinado comportamiento. El modelo de (Tinto, 1975) apunta también a explicar la dualidad persistencia/abandono a través de las interacciones entre los sistemas sociales y académicos de los individuos quienes están continuamente modificando sus objetivos y compromisos institucionales. Avanzando en el tiempo tenemos el modelo de (Bean, 1985) quién relaciona factores académicos, psicosociales, ambientales y de socialización que indicarían la intención de permanecer o abandonar, lo cual el denomino el Síndrome de deserción. Otro modelo psicológico de mirada positiva hacia la persistencia del estudiante es el de (Ethington, 1990) quien parte del apoyo que la familia otorga, donde la autoevaluación y la percepción de dificultad académica del estudiante reflejará sus expectativas de éxito hacia la persistencia o permanencia en la carrera. En adición de acuerdo con cita de (Donoso & Schiefelbein, 2007) y (Torres Guevara, 2012) el modelo de Pascarella y Terenzini de 1985 es un modelo causal donde convergen las características institucionales y ambientales, donde el desarrollo del estudiante se suman sus

características personales: aptitudes, rendimientos, personalidad, aspiraciones y etnicidad; también mencionan que el Modelo de Weidman de 1989 basado en factores sicológicos y socio estructurales, se suma a Tinto y Pascarella en tanto las variables personales de los estudiantes aptitudes, intereses de estudio, aspiraciones, valores entre otros, que también involucra a los padres.

Luego de revisar los modelos de (Spady, 1970), (Fishbein & Ajzen, 1975), (Tinto, 1975), (Bean, 1985), (Ethington, 1990), en la Figura 10, podemos resaltar las variables o categorías que tiene alta incidencia tanto para la retención como para el abandono estudiantil universitario:

1970	1975	1975	1985	1990
Spady	**Fishbein y Ajzen**	**Tinto**	**Bean**	**Ethington**
Antecedentes familiares	Creencias	Antecedentes familiares	Desempeño académico	Antecedentes familiares
Potencial académico	Actitudes	Destrezas y habilidades	Integración académica	Rendimiento académico previo
Congruencia normativa		Rendimiento académico previo	Expectativas de éxito	Apoyo familiar
Desempeño académico		Expectativas de éxito	Interacción con pares	Autoconcepto académico
Desarrollo intelectual		Compromiso institucional	Interacciones con profesores	Percepción dificultad estudios
Apoyo de pares		Desempeño académico	Financiamiento	Nivel de aspiraciones
Integración social		Interacciones con profesores	Integración social	Valores
Satisfacción		Actividades extracurriculares	Compromiso institucional	Expectativas de éxito
Compromiso institucional		Integración social		
Persistencia - Decisión de abandonar				

Figura 12 Categorías o variables que inciden en la persistencia o abandono estudiantil por autores

Los modelos antes descritos, entre otros, son empleados de base para elaborar la Guía de Gestión de permanencia estudiantil de educación superior (Ministerio de Educación Nacional Colombia, 2015), donde son agrupados en un primer momento como factores, visto desde las áreas de conocimiento que influyen en la decisión del estudiante para su permanencia o abandono (deserción), que se resumen en la Figura 13.

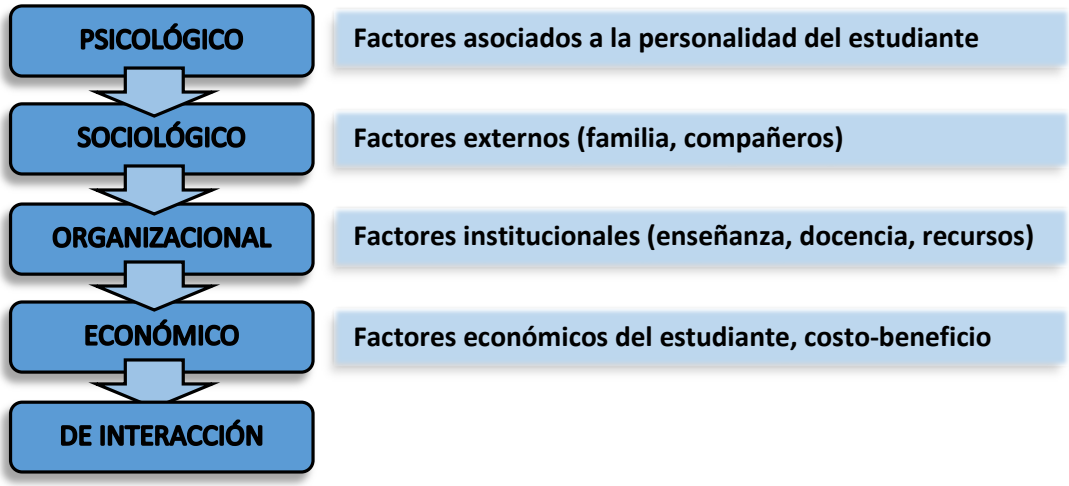


Figura 13 Factores asociados a la permanencia/abandono de los estudios superiores. Fuente: Adaptado de (Ministerio de Educación Nacional Colombia, 2015).

Figura 14 Variables que influyen en la permanencia o abandono de los estudiantes de educación superior Fuente: Adaptado de (Castaño, Gallón, Gómez, & Vásquez, 2004).

En la Figura 14, tenemos un listado de variables que influyen en la decisión de permanencia o abandono de los estudiantes de educación superior (Castaño et al., 2004), el cual es también tomado como referencia por (Ministerio de Educación Nacional Colombia, 2015) de Colombia para la elaboración de su Guía de Gestión de permanencia estudiantil, y la cual no será de utilidad al momento de seleccionar las más adecuadas para la

implementación de nuestro modelo de predicción para el seguimiento académico de los estudiantes.

En la revisión de (Shahiri, Husain, & Rashid, 2015) sobre la base de 30 estudios de investigación, determinan que los tres atributos mayormente utilizados para predecir el desempeño de los estudiantes son: las actividades extracurriculares, los antecedentes de secundaria y la red de interacción social. Asimismo, el modelo predictivo es el más utilizado para predecir el rendimiento de los estudiantes. La tarea más popular para predecir el rendimiento de los estudiantes es la clasificación. Los algoritmos de Clasificación que más se han utilizado para predecir el desempeño de los estudiantes son: El árbol de decisión, Redes neuronales artificiales, Bayes Naive, K-Vecino más cercano y Máquina de Vector de Apoyo. Esta revisión solo abarca los trabajos hasta el punto de predicción, más no inciden en el tratamiento de la mejora del desempeño estudiantil.

Enseguida realizaremos la revisión de la predicción del desempeño académico de estudiantes universitarios desde la minería de datos educativa (EDM), pues tiene mucho más tiempo de uso y sus aplicaciones han sido muy variadas. En el artículo de (Nandeshwar, Menzies, & Nelson, 2011) sobre la revisión de 14 autores significativos en el área, en el periodo desde los inicios allá por el año 1971 con (Spady, 1970) hasta el 2008 con (Pittman, 2008), entre ellos utilizaron 15 técnicas de predicción, que va desde regresión múltiple hasta redes neuronales, siendo la técnica más utilizada regresión logística con 57%, siguiéndole con 21% cada uno, la regresión múltiple, análisis discriminante y redes neuronales, las otras técnicas eran no significativas para esta muestra, como se puede ver en la Tabla 8.

Tabla 8

Técnicas de predicción utilizadas (1971-2008)

N°	Técnicas Utilizadas	Número	%
1	Logistic regression	8	57%
2	Multiple regression	3	21%
3	Discriminate analyzes	3	21%
4	Neural network	3	21%
5	C4.5	2	14%

Nota. Adaptado de (Nandeshwar et al., 2011)

Más adelante en el tiempo tenemos a (Hellas et al., 2018) quienes revisaron 357 artículos relacionados con la predicción del desempeño académico, publicados durante el periodo comprendido desde el 2010 hasta el 2018, siendo los temas más abordados con 21,60% la predicción, clasificación y la minería de datos educativa. Estando el tema de retención y estudiantes en riesgo en un tercer lugar con 16.70% ver Tabla 9, siendo un indicativo favorable para abordar el tema desde el punto de vista del Machine Learning.

Tabla 9

Temas de modelado

Tema	%
Predicción, clasificación y minería de datos educativa. Clasificación y precisión.	21.60%
Modelado de comportamiento y grado de predicción. Puntuaciones, exámenes y tareas.	20.80%
Predicción de grados, puntajes y éxito. Retención y estudiantes en riesgo.	16.70%
Modelado de datos, enfoques de cálculo, algoritmos y entrenamiento.	13.70%
Actividad en linea, tiempo y rendimiento. Factores sociales y motivación.	12.30%
Educación STEM, autoeficacia, factores de persistencia, motivación y género.	9.90%

Nota. Adaptado de (Hellas et al., 2018)

También nos indican que la variable de predicción: calificación de las asignaturas fue la más utilizadas con 24.4% y el número de asignaturas aprobadas / desaprobadas fue el menos utilizado con 1.10% ver Tabla 10.

Tabla 10

Variables de predicción más empleadas

Etiqueta	Cantidad	%
Curso Calificación o puntaje	88	24.40%
Examen/Post-prueba Grado o puntaje	53	14.70%
Rango de calificación del curso	49	13.60%
Programa o Módulo Graduación/Retención	48	13.40%
Desempeño impreciso o impreciso	44	12.20%
Rango de GPA o GPA (incluidos CGPA, SGPA)	44	12.20%
Rendimiento de la tarea	41	11.40%
Retención/abandono del curso	20	5.50%
Ganancia de conocimiento	8	2.20%
Número de cursos aprobados o reprobados	4	1.10%

Nota. Adaptado de (Hellas et al., 2018)
GPA=promedio de calificaciones

(Hellas et al., 2018) nos muestra también las características predictoras investigadas más frecuentes en relación a los valores pronosticados que son los que se muestran a la derecha en la Figura 15, para el caso del pronóstico de retención/abandono, sus variables predictoras más comunes son: el desempeño en las pre-asignaturas, en las asignaturas y la escuela secundaria, así como los datos demográficos, genero, edad, familia y los datos de personalidad, que nos servirán de guía para el modelo que propondremos.

Tabla 11

Métodos de predicción más empleados (2010-2018)

Método de predicción	Total	%
Statistical : Linear Modeling	110	17.71%
Classification: Probabilistic Graphical Model	80	12.88%
Classification: Decision Trees	74	11.92%
Statistical : Correlation	57	9.18%
Classification: Neural Network	51	8.21%
Classification: SVM	45	7.25%
Classification: Classification	42	6.76%

Nota. Adaptado de (Hellas et al., 2018)

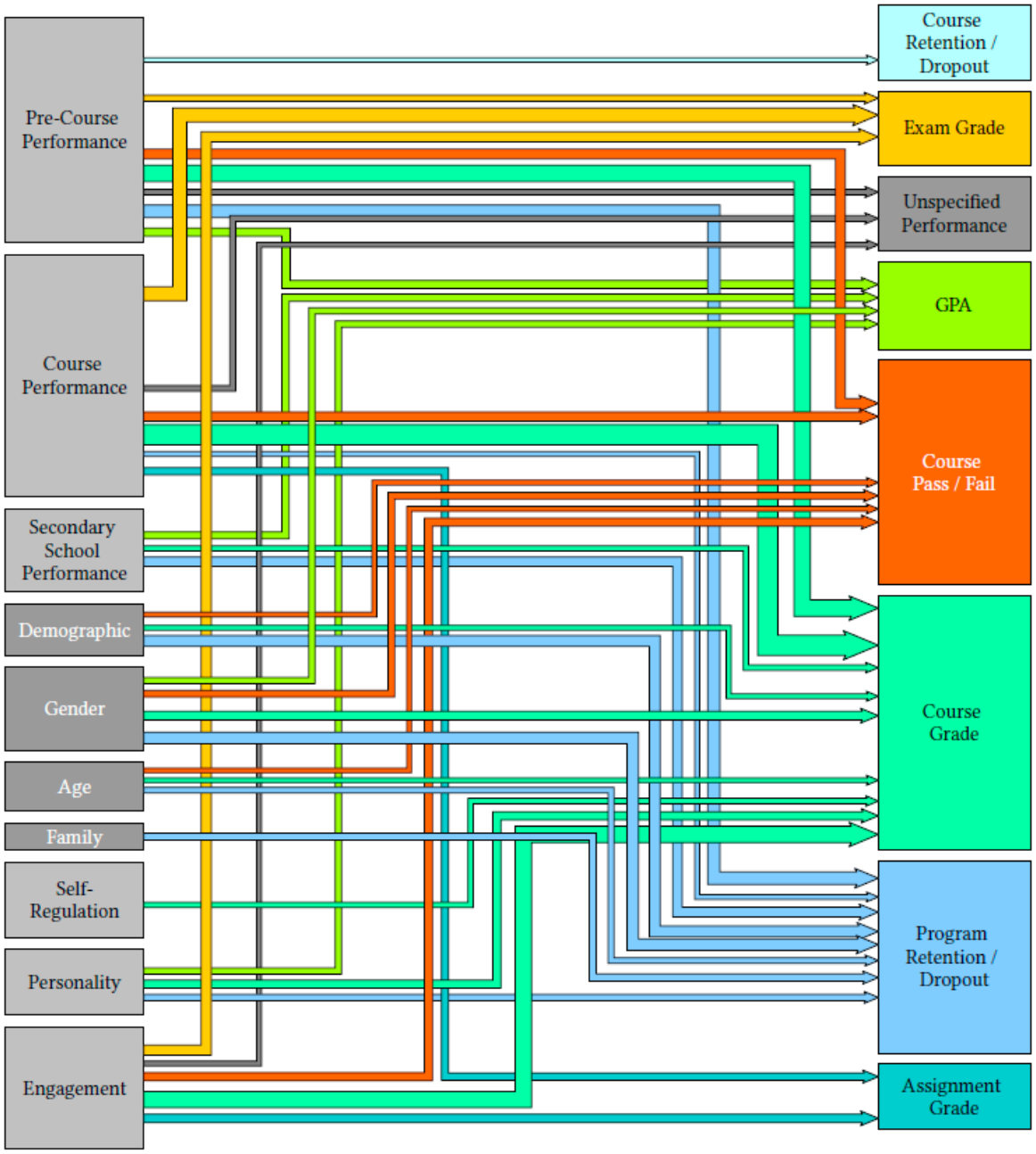


Figura 15. Características más frecuentemente investigadas como predictores (lado izquierdo) para los valores pronosticados (lado derecho). Tomado de (Hellas et al., 2018)

Mirando la Tabla 11 de Hellas y la Tabla 8 de Nandeshwar, vemos que las técnicas o métodos de predicción han ido cambiando su uso y/o aceptación, y en ambos se mantiene que las redes neuronales se mantienen a lo largo del tiempo, aun cuando no es el más utilizado, pero su uso es significativo y prolongado.

Ahora revisaremos que categorías o variables fueron más utilizadas y los temas más abordados, pero a nivel de machine Learning.

Según (Kučak, Juricic, & Đambić, 2018) la aplicación de Machine Learning en educación de acuerdo con sus 77 casos revisados, se distribuyen mayormente en la predicción del desempeño del estudiante con 54.55%, le sigue en aplicación respecto a la retención de estudiantes con 22.08%, en menor escala se aborda la graduación de estudiantes con 15.58% y al final se le emplea para la evaluación de los estudiantes con 7.79% de los casos estudiados. Notamos que es bajo el abordaje del tema de graduación de los estudiantes, que es el campo donde queremos incursionar con el presente trabajo.

(González Nespereira, Elhariri, El-Bendary, Fernández Vilas, & Díaz Redondo, 2016) nos presentan el aprendizaje automático con enfoque basado en algoritmos de clasificación de máquinas (SVM) y bosques aleatorios (RF) como modelos de predicción con el fin de descubrir la relación entre los estudiantes de cursos pasados y el Sistema de Administración del Aprendizaje (LMS) y sus tendencias a pasar o fallar, detectando que el RF superaba al SVM.

Para (Kostopoulos, Lipitakis, Kotsiantis, & Gravvanis, 2017) la predicción del rendimiento académico de los estudiantes universitarios es un problema de investigación importante para el aprendizaje automático, en él se incluye la minería de datos educativa para poder analizar el comportamiento académico de los alumnos y ellos lo realizan a través de métodos de clasificación supervisados, asociados con metodologías de aprendizaje activo demostrando la eficacia para el juego de datos etiquetados y no etiquetados que utilizaron.

En el análisis comparativo de evaluación de Algoritmos de Clasificación utilizados en la predicción del rendimiento de los estudiantes realizado por (Anuradha & T, 2015) entre las técnicas como árboles de decisión, K-Vecino más cercano, Bayes Naive y reglas de aprendizaje; sobre la base de un conjunto de atributos seleccionados, revelan que las tasas de

predicción entre ellos no son uniformes y varían entre 61 y 75% para las 5 clases definidas (entre fallar y destacar). Este análisis abarca la determinación de las variables, el recojo y tratamiento de la data, tratando de encontrar el algoritmo de clasificación con mejor tasa de predicción, que apuntan a la mejora del desempeño estudiantil, pero nada indica que se dio tratamiento para lograr mejorar el desempeño.

En cuanto a las categorías o variables más empleadas en los últimos años es muy heterogénea, así tenemos, el género del estudiante, promedio de secundaría, el programa, plan de estudio, como los de mayor coincidencia o más generales (Oancea, Dragoescu, & Ciucu, 2013), (Jia & Mareboyana, 2014), (Bendangnuksung & Prabu, 2018), (Forero Zea, Piñeros Reina, & Rodríguez Molano, 2019); diferenciándose en adición edad del estudiante, diferencia años entre secundaria-universidad de (Oancea et al., 2013); financiamiento, retención (Jia & Mareboyana, 2014); estrato socio económico, temas aprobados, temas no aprobados (Forero Zea et al., 2019).

e) La retención y abandono en Latinoamérica y el Perú

Más cerca en el tiempo, vemos que, ante el problema no resuelto, muy a pesar de los sendos modelos de estudio de la retención y abandono, y dado los cambios tecnológicos de los últimos tiempos, en una mirada a Latinoamérica es muy poco lo que encontramos, por ejemplo en Colombia (Torres Guevara, 2012) nos refieren que han implementado el sistema de información Spadies que es una herramienta para hacer seguimiento a las cifras de deserción de estudiantes de educación superior, que luego de aplicar acciones para combatir la deserción estudiantil en los factores: individuales, académicos, socioeconómicos e institucionales han logrado disminuir dicha deserción universitaria hasta un 9%. Posteriormente también en Colombia siempre vez bajo el patrocinio del (Ministerio de Educación Nacional Colombia, 2015) presentan su Guía para la implementación del Modelo de gestión de permanencia y graduación estudiantil en instituciones de educación superior, en el cual proponen 8

ejes para la gestión de permanencia y graduación estudiantil: compromiso del núcleo familiar, trabajo colaborativo, posicionamiento y formalización, cultura de la información, mejoramiento de la calidad, Trabajo conjunto con las Instituciones de la educación media (IEM), programas de apoyo y gestión de recursos. En Chile por su parte, (Paredes Esparza, Aguirre Larrain, & Quense Abarzúa, 2017) presentan el modelo de retención de la Universidad Andrés Bello, el cual tiene 4 dimensiones: diagnóstico, desarrollo de habilidades de aprendizaje, apoyos académicos extracurriculares y, acompañamiento y apoyo integral, a los cuales de manera transversal se le tiene que realizar seguimiento, evaluación y replicabilidad, reto que están siguiendo para su implementación total.

En el Perú, no existen cifras oficiales de los porcentajes de retención/abandono de los estudiantes universitarios, cuando uno revisa el informe bienal de (SUNEDU I, 2018) solo se encuentra la propuesta de un indicador que será contemplado en el Sistema de Información de Educación Superior, el cual se denomina "*deserción de estudiantes por cohorte"*, y en su Segundo Informe Bienal (SUNEDU II, 2020) solo hay dos referencias bibliográficas sobre el tema, pero sin ser abordado.

f) Seguimiento académico en la Universidad Nacional de Santa

Para nuestro caso de estudio, en la Universidad Nacional del Santa, tampoco se aborda el tema de la retención/deserción estudiantil, dado que seguimiento del desempeño de estudiante está circunscrito al ámbito de cada asignatura en particular. Si bien es cierto contamos con boletines estadísticos, entonces el desempeño académico se reduce a solo cifras estadísticas sobre las cuales no hay una acción definida al respecto. Sabemos cuántos ingresan y cuantos se titulan, pero no hay mayor análisis, ni acciones que puedan apuntar a reforzar la retención y las acciones que eviten y/o disminuya la deserción en cualquiera de sus formas.

Debemos tener en cuenta el Reglamento de la Actividad Docente (Vicerrectorado Académico, 2017) cuyo objetivo es regular el seguimiento

y supervisión del desarrollo de las actividades lectivas y no lectivas de los docentes, entre las no lectivas está la Tutoría y Consejería que es obligatoria para todos los alumnos (art°51), donde el docente tutor orienta al estudiante en sus actividades académicas y de haber problemas los deriva dependiendo del caso al especialista correspondiente, hecho del cual no hay históricos de las incidencias, de haber existido, luego que el docente emite su informe mensual este es derivado a la Dirección de Departamento (art° 53) y con el art°89 el Presidente de la Comisión de Tutoría y Consejería supervisa y evalúa esta labor docente quien eleva informe periódico al Director de Escuela y este a su vez lo remite al Director del Departamento, no hay mayor indicación que hacer con dichos informes, solo sirve para acusar cumplimiento.

Así mismo, a través del (Vicerrectorado Académico, Transparencia - Reglamento general de grados y títulos, 2017), para la etapa de titulación es el asesor, el encargado del seguimiento de la tesis hasta su culminación (art° 47), a través de un cronograma acordado con el Tesista (art° 49), donde el Tesista está obligado a asistir a las reuniones de asesoramiento (art° 50), ante el 30% de faltas el asesorado pierde su derecho al asesoramiento (art° 51). Son pocas las escuelas de la UNS que han implementado este asesoramiento con estricto cumplimiento del cronograma, es por ello que, en la Escuela Profesional de Sistemas e Informática, solamente se cuenta con un 23% de titulados con respecto al volumen de ingresantes.

g) Machine Learning

En español Aprendizaje Automático, (Mitchell, 1997) lo define como “un programa de computadora aprende de la experiencia E con respecto a alguna clase de tareas T y la medida de desempeño P, si su desempeño en tareas en T, medido por P, mejora con la experiencia E”, el aprendizaje automático es una forma de Inteligencia Artificial (IA) que permite que un sistema aprenda de datos en lugar de a través de programación explícita (Hurwitz & Kirsch, 2018), machine learning como parte de la ciencia de los datos (data science) y a su vez de la inteligencia artificial, es un campo

multidisciplinario el cual se vincula con las matemáticas, las estadísticas, ciencia de las computadoras (computer science), minería de datos (data mining), procesamiento en lenguaje natural (natural lenguaje processing) y el aprendizaje profundo (deep learning) (Sarkar, Bali, & Sharma, 2018), como se ve en la Figura 16.

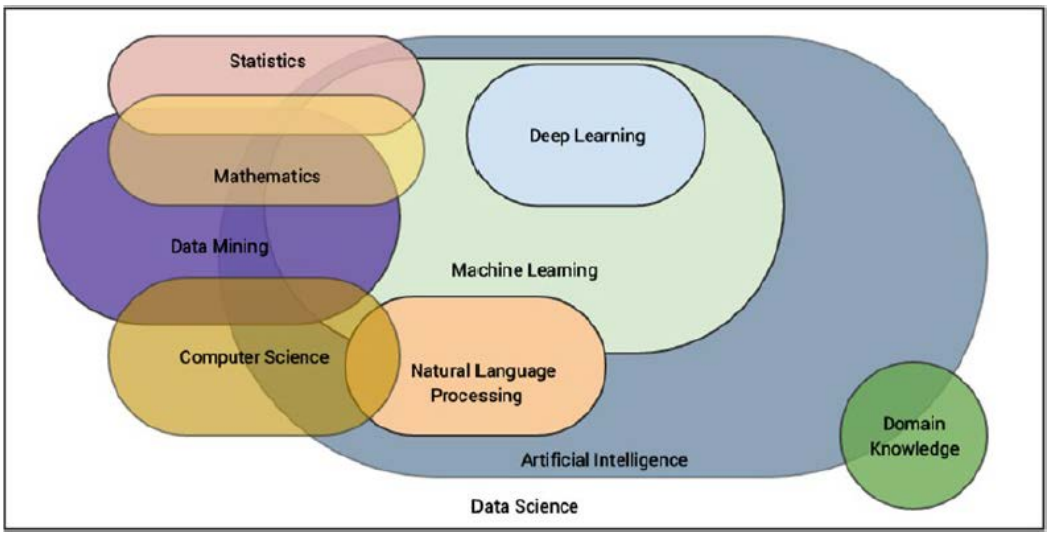


Figura 16 Machine Learning un campo multidisciplinario
Fuente: (Sarkar, Bali, & Sharma, 2018)

Tipos de Machine Learning

Entre los tipos de algoritmos de Machine Learning, Figura 15, tenemos:

- Aprendizaje supervisado, el cual contiene una variable de respuesta (o etiqueta), y esta puede ser continua o categórica, aquí el algoritmo aprende de la variable de respuesta contra el conjunto de variables predictoras.
- Aprendizaje no supervisado, cuando no hay variable de respuesta, por lo que el aprendizaje se da sobre la base de alguna medida de similitud o distancia entre cada fila en el conjunto de datos.
- Aprendizaje semi-supervisado, es una mezcla de las dos anteriores, donde hay o no hay variables de respuesta para todas las observaciones, y aun cuando estos sean desconocidos, los datos contienen información importante para el grupo.
- Aprendizaje de refuerzo, dada la necesidad de contar con datos limpios y precisos para obtener buenos resultados, y para evitar falsos

resultados, es por ellos la necesidad del refuerzo de tal manera que este aprenda continuamente del entorno en forma iterativa. (Ramasubramanian & Singh, 2017).

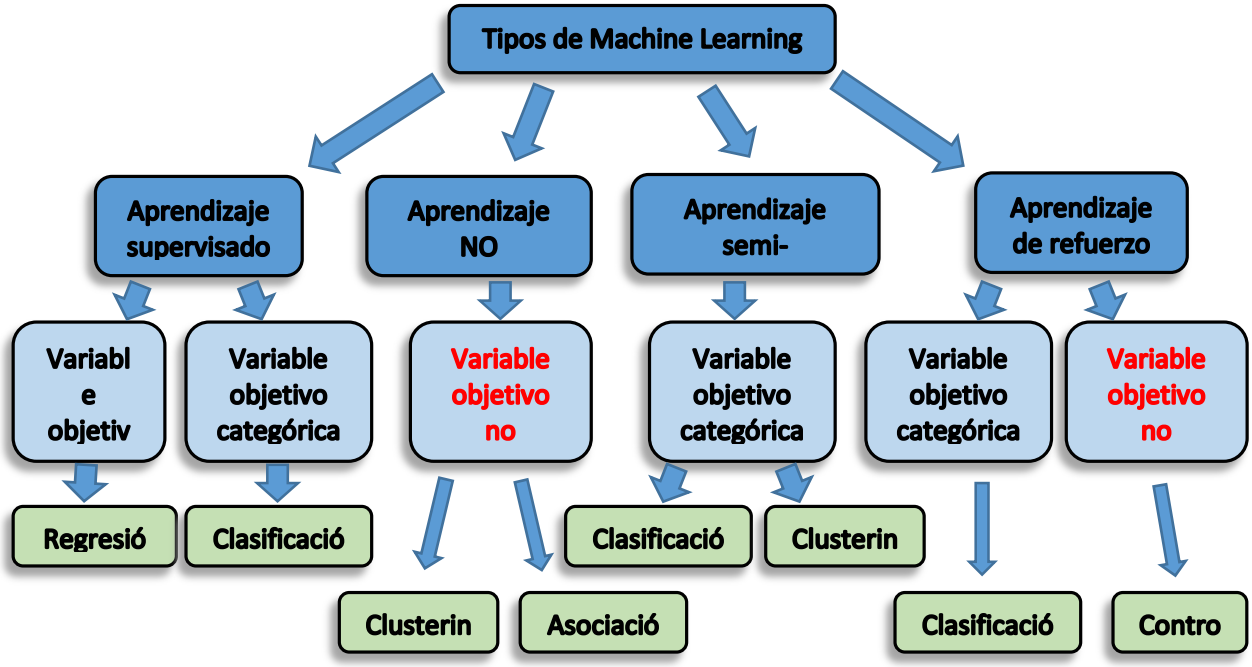


Figura 17 Tipos de Machine Learning
Fuente: Adaptado de (Ramasubramanian & Singh, 2017)

h) Plataformas para Machine Learning

Figura 18 Cuadrante mágico plataformas de Machine Learning
Fuente: Tomado de (Gartner, 2021)

De acuerdo con el Cuadrante mágico para plataformas de ciencia de datos y aprendizaje automático de (Gartner, 2021) ver Figura 16, la plataforma líder para Machine Learning es SAS, seguido de Alteryx, Databricks, TIBCO Software, MathWorks y Dataiku; en el cuadrante de visionarios están, Microsoft, H2O.ai, Datarobot, RapidMiner, Google, Knime y Domino .

En la Tabla 12 observamos una comparación entre las versiones de prueba y las versiones de pago de estas plataformas líderes y visionarias del cuadrante mágico de Gartner (RapidMiner, 2020), (TIBCO, 2020), (KNIME, 2020), (SAS, 2020), (MathWorks, 2020), (Databricks, 2020), (H2O.ai, 2020), (Microsoft Azure, 2020), (Google Cloud, 2020), (DataRobot, 2020), dado que esto será una limitante para la selección de la herramienta a utilizar.

Tabla 12

Comparación Costos Plataformas Machine Learning

Plataforma	**Versión Prueba (días/$/GB)**	**Versión paga ($/hora)**
Rapidminer	30 días	7.91/instancia+AWS
TIBCO	7 días	16.00+0.384 EC2/hr+AWS
KNIME	ND	1.16+AWS
SAS	28 días	Versión student
Mathworks	30 días	Versión student
Databricks	14 días	0.40+0.16
H2O.ai	21 días	3.06
IBM	$200	0.50/1000 predicciones
Microsoft	10GB	9.99
Google	$300/año	1.98 entrenamiento+0.0791 predicción
Datarobot	ND	ND

Se seleccionó IBM cloud con Watson Studio, de las plataformas Challengers de acuerdo con el cuadrante de Gardner, dado que tiene opciones de uso libres. Y en la misma línea de uso libre, se seleccionó Anaconda, la cual se encuentra en el primer cuadrante, pero que tiene las mismas herramientas, pero con la limitante del hardware personal que uno posee.

i) Metodología de analítica de datos

Existen diferentes marcos y metodologías provenientes de la minería de datos para hacer el análisis de datos, como: KDD, SEMMA, CATALYST y CRISP-DM, pero considerando que, para el modelamiento predictivo de la deserción de estudiantes universitarios, necesitamos cubrir aspectos del entendimiento del entorno universitario que hemos desplegado en el estado del arte de este estudio, se propone un marco mostrado en la Fig. 19

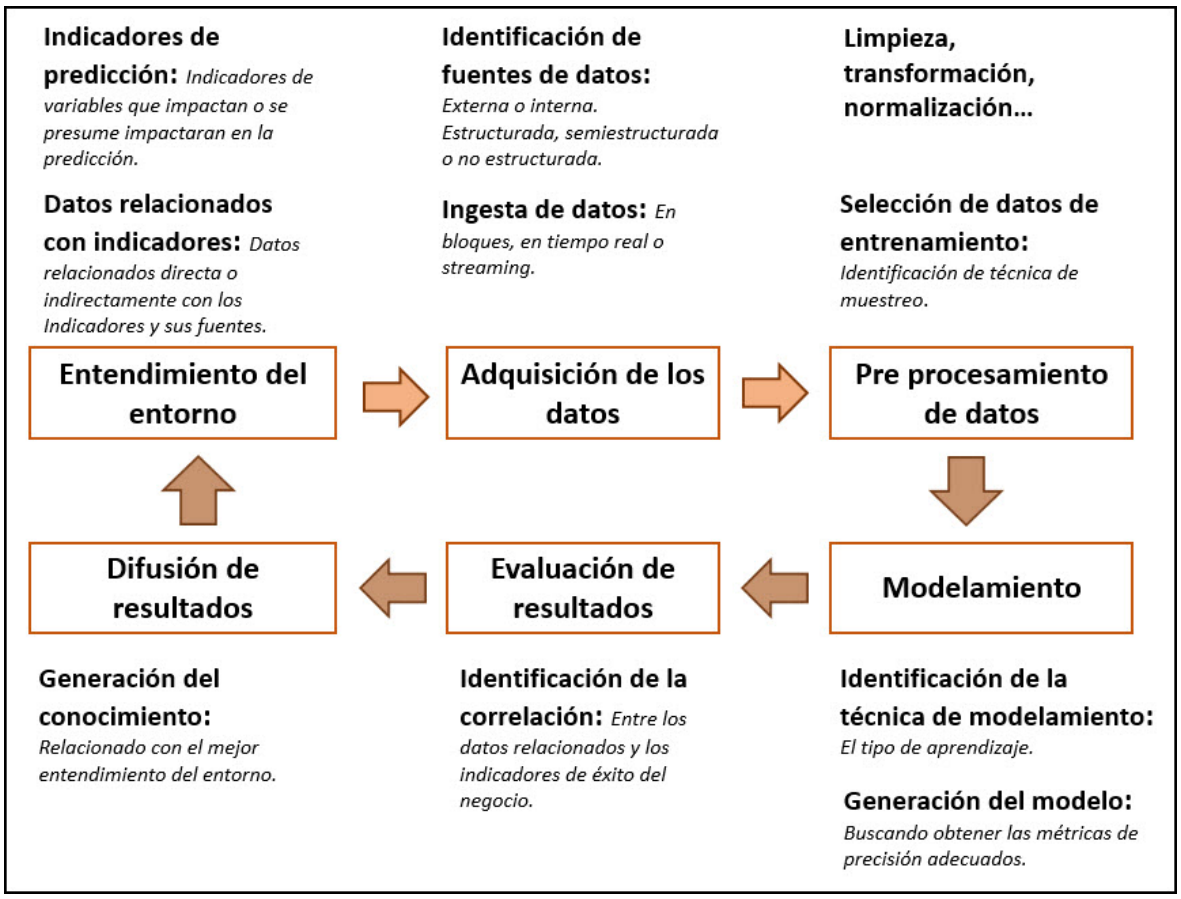


Figura 19 Modelo propuesto para desplegar el Proyecto de analítica predictiva del nivel de compleción de la carrera de estudiantes de la Universidad Nacional del Santa.

En el cual hacemos énfasis en la libertad de elección de las variables que impactaran en la predicción, dado que cada institución educativa tiene y obtiene data de sus estudiantes de manera distinta, y remarcamos que el modelo se basa en la variable objetivo multiclase que se propone por primera vez: estudiante, egresado, graduado (bachiller) y titulado.

2.2. Marco conceptual

2.2.1. Modelo

De las tantas acepciones del Diccionario de la lengua española (Real Academia Española, 2019), para nuestro contexto sería: "*arquetipo o punto de referencia para imitarlo o reproducirlo*", también, "*representación en pequeño de alguna cosa*", inclusive, "*Esquema teórico, generalmente en forma matemática, de un sistema o de una realidad compleja, como la evolución económica de un país, que se elabora para facilitar su comprensión y el estudio de su comportamiento*"

2.2.2. Modelamiento predictivo

Cuando hablamos de predicción de manera general, se entiende como pronosticar el futuro o lo que desconocemos, esto se entendió así desde mucho antes que el método científico apareciera, quien realizaba esta tarea era el astrologo o el brujo. Si partimos de los anterior, el modelamiento predictivo o análisis predictivo colegiremos que este se realiza empleando una serie de técnicas estadísticas y computacionales de manera ordenada para pronosticar resultados futuros a partir de los datos del pasado (Berea, 2017).

2.2.3. Desempeño académico

También entendido como rendimiento académico, (Salvador Blanco & Garcia-Valcarcel Muñoz-Repiso, 1989) es aquel que puede ser medido al finalizar los estudios con las tasas de promoción, la repetición de la misma asignatura y el abandono cuando dejan de matricularse en las asignaturas de la carrera.

2.2.4. Seguimiento académico

Partimos con la (Real Academia Española, 2019) donde seguimiento es la "*acción y efecto de seguir o seguirse*", desde el origen y también con la (Real Academia Española, 2019), seguir significa "*ir después o*

detrás de alguien", también, "*ir en busca de alguien o algo; dirigirse, caminar hacia él o ello*". Por otro lado, el termino académico es definido como "*perteneciente o relativo a centros oficiales de enseñanza, especialmente a los superiores*", también, "*individuo perteneciente a una corporación académica*", por lo tanto, de la suma de ambos términos, si tenemos en cuenta que los individuos dentro de la academia son los estudiantes, entonces la acción de seguimiento se ejercería sobre ellos y orientado a mirar su desempeño de acuerdo con 2.2.3.

2.2.5. Clasificación multiclase

La clasificación multiclase o multinomial deben tener tres o más categorías. Por ejemplo, para predecir las condiciones meteorológicas, puede tener cinco categorías: lluvioso, nublado, soleado, nevado y ventoso (Quinto, 2020).

2.2.6. Perceptrón multicapa

Es una red artificial de retroalimentación que consta de varias capas de nodos completamente conectados. Los nodos de la capa de entrada corresponden al conjunto de datos de entrada.

Los nodos de las capas intermedias utilizan una función logística (sigmoidea, RELU), mientras que los nodos de la capa de salida final utilizan una función softmax para admitir la clasificación multiclase. El número de nodos en la capa de salida debe coincidir con el número de clases (Quinto, 2020).

2.2.7. IBM Watson Studio

Watson es el conjunto de herramientas, aplicaciones y servicios de inteligencia artificial de IBM. Es una colección de servicios en la nube que se ejecutan en la plataforma IBM Cloud. Una de las herramientas es Watson Studio que ayuda a extraer valor y conocimientos de los datos al permitir entornos colaborativos de ciencia de datos y aprendizaje automático para crear y entrenar modelos de IA, y preparar

y analizar datos en un único entorno integrado (Sabharwal, Barua, Anand, & Aggarwal, 2020).

2.2.8. XGBoost

Son las siglas de eXtreme Gradient Boosting, se refiere al objetivo de ingeniería de empujar el límite de recursos de cálculo para algoritmos de árboles potenciados, es una implementación de máquinas de aumento de gradiente creadas por Tianqi Chen. Cuenta con implementación Sparse Aware con manejo automático de valores de datos faltantes, estructura de bloques para apoyar la paralelización en la construcción de árboles, y entrenamiento continuo que impulsa aún más un modelo ya configurado con nuevos datos (Brownlee, 2018).

2.2.9. Optimización de hiperparámetros

Los sistemas de aprendizaje automático (ML) tiene hiperparámetros, la tarea básica en ML automatizado (AutoML) es configurar automáticamente estos hiperparámetros para optimizar el rendimiento. Las redes neuronales profundas dependen de una amplia gama de opciones de hiperparámetros, la regularización y la optimización de la red neuronal. La optimización de hiperparámetros automatizada (HPO) puede reducir el esfuerzo humano para aplicar el ML automático pues mejora el rendimiento de los algoritmos de aprendizaje automático (adaptándolos al problema en cuestión); mejora la reproducibilidad y la equidad de los estudios científicos. La HPO automatizada es claramente más reproducible que la búsqueda manual (Feurer & Hutter, 2019).

2.2.10. Anaconda

Es una distribución de Python para el procesamiento de datos a gran escala, el análisis predictivo y la computación científica. Anaconda incluye NumPy, SciPy, matplotlib, pandas, IPython, Jupyter Notebook y scikit-learn (Müller & Guido, 2016).

2.2.11. Jupyter Notebook

Es un entorno interactivo para ejecutar código en el navegador, es una herramienta para el análisis de datos exploratorios y es ampliamente utilizada por los científicos de datos. Jupyter Notebook admite muchos lenguajes de programación, pero basta el soporte de Python, además facilita la incorporación de código, texto e imágenes.

2.2.12. Numpy

Un paquete fundamental para la computación científica en Python. Trabaja con matrices multidimensionales, funciones matemáticas de alto nivel, tales como operaciones de álgebra lineal, transformada de Fourier, y generadores de números pseudoaleatorios (Müller & Guido, 2016).

2.2.13. Pandas

Es una biblioteca de Python para gestión y análisis de datos. Trabaja con una estructura de datos llamada DataFrame que es una tabla, similar a una hoja de cálculo de Excel. Pandas proporciona una gran variedad de métodos para modificar y operar en esta tabla; permite consultas de tipo SQL y uniones de tablas. A diferencia de NumPy, que requiere que todas las entradas en una matriz sean del mismo tipo, pandas permite que cada columna tenga un tipo separado (por ejemplo, números enteros, fechas, números de punto flotante y cadenas). Otra herramienta valiosa proporcionada por pandas es su capacidad para ingerir desde una gran variedad de formatos de archivo y bases de datos, como SQL, archivos de Excel y archivos de valores separados por comas (CSV) (Müller & Guido, 2016).

2.2.14. Tensorflow

Es una plataforma de aprendizaje automático de código abierto con enfoque particular en las redes neuronales, desarrollada por el equipo de Google Brain, fue lanzada la biblioteca bajo la Licencia Apache 2.0 a finales del 2015, siendo una biblioteca de código abierto. Aunque los

casos de uso de TensorFlow no se limitan a las aplicaciones de aprendizaje automático, el aprendizaje automático es el campo donde vemos la fuerza de TensorFlow. Hay dos lenguajes de programación con API de TensorFlow estables y oficiales son Python y C (Yalçın, 2021).

2.2.15. Softmax

La función softmax nos permite generar una distribución de probabilidad categórica sobre K clases.

$$f(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

Es un caso particular para las funciones de activación, ya que rara vez se ve como una activación que ocurre entre capas. Por lo que, softmax casi siempre es la última capa de una red para la clasificación multiclase en lugar de una función de activación (Kamath, Liu, & Whitaker, 2019).

2.2.16. Cross-entropy

La entropía cruzada es la medida de rendimiento de un clasificador, es una función continua que siempre es positiva y es igual a cero si la salida predicha coincide exactamente con la salida deseada. Por lo tanto, el objetivo de esta optimización es minimizar la entropía cruzada, por lo que es lo más cercano a cero posible, cambiando las variables en las capas de la red. TensorFlow tiene una función incorporada para calcular la entropía cruzada (Zaccone & Karim, 2018).

2.2.17. Adam Optimizer

Es un algoritmo para la optimización que se basa en gradientes de primer orden de funciones objetivas estocásticas, y también en estimaciones adaptativas de momentos de orden inferior. Es sencillo de implementar, computacionalmente eficiente, con pocos requisitos de memoria, es invariante al cambio de escala diagonal de los gradientes y es muy adecuado para problemas que son grandes en términos de

datos y/o parámetros. Adam también es apropiado para objetivos no estacionarios y problemas con gradientes muy ruidosos y/o dispersos (Kingma & Ba, 2015).

2.2.18. Epoch

Cada iteración sobre todos los datos de entrenamiento se denomina una época (Chollet, 2017)

2.2.19. Learning Rate

O tasa de aprendizaje, es un parámetro en algoritmos de optimización que regula el tamaño del paso tomado en cada iteración mientras avanza un mínimo de una función de costo o pérdida. Con una tasa de aprendizaje rápida, el modelo converge alrededor del mínimo más rápido, pero puede sobrepasar el punto mínimo real. Con una tasa de aprendizaje lenta, la optimización puede llevar demasiado tiempo. Por lo que se debe elegir la tasa de aprendizaje óptima, que le permite al modelo encontrar el punto mínimo deseado en un tiempo razonable (Yalçın, 2021).

2.2.20. Mini-Batch o Batch,

Es un pequeño conjunto de muestras (normalmente entre 8 y 128) que el modelo procesa simultáneamente. El número de muestras suele ser una potencia de 2, lo cual facilita la asignación de memoria en la GPU. Durante el entrenamiento, se usa un mini-lote para calcular una única actualización de descenso de gradiente aplicada a los pesos del modelo (Chollet, 2017).

CAPÍTULO III

MARCO METODOLÓGICO

3.1. Hipótesis central de la investigación

Un modelo basado en Machine Learning permite predecir el nivel de compleción de la carrera de los estudiantes de las Escuelas Profesionales de Ingeniería de la Universidad Nacional del Santa con un nivel de precisión mayor al 85%

3.2. Variables e indicadores de la investigación

Tabla 13
Operacionalización de variables

Variable	Definición Conceptual	Indicadores	Tipo	Técnica	Instrumento
VI Modelo predictivo basado en Machine Learning	Conjunto de normas y herramientas que generará perfiles de desempeño ad-hoc (estudiante, egresado, grado y/o título)	Modelos a comparar	Cuantitativo	Modelado	Software Machine learning
		Numero de capas	Cuantitativo	Modelado	Software Machine learning
		Número de nodos	Cuantitativo	Modelado	Software Machine learning
		Número de iteraciones	Cuantitativo	Modelado	Software Machine learning
VD Nivel de compleción del estudiante universitario	Gestión del desempeño de los estudiantes con modelo predictivo	Precisión del conjunto de entrenamiento	Cuantitativo	Caso de estudio	Modelo seleccionado
		Precisión del conjunto de prueba	Cuantitativo	Caso de estudio	Modelo Seleccionado
		Característica que más influye	Cuantitativo	Caso de estudio	Modelos comparados

3.3. Método de la Investigación

1. Revisaremos la bibliografía pertinente, así como las normas al interno de la UNS.

2. Se seleccionará la muestra sobre a la cual se le aplicará piloto del modelo del estudio
3. Selección de atributos pertinentes para desarrollar el modelo
4. Recolección y depuración de la data pertinente. De las fuentes históricas, de encuestas a la muestra seleccionada.
5. Proceso de aprendizaje del modelo
6. Aplicación del piloto sobre datos aleatorios de prueba de precisión
7. Resultados y discusión
8. Conclusiones recomendaciones

3.4. Diseño

Cuasi experimental

Pretest → X → Postest

3.5. Población y Muestra

Población: Estudiantes de la Universidad Nacional del Santa (UNS)

Muestra: Dirigida, dado que los estudiantes de las 04 Escuelas Profesionales de la Facultad de Ingeniería de la UNS entre los semestres 2004-2018 tienen la más baja proporción de bachilleres vs ingresantes (39.13%), así como la de titulados vs ingresantes (20.06%) con respecto a las otras 2 Facultades, tal como se observa en la Tabla 7.

3.6. Técnicas e Instrumentos de Recolección de Datos

1. Análisis documental de la reglamentación académica UNS
2. Archivo réplica de la base de datos socio económico de los estudiantes de la Facultad de Ingeniería de la UNS del sistema SIIGAA
3. Archivo réplica de estudiantes egresados y promedio por semestre de la Facultad de Ingeniería de la UNS del sistema SIIGAA
4. Archivos réplica de estudiantes graduados y titulados de la Facultad de Ingeniería de la UNS de la oficina de grados y títulos.

3.7. Procedimiento de la Recolección de Datos.

Se recolectarán las notas y asistencias de los módulos respectivos del sistema de gestión académica de la UNS.

Parte de la data socio económica se obtendrá del sistema de bienestar universitario con la confidencialidad respectiva para uso de investigación de manera anónima.

3.8. Técnicas de Procesamiento y análisis de Resultados

Se evaluará el modelo sin entrenamiento y con entrenamiento por juicio de expertos en función a los rendimientos esperados para la variable dependiente. Para ello se utilizará las herramientas software libre: Júpiter, Anaconda, Python, Pandas.

3.9. Colección de datos caso estudio

La data es el elemento esencial para poder abordar un modelo predictivo con machine Learning.

A continuación, mostraremos como se obtuvo y como se transformó la data para poder generar los conjuntos de datos (datasets), que permitan entrenar, testear y validar el modelo de predicción propuesto en la presente investigación. El código Python empleado se adjunta en el Anexo 1.

a) Data socio-económica
   La data obtenida del sistema de bienestar universitario y proporcionada en una exportación de la base de datos en Excel, cuenta con 15 características las cuales se encuentran en la Tabla 14, en la cual observamos que la característica servicios es multivaluada, pues contiene los siguientes servicios separados por comas: luz, agua, desagüe, teléfono, cable e internet.

b) Data Graduados y titulados
   Proporcionada por la Oficina de grados y títulos de la UNS, anonimizada de su base de datos en Excel, con 03 características relevantes cada una se muestran en las Tablas 15 y 16

Tabla 14

Data socio económica

Número	Característica	Registros	Tipo de dato
1	Código	767	Int64
2	Fecha de Nacimiento	767	string
3	Lugar de Nacimiento	767	string
4	Sexo	767	string
5	Estado civil	767	string
6	Celular	767	string
7	Dependencia económica	767	string
8	Condición de trabajo responsable familia	767	string
9	Alumno trabaja	767	string
10	Ingreso total familiar	767	float64
11	Lugar de procedencia	767	string
12	Material de vivienda	767	string
13	Servicios	767	string
14	Tipo de colegio	767	string
15	Veces que postulo	767	Int64

Fuente: Oficina de tecnologías de la información UNS

Tabla 15

Data de graduados UNS

Número	Característica	Registros	Tipo de dato
1	Código	2873	Int64
2	Escuela profesional	2873	string
3	Fecha diploma	2873	string

Fuente: Oficina de grados y títulos UNS

Tabla 16

Data de titulados UNS

Número	Característica	Registros	Tipo de dato
1	Código	2068	Int64
2	Escuela profesional	2068	string
3	Fecha diploma	2068	string

Fuente: Oficina de grados y títulos UNS

c) Análisis de los datos

La data necesaria ha sido colectada de las diversas bases de la institución del caso de estudio, tal como se muestra en la Figura 20

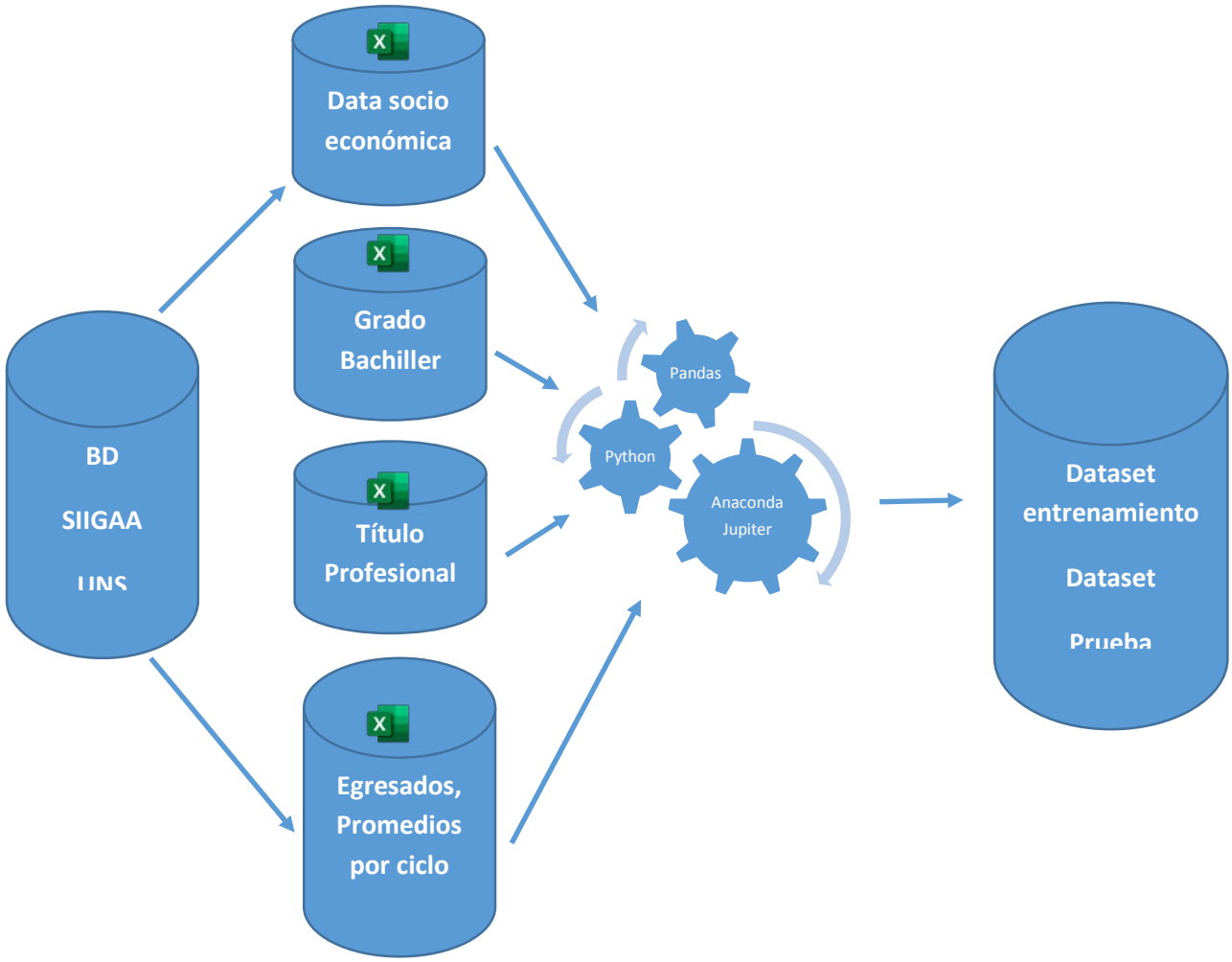


Figura 20 Análisis de la data - caso de estudio

En muchos casos la data que se ha colectado tiene deficiencias, dado contiene muchos valores nulos, el mal diseño de las bases de datos transaccionales, hace que se tenga campos multivalores con hasta 6 términos distintos, los identificadores de los registros tienen diferente formato y tipo de dato en la medida que la data viene de fuentes distintas, luego se unió la data a través de su identificador ya estandarizado, y se mejoró la calidad de los datos, eliminando los campos sin valor y categorizando aquellos que se encuentran muy dispersos.

Para ello hemos utilizado Python, que de acuerdo con (Müller & Guido, 2016) es la lengua franca para las aplicaciones de la ciencia de datos, ya que tiene librerías para carga de data, visualización, estadísticas y más, asociada con la biblioteca scikit-learn, el cual es un proyecto de código abierto dependiente de los paquetes Python: NumPy y SciPy, los cuales deben ser asociados con pandas, matplotlib y Júpiter Notebook, todos estos trabajan en la distribución de Python: Anaconda.

1. Trabajando con la data socio económica
    1° Se separó la característica multivaluada Servicios, en cada uno de sus servicios por separado: luz, agua, desagüe, teléfono, cable e internet. Y se eliminó la característica fecha de nacimiento por no ser relevante para el estudio.
    2° Se analizó su mapa de calor para revisar la calidad de la data en cuanto a valores nulos o faltantes, Figura 21.

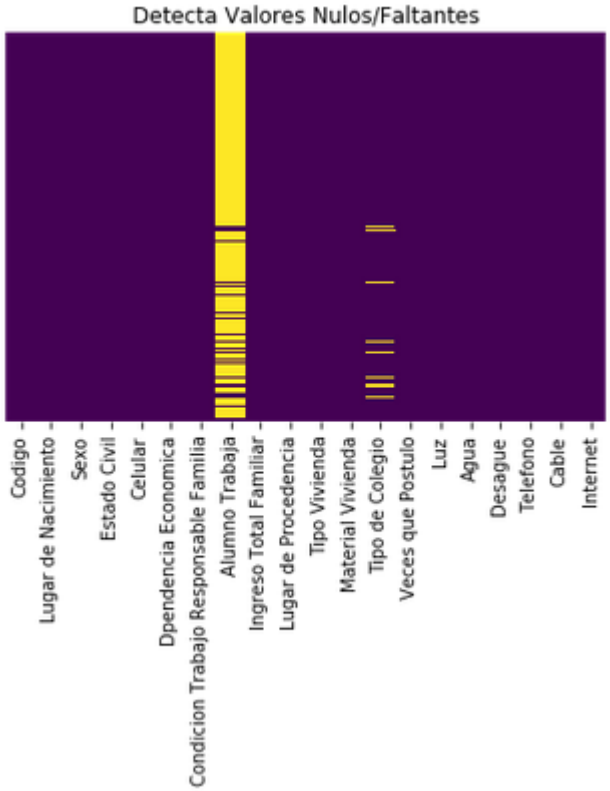


Figura 21 Mapa de calor de la data socio económica

Se detectó que la característica: Alumno trabaja, más del 90% de color amarillo, eran valores faltantes o no ingresados, por lo que se descartó esta característica.

Al evaluar la característica Tipo de colegio, también con pintas amarillas, Figura 22

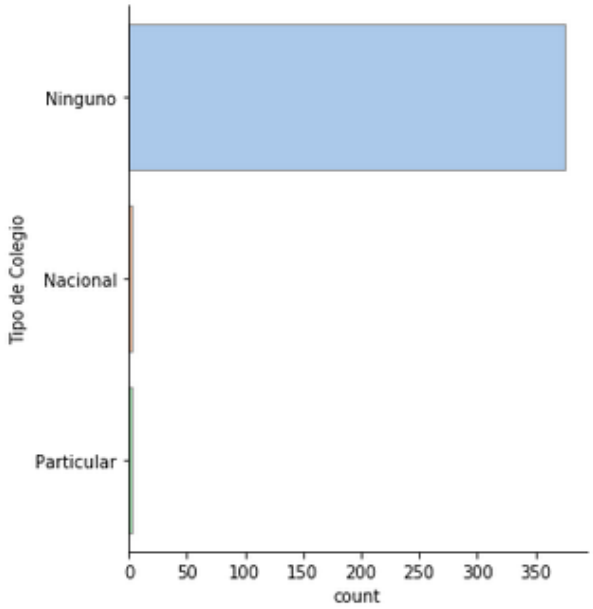


Figura 22 Característica: Tipo de Colegio

Se encontró que el 90% de la referencia decía ninguno, motivo por el cual se descartó esta característica, por tener demasiado sesgo.

3° De igual manera, las características: Lugar de nacimiento, Estado civil, Tipo de vivienda, Veces que postuló, Material de la vivienda, se descartaron porque tenían demasiado sesgo, tal como se ve en las Figuras 23, 24, 25, 26 y 27

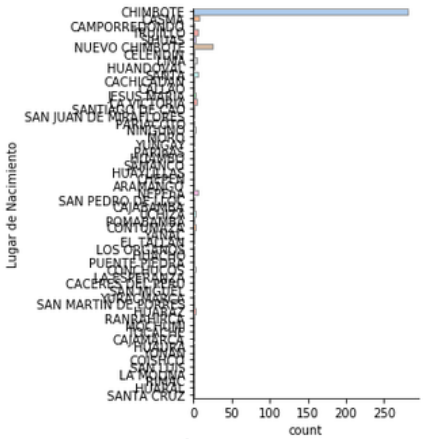


Figura 23 Característica: Lugar de nacimiento

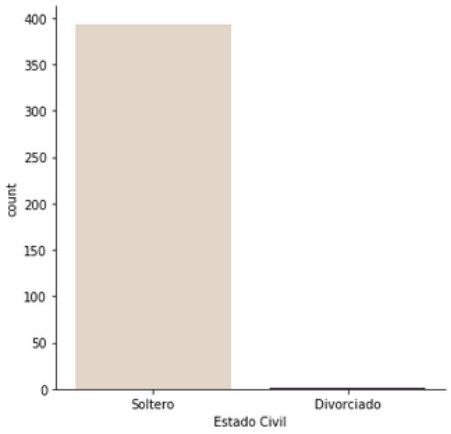


Figura 24 Característica: Estado civil

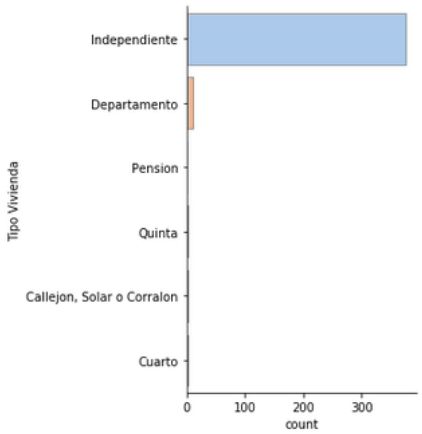


Figura 25 Característica: Tipo de vivienda

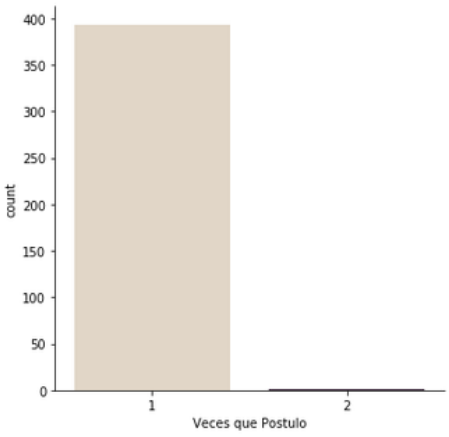


Figura 26 Característica: Veces que postuló

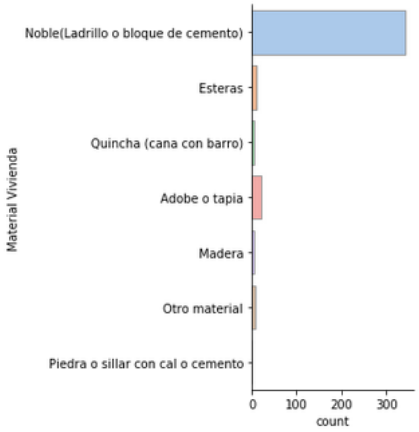


Figura 27 Característica: Material de la vivienda

4° Se reajustaron las siguientes características a valores booleanos, como el sexo, (tiene) celular a 0 y 1

5° Las características Dependencia económica y Condición trabajo responsable familia y Lugar de procedencia, de tipos categóricos, fueron normalizados a valores numéricos.

Todos estos cambios se resumen en la Tabla 17

Tabla 17

Data Socio Económica depurada

Número	Característica	Registros	Tipo de dato
1	Código	767	String
2	Sexo	767	Int64
3	Celular	767	Int64
4	Dependencia económica	767	Int64
5	Condición de trabajo responsable familia	767	Int64
6	Ingreso total familiar	767	category
7	Lugar de procedencia	767	Int64
8	Luz	767	Int64
9	Agua	767	Int64
10	Desagüe	767	Int64
11	Teléfono	767	Int64
12	Cable	767	Int64
13	Internet	767	Int64

2. Trabajando con la data de graduados y titulados
   En ambos casos se eliminó la característica Fecha diploma y Escuela profesional.

3. Merge data adquirida de diversas fuentes
   Con la data socio económica, graduados y titulados depurada, así como los promedios de notas de todos los semestres matriculados y la cantidad de semestres matriculados se procedió a efectuar un merge de estas fuentes de datos con pandas a través del identificador código, clave primaria estandarizada en todos los casos. Obteniéndose el siguiente dataset maestro en la Tabla 18

Tabla 18

Dataset maestro

Número	Característica	Registros	Tipo de dato
1 (*)	Código	767	String
2	Sexo	767	Int64
3	Celular	767	Int64
4	Dependencia económica	767	Int64
5	Condición de trabajo responsable familia	767	Int64
6	Ingreso total familiar	767	category
7	Lugar de procedencia	767	Int64
8	Luz	767	Int64
9	Agua	767	Int64
10	Desagüe	767	Int64
11	Teléfono	767	Int64
12	Cable	767	Int64
13	Internet	767	Int64
14	Promedios	767	Float64
15	Número de semestres	767	Float64
16	Estudiante	767	Int64
17	Egresado	767	Int64
18	Graduado	767	Int64
19	Titulado	767	Int64

(*) El código se excluye para efectos de evaluación

4. Separando dataset de entrenamiento y de prueba
   Del total de registros resultantes del merge, se han separado 394 registros como parte del dataset de entrenamiento y 372 registros para el dataset de prueba. Los cuales fueron guardados en archivos con formato csv.

5. Gráficos estadísticos del dataset de entrenamiento
   A continuación, visualizaremos gráficamente cada característica del dataset de entrenamiento maestro, con la dispersión de los datos que contienen.

- Mapa de calor de las 17 características del dataset maestro

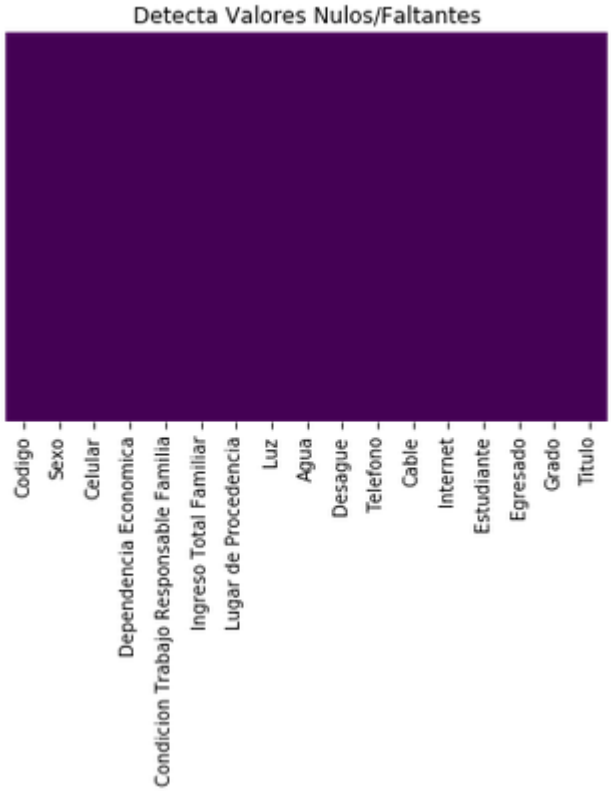


Figura 28 Mapa de calor dataset de entrenamiento

En la Figura 28, se nota que cada característica está poblada, lo que se verificará con las siguientes gráficas de cada una de ellas en las Figuras de la 29 a la 42.

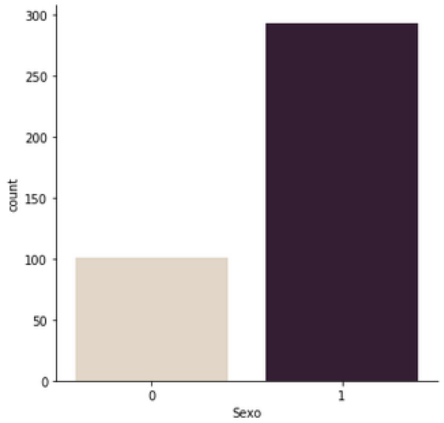


Figura 29 Característica: Sexo

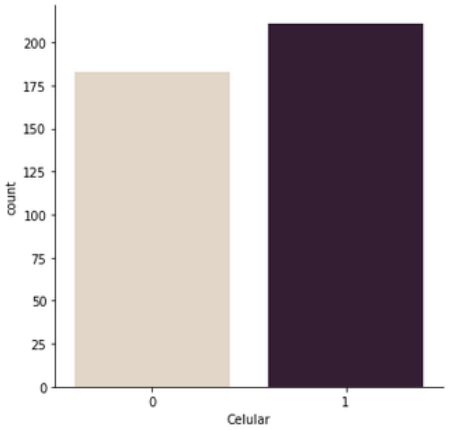


Figura 30 Característica: Celular

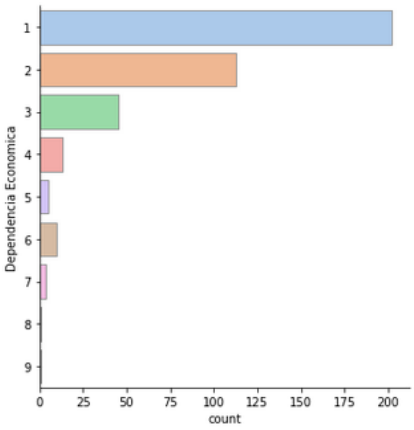


Figura 31 Dependencia Familiar

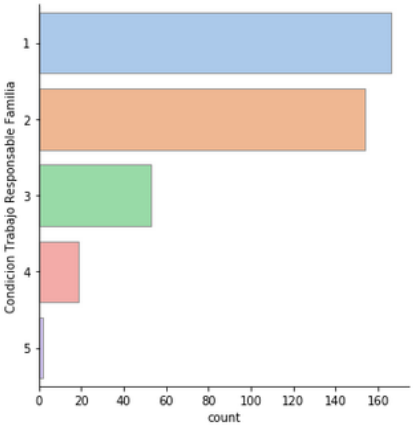


Figura 32 Condición trabajo responsable

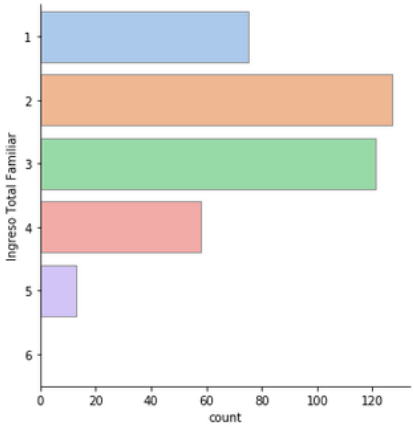


Figura 33 Ingreso total familiar

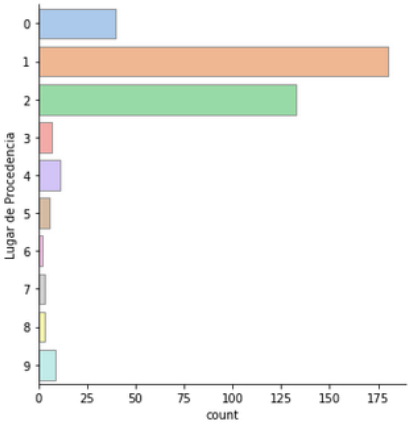


Figura 34 Lugar de procedencia

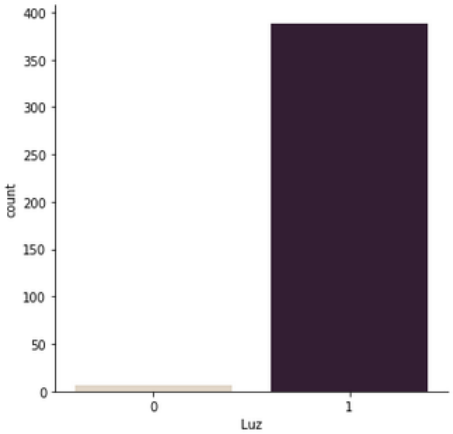


Figura 35 Característica: Luz

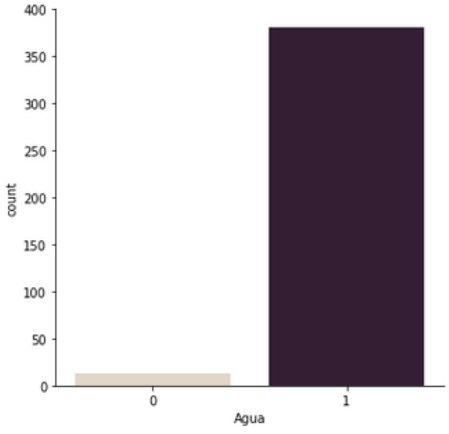


Figura 36 Característica: Agua

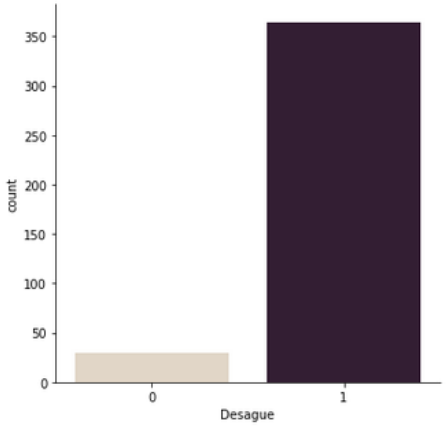


Figura 37 Característica: Desagüe

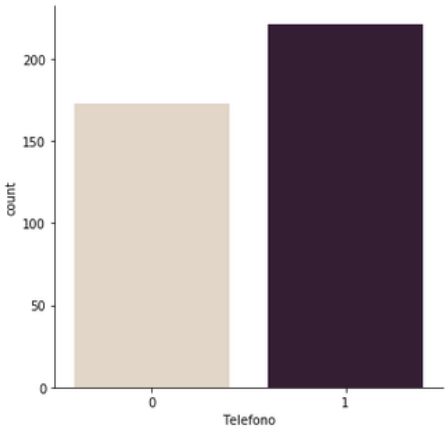


Figura 38 Característica: Teléfono

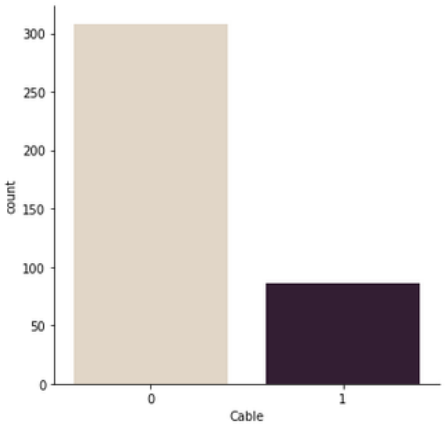


Figura 39 Característica: Cable

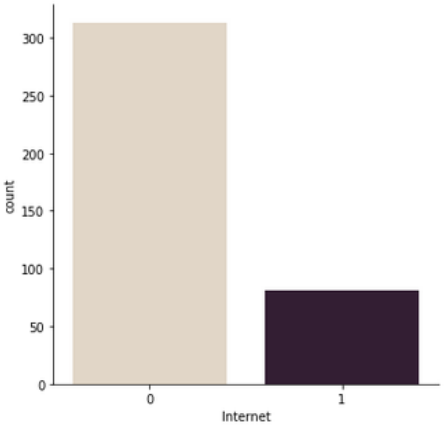


Figura 40 Característica: Internet

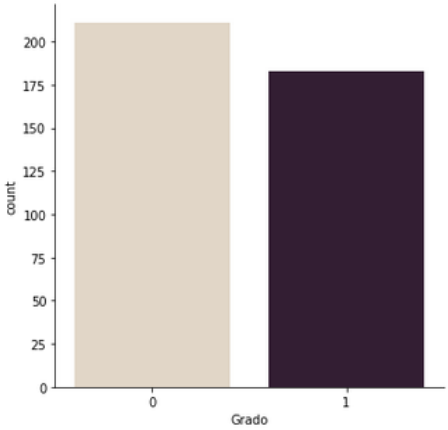


Figura 41 Característica: Graduado

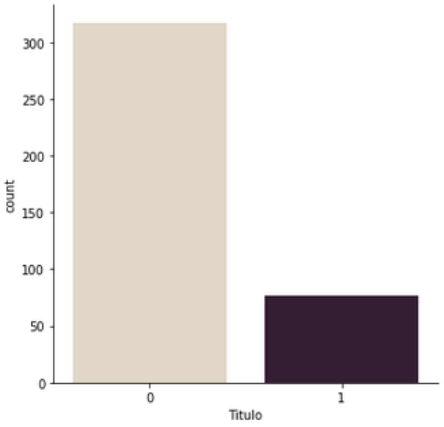


Figura 42 Característica: Titulado

CAPÍTULO IV.- Resultados y Discusión

RESULTADOS

1. Seleccionar características

   Luego del procesamiento de la data recogida, esta fue depurada, descartando las características que ofrecían sesgo elevado, así como aquellos que tenían exceso de nulos y/o faltantes. Para evitar el sesgo de ciertas características que contenían valores muy alto con respecto a otras, estas fueron regularizadas de tal manera que todas se encuentren en el rango de 0 a 1. Finalmente, el código del alumno, es un dato que se descarta pues no aporta a la predicción, salvo su valor de identificador. En la Tabla 19 se listan las 14 características de entrada, así como las 4 últimas que conforman la función objetivo, las que están incluidas en el dataset correspondiente.

Tabla 19 Características seleccionadas

Número	Categoría	Característica	Valores
1	Características no académicas (Demográficas)	Sexo	M/F
2		Celular	V/F
3		Dependencia económica	Rango
4		Condición de trabajo responsable familia	Rango
5		Ingreso total familiar	Rango
6		Lugar de procedencia	0-9
7		Luz	V/F
8		Agua	V/F
9		Desagüe	V/F
10		Teléfono	V/F
11		Cable	V/F
12		Internet	V/F
13	Características académicas	Promedios	Rango
14		Número de semestres	Rango
15		Estudiante	V/F
16		Egresado	V/F
17		Graduado	V/F
18		Titulado	V/F

2. Diseñar modelo predictivo

   Teniendo en cuenta la problemática del caso de estudio, es que vamos a plantear un modelo de red neuronal de predicción multiclase, para poder determinar, que dada ciertas características académicas y socio económicas se verificará la condición que son estudiantes, egresados, o logran obtener el grado de bachiller o titularse, para ello empleando aprendizaje profundo diseñaremos una red neuronal perceptrón multicapa la cual será comparada

con los modelos de clasificación generados por la herramienta de experimentación AutoAI de Watson Studio de IBM.

El modelo inicial (03-00), consta de 14 características, 1 capa oculta de 14 nodos y una capa de salida con 4 nodos, que coinciden con las 4 clases motivo de predicción. Para la capa oculta se ha empleado la activación lineal Rectified Linear Unit (ReLU) y para la capa de salida la activación Softmax, ver Figura 43

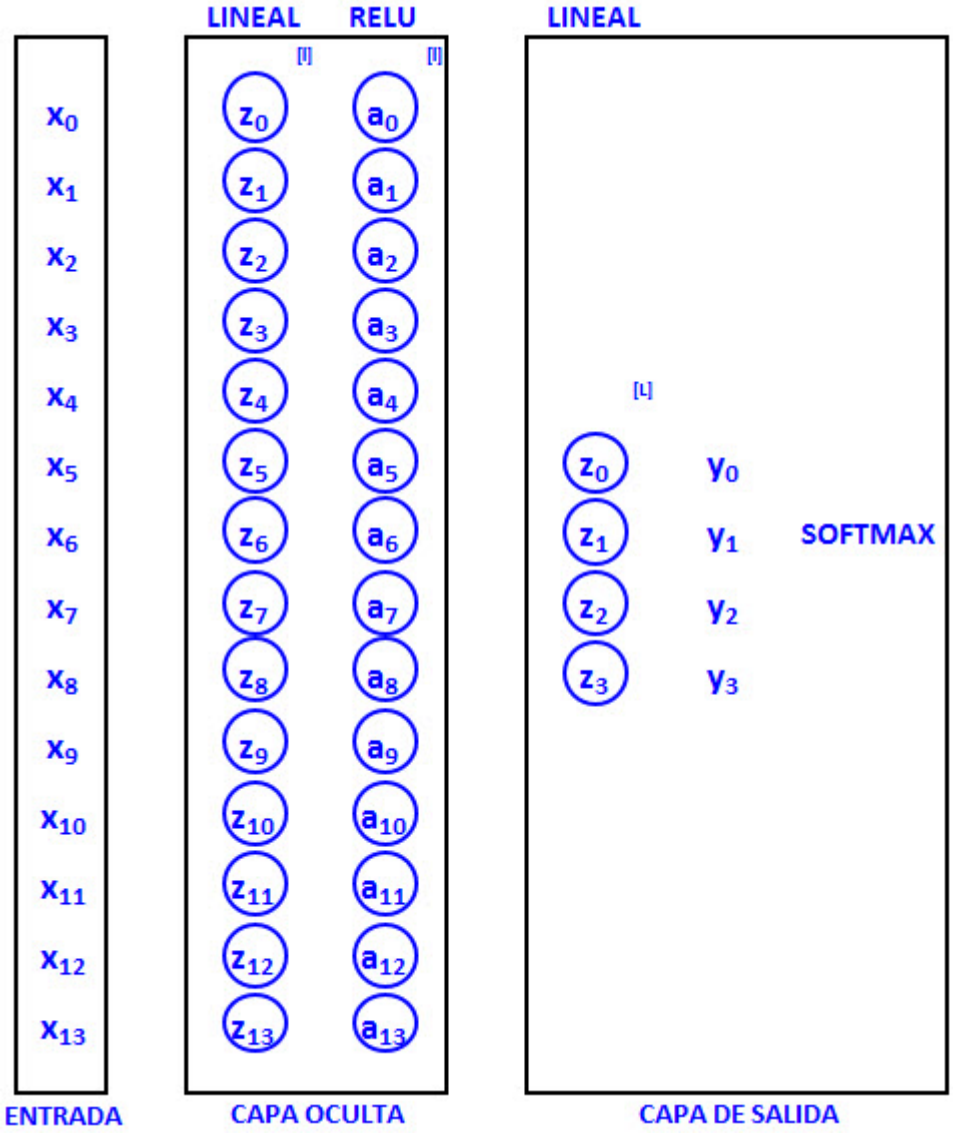


Figura 43 Modelo de 2 capas (14-4)

Manteniendo la estructura del modelo 03-00, para efectos de seleccionar el modelo que mejor precisión proporcione, proponemos 14 modelos adicionales que varíen el número de capas ocultas y cantidad de nodos en cada capa oculta, tanto la capa de entrada (14 características) como la capa de salida (4 clases de predicción) se mantienen invariables, las capas ocultas para la

propagación hacia adelante empleará una función lineal, con una activación ReLU, reservándose la activación Softmax para la capa de salida para cualquiera de los casos, Figuras 44 hasta 48, consolidados en la Tabla 20.

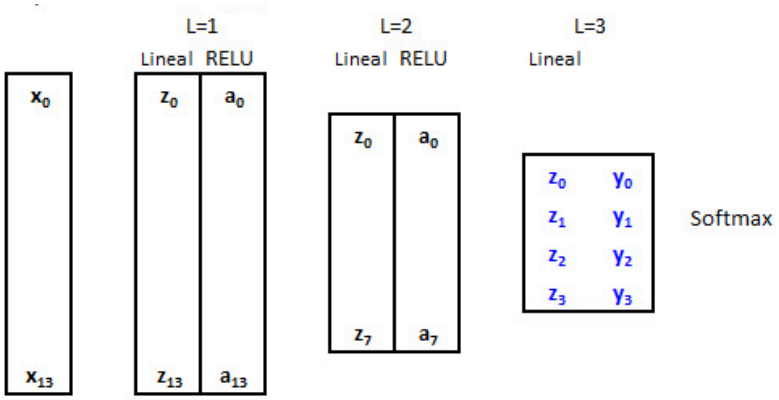


Figura 44 Modelo de 3 capas (14-8-4)

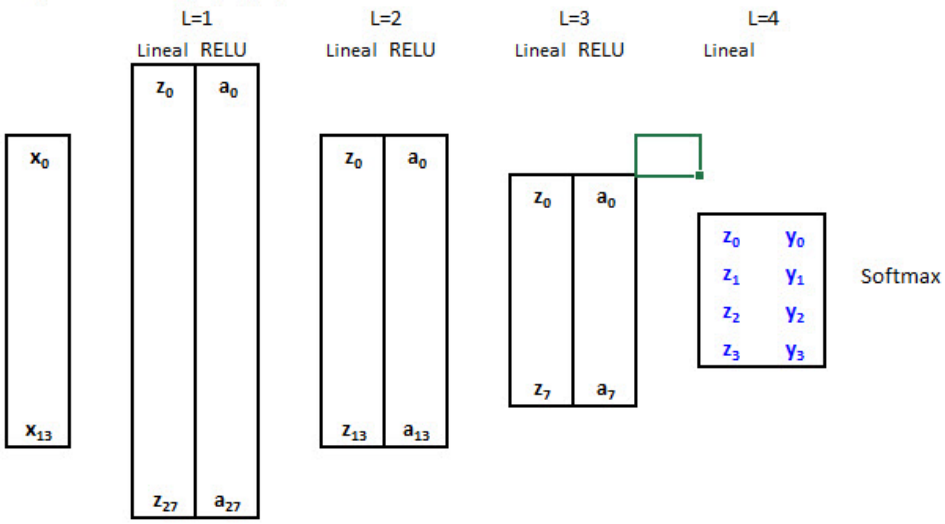


Figura 45 Modelo de 4 capas (28-14-8-4)

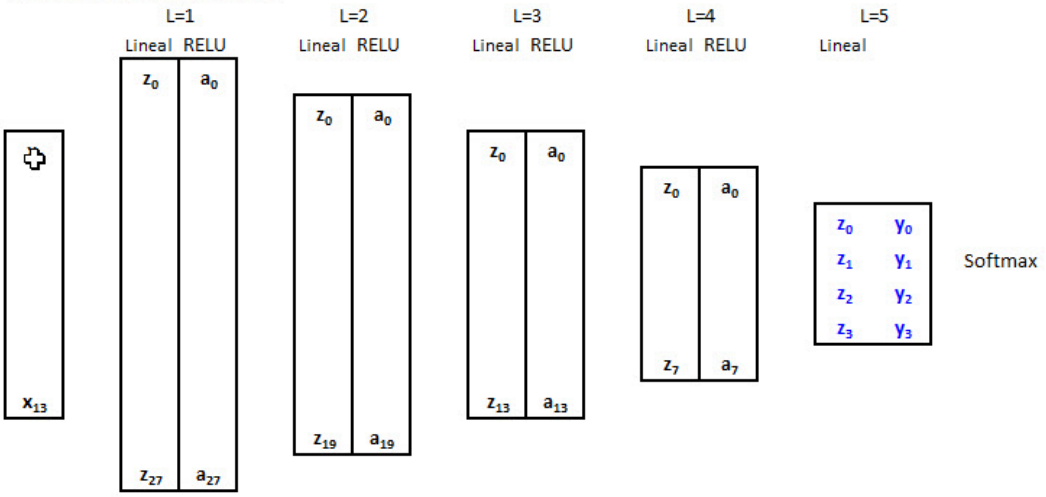


Figura 46 Modelo de 5 capas (28-20-14-8-4)

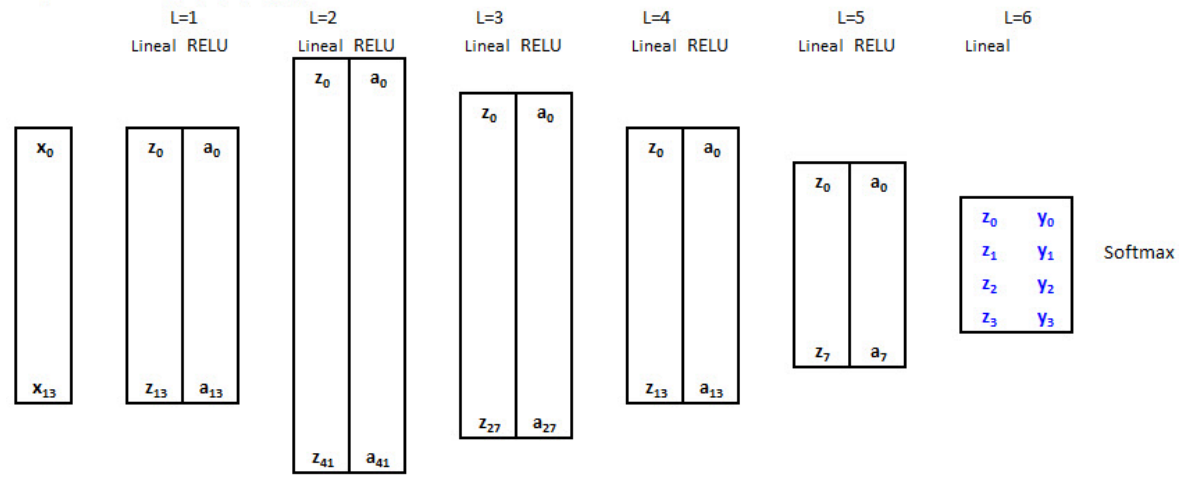


Figura 47 Modelo de 6 capas (14-42-28-14-8-4)

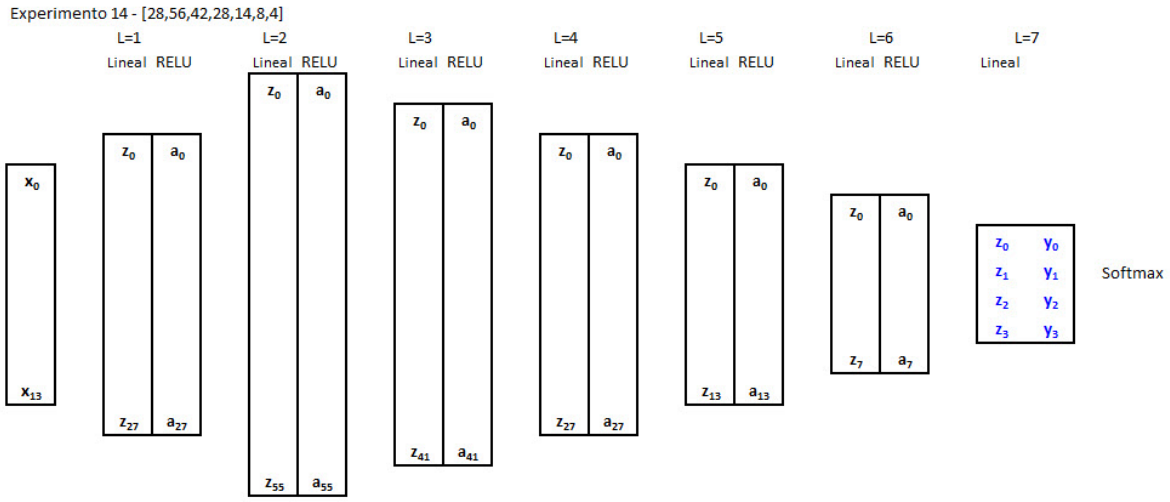


Figura 48 Modelo de 7 capas (28-56-42-28-14-8-4)

Tabla 20 Consolidado de modelos RNA para experimentar

N° Modelo	Capas y Nodos	N° Modelo	Capas y Nodos
03-00	[14, 4]	03-08	[42, 28, 8, 4]
03-01	[14, 8, 4]	03-09	[42, 14, 10, 4]
03-02	[28, 14, 4]	03-10	[42, 28, 14, 8, 4]
03-03	[28, 14, 8, 4]	03-11	[14, 42, 28, 14, 8, 4]
03-04	[28, 28, 14, 8, 4]	03-12	[42, 28, 28, 14, 14, 4]
03-05	[28, 20, 14, 8, 4]	03-13	[56, 42, 28, 14, 8, 4]
03-06	[28, 14, 14, 8, 4]	03-14	[28, 56, 42, 28, 14, 8, 4]
03-07	[42, 14, 8, 4]		

Para un primer test con la herramienta de experimentación AutoAI de Watson Studio de IBM con dataset limitado a datos de una Escuela Profesional de la UNS con 767 registros, esta herramienta ajustó automáticamente los datos a los algoritmos clasificador XGB, y el clasificador LGBM. Luego con data de 04 Escuelas Profesionales (mayor volumen de datos) los datos se ajustaron al algoritmo clasificador XGB y al algoritmo clasificador de árbol de decisiones; en ambos casos los emplearemos como control y poder comparar la precisión del modelo de aprendizaje profundo que seleccionaremos luego de experimentar.

En la Figura 49, vemos el desarrollo de los 02 algoritmos de clasificación: XGB y LGBM, los cuales han tenido 4 interconexiones y probado con 28 transformadores de características con dataset de 01 Escuela Profesional de la UNS.

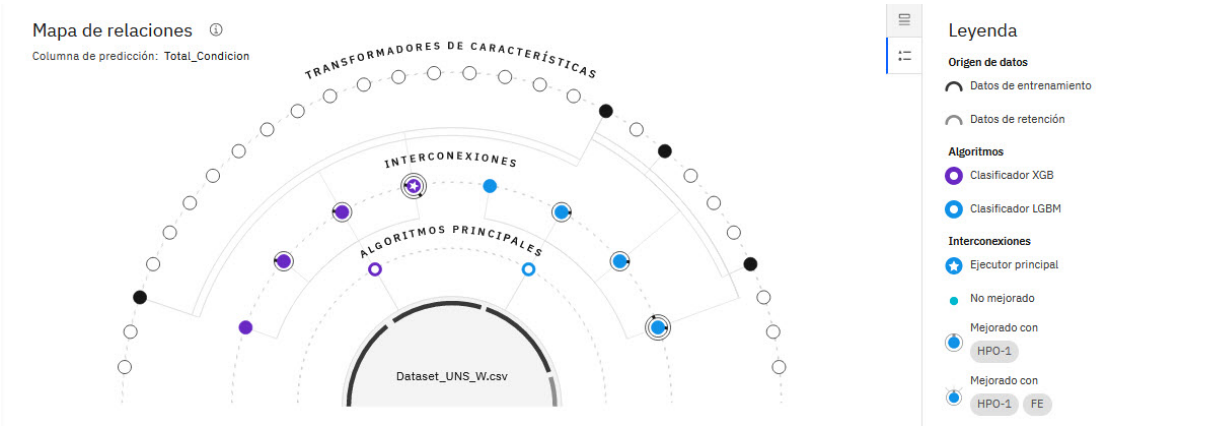


Figura 49 Mapa de relaciones Clasificación XGB y LGBM - Dataset UNS

Con dataset ampliado a 04 Escuelas Profesionales de Ingeniería con 3529 registros, el mapa de relaciones con la herramienta de experimentación AutoAI de Watson Studio de IBM seleccionó igualmente el clasificador XGB y cambió al clasificador de árbol de decisiones como se ve en la Figura 50.

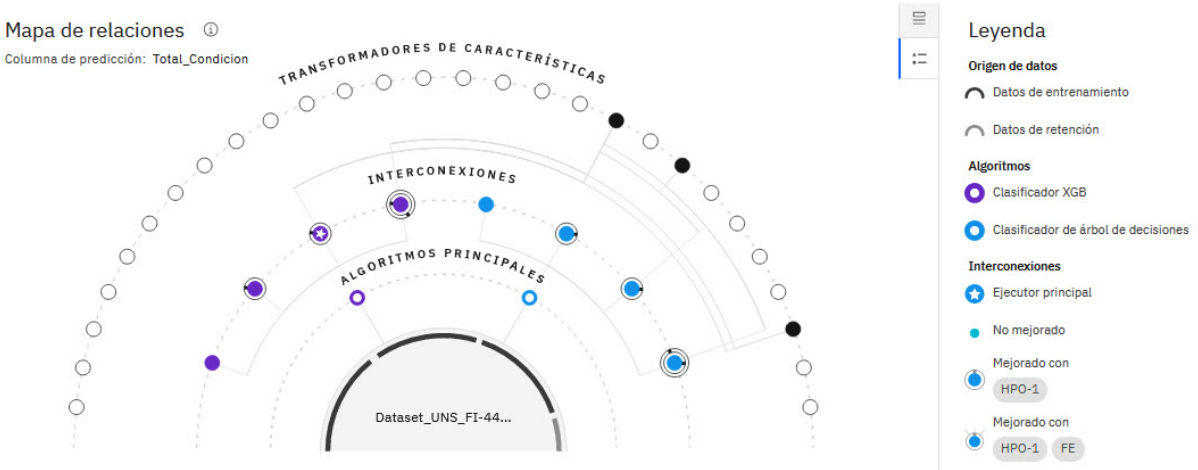


Figura 50 Mapa de relaciones Clasificación XGB y Árbol de decisiones - Dataset UNS FI

3. Implementar propuesta de modelo predictivo

La implementación del modelo se realizó con las librerías de machine Learning de Python: Scikit-learn; para el aprendizaje profundo se usó Tensorflow y previamente para el procesamiento de los datos scipy. Para reducir la perdida (costo) se utilizó: Softmax cross entropy con logits, el modelo se configuró con un ratio de aprendizaje de 0.0001, con mini lotes (minibatchs) de tamaño 32 e inicialmente se trabajó con 1500 iteraciones (epochs). La propagación hacia atrás se consideró el optimizador Adam optimizer.

En un primer lote de experimentos, estos se realizaron con un dataset de la Escuela Profesional de Ingeniería de Sistemas de la UNS con 767 registros, dado que de acuerdo con la data histórica es la tiene el menor porcentaje de graduados y titulados. Se consideró 80% para el conjunto de entrenamiento y el 20% restante para el conjunto del test, luego de ejecutados los experimentos los resultados presentamos a continuación:

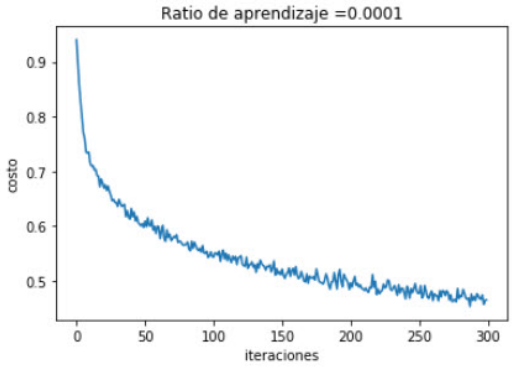


Figura 51 Experimento 03-00 [14-4]

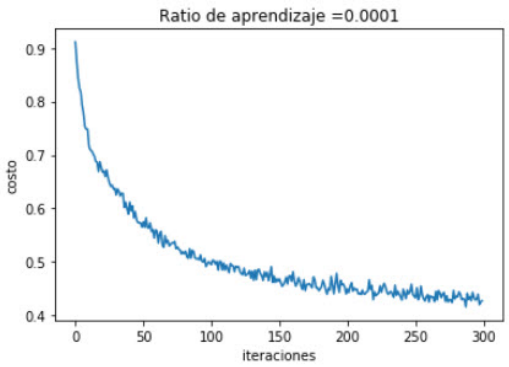


Figura 52 Experimento 03-01 [14-8-4]

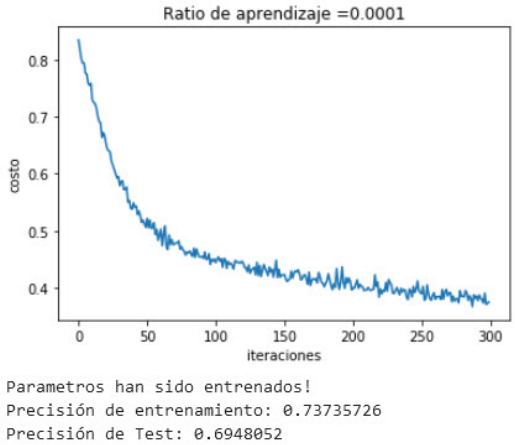


Figura 53 Experimento 03-03 [28-14-8-4]

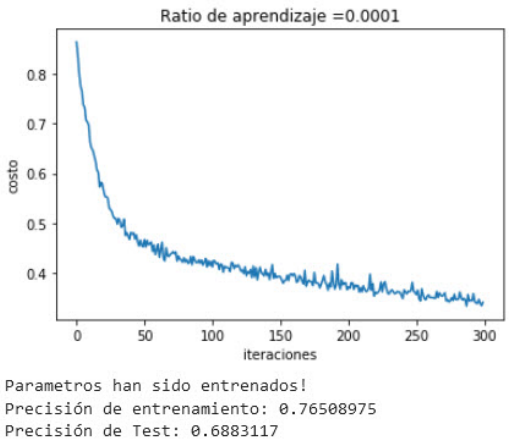


Figura 54 Experimento 03-10 [42-28-14-8-4]

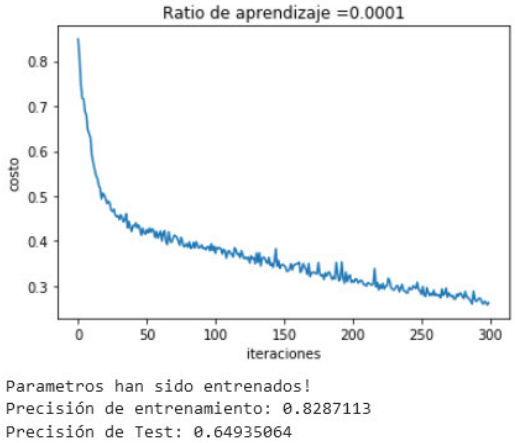


Figura 55 Experimento 03-13 [56-42-28-14-8-4]

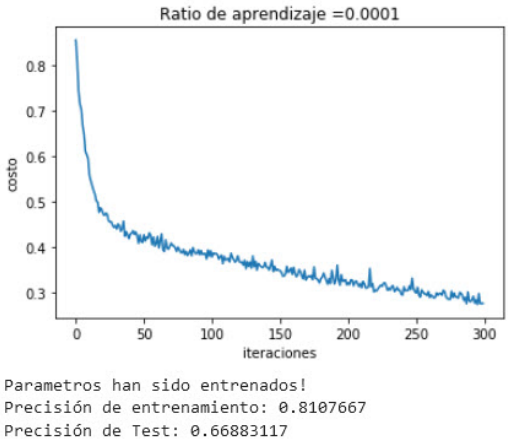


Figura 56 Experimento 03-14 [28-56-42-28-14-8-4]

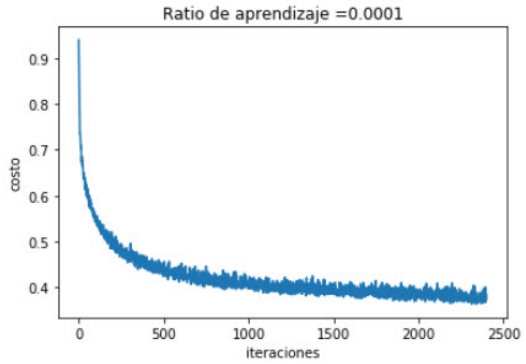
Los resultados de las Figuras 51 al 56, se consolidan en la Tabla 21, donde se incluyen los experimentos adicionales para efectos de determinar cuál es el mejor escenario.

Tabla 21 Exactitud Experimentos con 1500 iteraciones

Experimento	Capas-Nodos	Conjunto de Entrenamiento	Conjunto de Prueba
13	[56,42,28,14,8,4]	82.87%	64.94%
14	[28,56,42,28,14,8,4]	81.08%	66.88%
12	[42,28,28,14,14,4]	80.26%	64.29%
11	[14, 42, 28, 14,8,4]	76.83%	70.13%
10	[42, 28, 14,8,4]	76.51%	68.83%
09	[42, 14,10,4]	75.86%	67.53%
07	[42, 14,8,4]	75.69%	72.08%
08	[42, 28,8,4]	75.69%	68.83%
04	[28, 28, 14,8,4]	75.20%	68.18%
03	[28, 14,8,4]	73.74%	69.48%
05	[28, 20, 14,8,4]	73.57%	66.88%
02	[28,14,4]	72.92%	70.78%
06	[28, 14, 14,8,4]	72.43%	69.48%
01	[14,8,4]	70.96%	68.83%
00	[14,4]	69.00%	62.34%

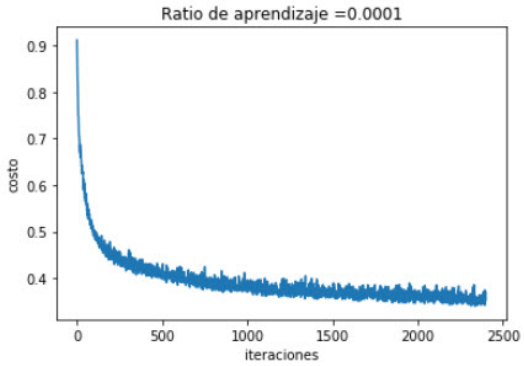
Aun cuando el conjunto de entrenamiento del experimento 13 (6 capas) tiene una exactitud de 82.87%, que es relativamente alto, para el conjunto de prueba tiene una exactitud demasiada baja 64.94%. Ante esta circunstancia, de alta varianza, vamos a aumentar inicialmente el número de iteraciones, entrenar por más tiempo el modelo.

En el segundo lote de los experimentos, se varió los epochs de 1500 a 12000, naturalmente que se fueron realizando pruebas intermedias, hasta llegar a este límite de mejora, que presentamos a continuación.



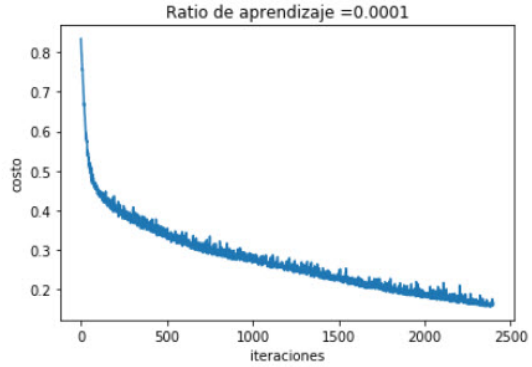

```
Parametros han sido entrenados!
Precisión de entrenamiento: 0.7585644
Precisión de Test: 0.72727275
```

Figura 57 Experimento 03-00 [14-4] 12000 epochs



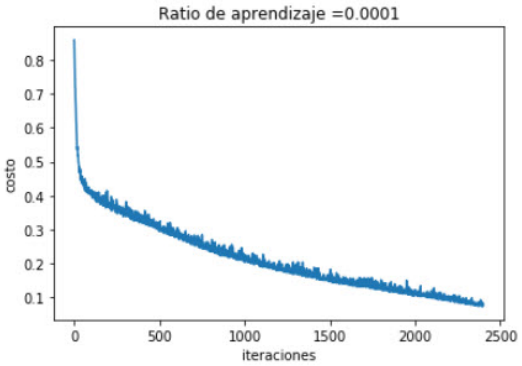

```
Parametros han sido entrenados!
Precisión de entrenamiento: 0.7569331
Precisión de Test: 0.6818182
```

Figura 58 Experimento 03-01 [14-8-4] 12000 epochs



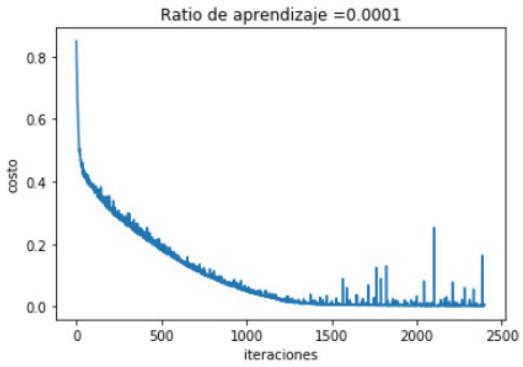

```
Parametros han sido entrenados!
Precisión de entrenamiento: 0.9102773
Precisión de Test: 0.6363636
```

Figura 59 Experimento 03-03 [28-14-8-4] 12000 epochs



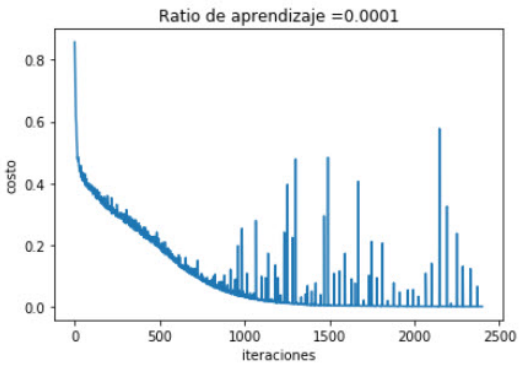

```
Parametros han sido entrenados!
Precisión de entrenamiento: 0.9477977
Precisión de Test: 0.64935064
```

Figura 60 Experimento 03-10 [42-28-14-8-4] 12000 epochs




```
Parametros han sido entrenados!
Precisión de entrenamiento: 1.0
Precisión de Test: 0.5779221
```

Figura 61 Experimento 03-13 [56-42-28-14-8-4] 12000 epochs




```
Parametros han sido entrenados!
Precisión de entrenamiento: 1.0
Precisión de Test: 0.6168831
```

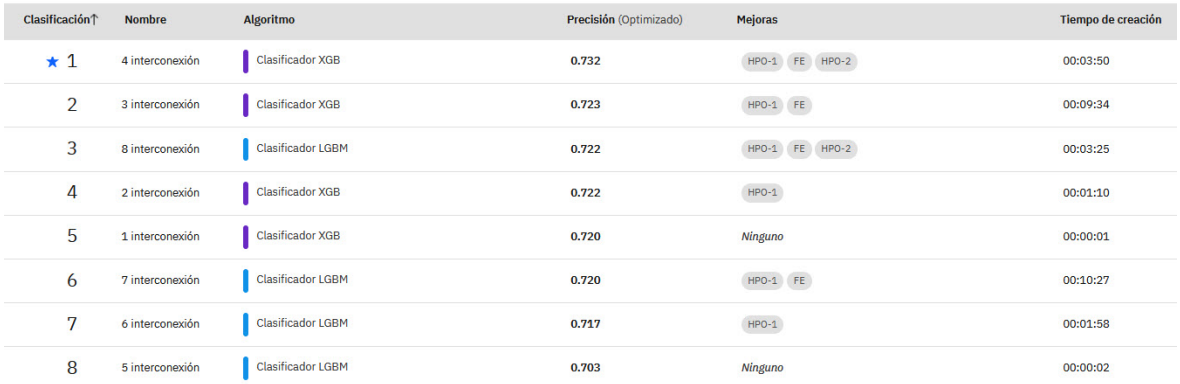
Figura 62 Exp03-14 [28-56-42-28-14-8-4] 12000 epochs

Los resultados de las Figuras 57 al 62, se consolidan en la Tabla 22, donde se incluyen los experimentos adicionales para efectos de determinar si el aumento de iteraciones mejora el aprendizaje.

Tabla 22 Precisión Experimentos Dataset EPISI con 12000 iteraciones

Experimento	Capas-Nodos	Conjunto de Entrenamiento	Conjunto de Prueba
14	[28,56,42,28,14,8,4]	100.00%	61.69%
13	[56,42,28,14,8,4]	100.00%	57.79%
11	[14, 42, 28, 14,8,4]	99.86%	64.94%
12	[42,28,28,14,14,4]	99.84%	61.69%
10	[42, 28, 14,8,4]	94.78%	64.94%
08	[42, 28,8,4]	93.15%	67.53%
04	[28, 28, 14,8,4]	92.50%	61.04%
09	[42, 14,10,4]	91.35%	64.29%
05	[28, 20, 14,8,4]	91.21%	68.18%
03	[28, 14,8,4]	91.03%	63.64%
07	[42, 14,8,4]	90.21%	67.53%
02	[28,14,4]	90.21%	64.94%
06	[28, 14, 14,8,4]	85.15%	62.34%
00	[14,4]	75.86%	72.73%
01	[14,8,4]	75.69%	68.18%

Al revisar los resultados encontramos que en todos los casos ha ocurrido una mejora en la exactitud en el conjunto de entrenamiento, pero con gran sobreajuste en este conjunto de datos, pero la exactitud del conjunto de prueba no ha mejorado haciéndose notoria la alta varianza y probable sesgo de los datos.

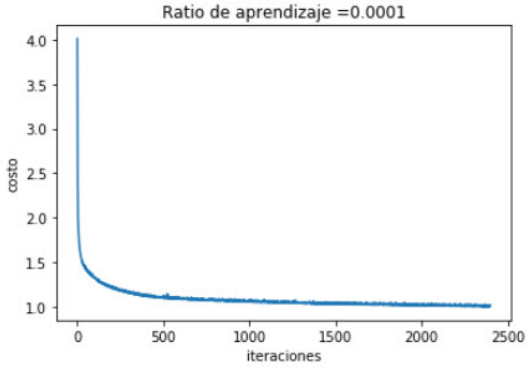


Clasificación↑	Nombre	Algoritmo	Precisión (Optimizado)	Mejoras	Tiempo de creación
★ 1	4 interconexión	Clasificador XGB	0.732	HPO-1 FE HPO-2	00:03:50
2	3 interconexión	Clasificador XGB	0.723	HPO-1 FE	00:09:34
3	8 interconexión	Clasificador LGBM	0.722	HPO-1 FE HPO-2	00:03:25
4	2 interconexión	Clasificador XGB	0.722	HPO-1	00:01:10
5	1 interconexión	Clasificador XGB	0.720	*Ninguno*	00:00:01
6	7 interconexión	Clasificador LGBM	0.720	HPO-1 FE	00:10:27
7	6 interconexión	Clasificador LGBM	0.717	HPO-1	00:01:58
8	5 interconexión	Clasificador LGBM	0.703	*Ninguno*	00:00:02

Figura 63 Ranking de Clasificadores Watson Studio Dataset-UNS

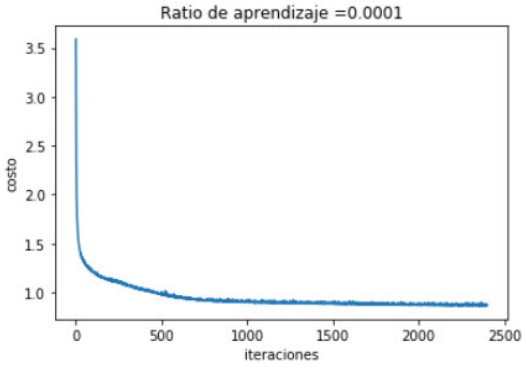
El primer test con el mismo dataset limitado a datos de una Escuela Profesional de la UNS con 767 registros, implementado en la herramienta de experimentación AutoAI de Watson Studio de IBM, la cual requiere que la columna de predicción ŷ contenga las cuatro clases, lo cual difiere con modelo en Tensorflow que trabaja con cuatro vectores one-hot que representan las clases en estudio, se obtuvieron los resultados de la Figura 63, donde el Clasificador XGB que incluye las mejoras HPO-1, FE y HPO-2 el que ofrece la mejor precisión de entrenamiento: 73.2%, los siete restantes están por debajo de esa precisión. Y por cierto también se encuentran por debajo de las precisiones de la red neuronal de aprendizaje profundo de la Tabla 21.

Por lo visto anteriormente de tener aún una baja precisión de entrenamiento, con la sola data de una Escuela Profesional de la UNS; es necesario que se incremente los ejemplos de entrenamiento, por lo que ampliaremos la data a 4 Escuelas Profesionales de la UNS que tienen características similares, para ello utilizaremos el dataset ampliado: Dataset UNS FI con 3529 registros. Los cuáles serán probados en los 15 modelos prediseñados previamente con 12000 epochs.



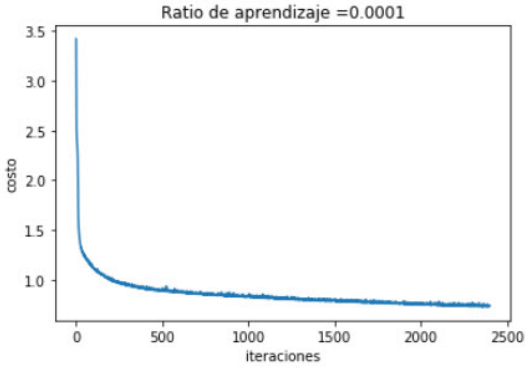

```
Parametros han sido entrenados!
Precisión de entrenamiento: 0.8682253
Precisión de Test: 0.84560907
```

Figura 64 Experimento 03-00-FI [14-4]



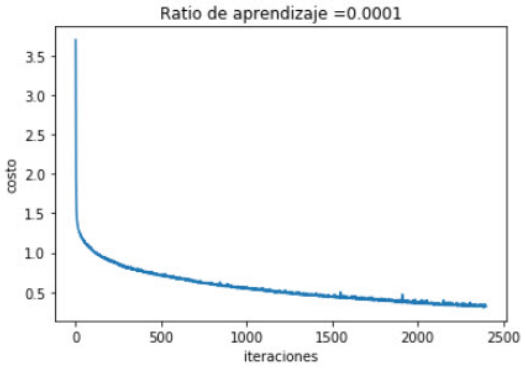

```
Parametros han sido entrenados!
Precisión de entrenamiento: 0.88699967
Precisión de Test: 0.8470255
```

Figura 65 Experimento 03-01-FI [14-8-4]




```
Parametros han sido entrenados!
Precisión de entrenamiento: 0.90187746
Precisión de Test: 0.8385269
```

Figura 66 Experimento 03-03-FI [28-14-8-4]




```
Parametros han sido entrenados!
Precisión de entrenamiento: 0.95997167
Precisión de Test: 0.786119
```

Figura 67 Exp. 03-10-FI [42-28-14-8-4]

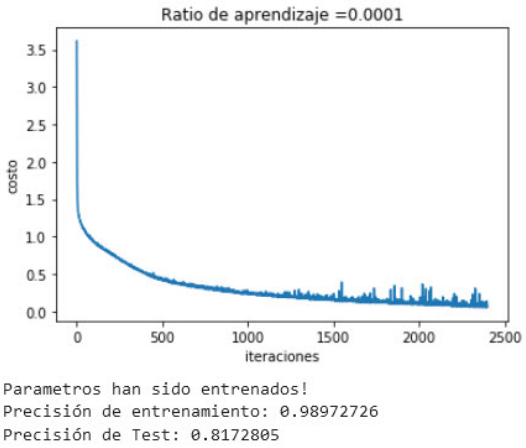


Figura 68 Exp. 03-13-FI [56-42-28-14-8-4]

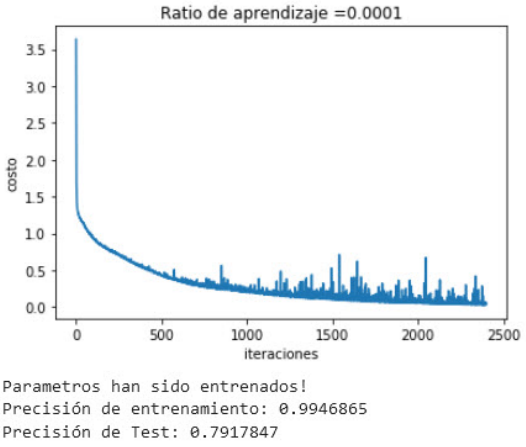


Figura 69 Exp03-14-FI [28-56-42-28-14-8-4]

De igual manera consolidamos los resultados de las Figuras 64 al 69 con los modelos adicionales que completan 15 y que se ve en la Tabla 23

Tabla 23 Precisión Experimentos Dataset UNS FI con 12000 iteraciones

Experimento	Capas-Nodos	Conjunto de Entrenamiento	Conjunto de Prueba
14	[28,56,42,28,14,8,4]	99.47%	79.18%
13	[56,42,28,14,8,4]	98.97%	81.73%
12	[42,28,28,14,14,4]	96.95%	77.05%
10	[42, 28, 14,8,4]	96.00%	78.61%
11	[14, 42, 28, 14,8,4]	93.59%	80.31%
04	[28, 28, 14,8,4]	93.27%	81.16%
08	[42, 28,8,4]	93.27%	80.45%
05	[28, 20, 14,8,4]	92.84%	83.00%
09	[42, 14,10,4]	92.49%	82.01%
07	[42, 14,8,4]	91.82%	81.16%
06	[28, 14, 14,8,4]	91.57%	83.14%
02	[28,14,4]	90.72%	83.14%
03	[28, 14,8,4]	90.19%	83.85%
01	[14,8,4]	88.70%	84.70%
00	[14,4]	86.82%	84.56%

Con más datos se tiene una notable mejoría.

De igual manera con este Dataset UNS FI, lo implementaremos en la herramienta de experimentación AutoAI de Watson Studio de IBM,

Clasificación ↑	Nombre	Algoritmo	Precisión (Optimizado)	Mejoras	Tiempo de creación
★ 1	3 interconexión	Clasificador XGB	0.871	HPO-1 FE	00:00:42
2	4 interconexión	Clasificador XGB	0.871	HPO-1 FE HPO-2	00:01:46
3	1 interconexión	Clasificador XGB	0.867	*Ninguno*	00:00:02
4	2 interconexión	Clasificador XGB	0.867	HPO-1	00:00:34
5	7 interconexión	Clasificador de árbol de decisiones	0.806	HPO-1 FE	00:00:17
6	8 interconexión	Clasificador de árbol de decisiones	0.806	HPO-1 FE HPO-2	00:00:05
7	5 interconexión	Clasificador de árbol de decisiones	0.798	*Ninguno*	00:00:01
8	6 interconexión	Clasificador de árbol de decisiones	0.798	HPO-1	00:00:02

Figura 70 Ranking de Clasificadores Watson Studio Dataset-UNS-FI

Para este segundo dataset ampliado a datos de cuatro Escuelas Profesionales de la UNS con 3529 registros, se obtuvieron los resultados de la Figura 70, donde el Clasificador XGB que incluye las mejoras HPO-1 y FE el que ofrece la mejor precisión de entrenamiento: 87.1%, mejorando con respecto al dataset-UNS. Con 767 registros, asimismo también se encuentran por debajo de las precisiones de la red neuronal de aprendizaje profundo de la Tabla 23.

4. Variable de mayor influencia en la predicción

Teniendo en cuenta que el Dataset UNS FI tiene más ejemplos de entrenamiento y nos está ofreciendo mejores resultados de predicción, sobre la base de lo trabajado en la herramienta de experimentación AutoAI de Watson Studio de IBM que indica que el clasificador XGB es el que mejor exactitud tiene, lo mismo nos dice el Dataset UNS con datos de una sola Escuela Profesional.

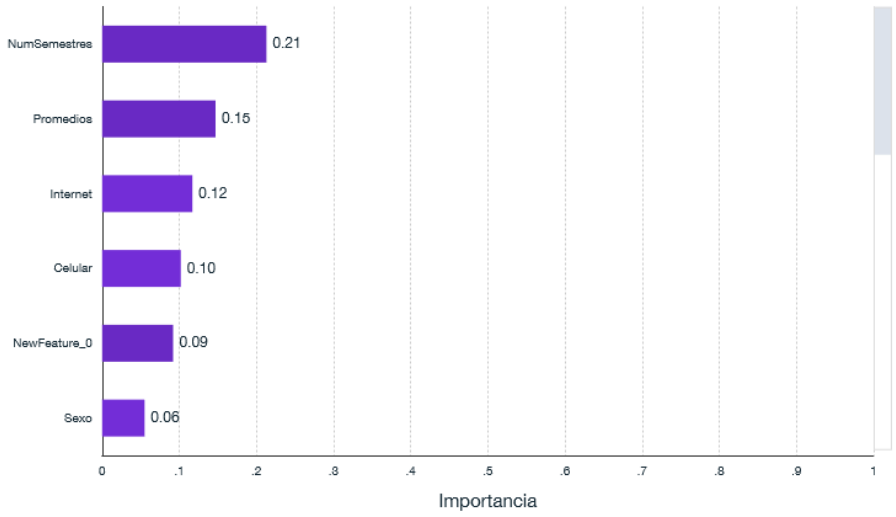


Figura 71 Influencia de variables en predicción con datos EPISI

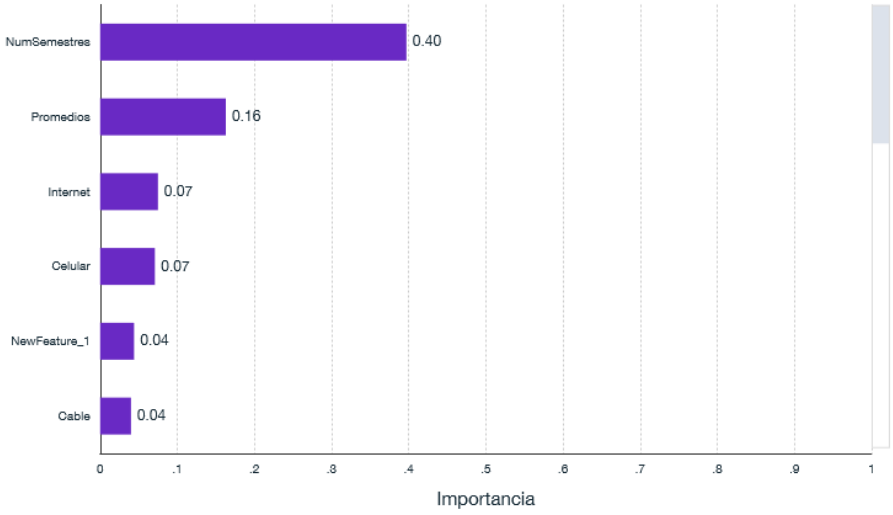


Figura 72 Influencia de variables en predicción con datos FI

De las Figuras 71 y 72, podemos ver que son cuatro características que más influyen en la predicción, Numero de semestres, promedio, internet, celular. La que está mejor posicionada en ambos casos es Numero de semestres que alcanza una ratio de 0.40 con el Dataset UNS FI, el cual contiene más ejemplos de entrenamiento.

5. Validación del modelo predictivo

Validaremos el modelo predictivo sobre la base de precisión o exactitud de la predicción tanto en el conjunto de entrenamiento como en el de prueba, para ello comparemos los diversos modelos examinados que determina qué modelo nos proporciona mejores resultados.

En la Tabla 24, observamos la exactitud de los diversos modelos de redes neuronales en tanto sus conjuntos de entrenamiento como de prueba, en el cual se ha incluido el Clasificador XGB que incluye las mejoras HPO-1, FE y HPO-2 que nos proporcionó la herramienta de experimentación AutoAI de Watson Studio de IBM, todos estos modelos probados con el mismo Dataset UNS FI con 3529 registros, encontrándose que el modelo del experimento 13, el cual tiene 6 capas y tantas neuronas en cada capa como sigue en orden de sucesión [56,42,28,14,8,4], con 98.97% de exactitud en el conjunto de entrenamiento y 81.73% en conjunto de prueba, y es el cual tiene una menor diferencia de precisión entre ambos conjuntos con 14,24 puntos porcentuales que reduce la impresión de sobreajuste, a diferencia del modelo del experimento 14 el cual tiene 7 capas y neuronas por capa sucesivas igual a [28,56,42,28,14,8,4] quien tiene una diferencia de 20.29 puntos porcentuales que agranda la impresión de un posible sobreajuste.

Tabla 24 Precisión de los diversos modelos con Dataset UNS FI

Experimento	Conjunto de Entrenamiento	Conjunto de Prueba
14	99.47%	79.18%
13	98.97%	81.73%
12	96.95%	77.05%
10	96.00%	78.61%
11	93.59%	80.31%
04	93.27%	81.16%
08	93.27%	80.45%
05	92.84%	83.00%
09	92.49%	82.01%
07	91.82%	81.16%
06	91.57%	83.14%
02	90.72%	83.14%
03	90.19%	83.85%
01	88.70%	84.70%
WS-XGB	87.10%	83.60%
00	86.82%	84.56%

6. Contrastación de la hipótesis

A continuación, vamos a demostrar que un modelo predictivo basado en Machine Learning ayuda a mejorar el seguimiento académico de estudiantes universitarios en las Escuelas Profesionales de Ingeniería de la Universidad Nacional del Santa.

Dado que es un diseño cuasiexperimental, la muestra seleccionada ha sido dirigida a la cual se le ha aplicado el estímulo del Modelo predictivo basado en Machine Learning.

En la Tabla 6 hemos presentado los porcentajes de graduados y titulados en las 04 escuelas de ingeniería de la muestra de estudiantes de la UNS, la que resumiremos en la Tabla 25, estos valores obtenidos de los datos históricos iniciales forman parte del pre-test, antes de la aplicación del estímulo, por lo que el resultado de la predicción obtenido después de aplicado el estímulo, será el valor con el cual contrastaremos y diremos si realmente se mejora el seguimiento al estudiante, de la Tabla 24 extraemos la precisión de predecir si el estudiante logra graduarse o titularse y la mostramos en la Tabla 26, siendo

la precisión del conjunto de entrenamiento 98.87% y la conjunto de prueba 81.73%. Para efectos de la contrastación lo haremos con el porcentaje de precisión del conjunto de prueba, que actúa para cualquiera de las clases graduación o titulación, que son parte de la función objetivo del modelo de predicción.

Tabla 25 Porcentaje de Graduados y Titulados

Escuela	% Graduados	% Titulados
Sistemas	31.60	23.26
Energía	38.53	27.32
Agroindustria	52.84	41.36
Civil	44.46	31.96

Tabla 26 Precisión del mejor modelo de predicción

Experimento	Conjunto de Entrenamiento	Conjunto de Prueba
13	98.97%	81.73%

En las Tabla 27 y 28, presentamos las diferencias entre los porcentajes de la precisión del conjunto de prueba, contra el de graduados y titulados delas 04 Escuelas Profesionales de Ingeniería, donde observamos que tal diferencia sería la mejora que proporciona la intervención del modelo predictivo con machine learning, de saber quiénes se gradúan y quienes se titulan y sobre los cuales se puede intervenir para poder darles el soporte adecuado.

Tabla 27 Porcentaje de mejora para graduados

Escuela	% Precisión conjunto de prueba	% Graduados	Mejora (Diferencia)
Sistemas	81.73	31.60	50.13
Energía	81.73	38.53	43.20
Agroindustria	81.73	52.84	28.89
Civil	81.73	44.46	37.27

Tabla 28 Porcentaje de mejora para titulados

Escuela	% Precisión conjunto de prueba	% Titulados	Mejora (Diferencia)
Sistemas	81.73	23.26	58.47
Energía	81.73	27.32	54.41
Agroindustria	81.73	41.36	40.37
Civil	81.73	31.96	49.77

Se acompaña en el Anexo 2, test de predicción con datos reales al azar y dirigidos que verifica las predicciones de acuerdo con la variable objetivo.

DISCUSION

En cuanto a las variables o características que inciden en la persistencia o abandono estudiantil, seleccionamos 12 de orden demográfico y 02 de tipo académico, aun cuando desde ya (Spady, 1970), (Tinto, 1975), (Bean, 1985), (Ethington, 1990) coinciden en que el desempeño académico y los antecedentes familiares tiene amplia influencia, como lo dice (Chanlekha & Niramitranon, 2018) es uno de los mayores desafíos en el desarrollo de un modelo de predicción la cantidad limitada de información disponible para ser utilizada como atributos de los estudiantes, coincidimos con (Hellas et al., 2018) en que la data demográfica, el género y la edad ayudan a la predicción de la retención o el abandono y que para el grado o titulación el desempeño de los cursos es vital. Nosotros hemos tenido que prescindir de datos como el colegio de procedencia porque era un atributo donde la data no fue coleccionada adecuadamente, demasiados faltantes, de manera similar campo que existía en la recogida de data pero que no era obligatorio el estudiante trabaja tuvo que ser descartado, este era un campo que se incluía en el antecedente familiar. En definitiva, tenemos que trabajar de acuerdo con el contexto de la realidad inmediata sobre el cual modelar, pues no existe juego de características valido para todas las realidades.

Para el diseño e implementación del modelo predictivo se utilizó las coincidencias entre (Musso et al., 2020), (Vijayalakshmi & Vengatachalapathy, 2019), quienes obtuvieron precisión máxima con redes neuronales de perceptron multicapa y capa de salida softmax, para predecir el desempeño de los alumnos y colegir si abandona o culminan satisfactoriamente, aun cuando sus modelos implementados tenían 2 o 3 capas ocultas, y dataset con alrededor de 600 registros, acusan buenos resultados en la predicción que superan el 80% en el conjunto de entrenamiento, se resalta que (Ojha et al., 2017) con un modelo de 3 capas, con 256 neuronas por capa, y con 16174 registros logra determinar el tiempo de graduación entre 0, 1 y 2 años logrando un ratio de precisión de hasta 88%. Nosotros enfocamos nuestro modelo de manera diferente y alternativa para predecir si el estudiante lograba culminar sus estudios (egresar), obtener el grado de bachiller, obtener el título profesional o abandonar y quedar en la condición de estudiante sin fecha de término, para ello modelamos una red neuronal de aprendizaje profundo, inicialmente de 02 capas, con 767 registros, en vista de tener baja precisión tanto

en el conjunto de entrenamiento como de prueba, se experimentó aleatoria y sucesivamente con 03, 04, 05, 06 y 07 capas, variando el número neuronas por capa, como se puede ver en la Tabla 19, alternativamente con el mismo conjunto de datos se modelo en la herramienta de experimentación AutoAI de Watson Studio de IBM la cual proporcionó dos algoritmos de clasificación XGB y LGBM, ambos también con baja precisión con máximo de 73.2% en el conjunto de entrenamiento. Fue necesario aumentar los datos, es por ello que se amplió a 3529 registros, mejorando los resultados, con la herramienta de experimentación AutoAI de Watson Studio de IBM nos devolvió el clasificador XGB por encima del clasificador árbol de decisiones, con una precisión de 87.1% en el conjunto de entrenamiento, probando con los mismos modelos de la primera parte nos quedó el modelo de 06 capas con la secuencia de nodos [56,42,28,14,8,4] como el seleccionado con 98.97% de precisión en el conjunto de entrenamiento y 81.73% de precisión en el conjunto de prueba.

Nos sumamos a ese espectro heterogéneo que determina la variable o categoría que más influencia en una predicción del desempeño estudiantil, (Oancea et al., 2013), (Jia & Mareboyana, 2014), (Bendangnuksung & Prabu, 2018), (Forero Zea et al., 2019) nos dijeron el promedio de secundaria, el género; (Chanlekha & Niramitranon, 2018) nos dice que los cursos que tienen prerrequisito, para (Vijayalakshmi & Vengatachalapathy, 2019) nos dicen que sorprendentemente son los días de ausencia de los estudiantes, para (Ruiz Palacios, 2018) la edad del estudiante, En función al conjunto de datos empleados, la variable que más influyó en la predicción fue el número de semestres, y esto tiene sentido dado que culminar una carrera debería tomar en forma mínima 10 semestres académicos pero el rango en esta característica se amplía hasta estudiantes con 34 semestres para concluir o no su carrera.

La evaluación del modelo predictivo, como ocurre con (Chanlekha & Niramitranon, 2018) que investigaron varias combinaciones de atributos, así como diferentes configuraciones de parámetros del modelo e informaron y analizaron el rendimiento de predicción de los mejores modelos, (Ojha et al., 2017) también evaluó sus modelos con el rendimiento de predicción, además de la métrica recall y f1-score, (Vijayalakshmi & Vengatachalapathy, 2019) evaluaron con el rendimiento de la precisión, la perdida y la matriz de confusión, pero al final siempre concluyen con

el rendimiento de la precisión, que en nuestro caso con el conjunto de datos ampliados a las 4 Escuelas Profesionales de la Facultad de Ingeniería con 3529 registros, el modelo aprendizaje profundo, la red neuronal artificial de 6 capas con la secuencia de nodos [56,42,28,14,8,4] como el seleccionado con 98.97% de precisión en el conjunto de entrenamiento y 81.73% de precisión en el conjunto de prueba.

La foto inicial de la data histórica nos dice de los bajos porcentajes de graduación y titulación en la UNS, con el modelo de predicción nos permitiría mejorar significativamente en rangos que van de 28.89% a 50.13% para el caso de los graduados, y para titulados la mejoría estaría en el rango de 40.37% a 58.47%, la diferencia de mejoría es mayor porque es menor el número de titulados. De darse el uso de nuestra propuesta de modelo predictivo con machine learning, y tratar de alcanzar ese 65% de graduación que es lo que está ocurriendo en México, Perú y EEUU de acuerdo con (Ferreyra et al., 2017) trabajo del Banco Mundial de ese año. Pero no solo se trata de predecir si el alumno graduará o se titulará, también es necesario adoptar las medidas de acompañamiento como por ejemplo la buena práctica del (Ministerio de Educación Nacional Colombia, 2015) los cuales tiene un modelo de gestión de permanencia y graduación estudiantil.

CAPÍTULO V Conclusiones y Recomendaciones

CONCLUSIONES

1. El aplicar el modelo de predicción basado en machine Learning, logra mejorar el seguimiento académico de los estudiantes de las Escuelas Profesionales de Ingeniería de la UNS, en rangos que van desde 28.89% (Tabla 27) a 58.47% (Tabla 28) por escuela ya sea graduación o titulación, con esto tendremos certeza de quienes pudieran fallar, pero es necesario se implemente un sistema de gestión de permanencia y graduación estudiantil, para poder concretar estos resultados.

2. Del amplio espectro de variables o características que los diversos autores proponen, desde el punto de vista pedagógico y de los mismos modeladores de inteligencia artificial, dada nuestra realidad, seleccionamos 12 características de orden demográfico y 02 de tipo académico (Tabla 19), por las limitaciones para adquirir data adecuada y buena, lo que nos obliga a adaptarnos al contexto del entorno en estudio.

3. Se logró diseñar un modelo predictivo con machine Learning y aprendizaje profundo a través de una red neuronal de perceptron multicapa con capa de salida softmax, siendo el diseño de 06 capas con la secuencia de nodos [56,42,28,14,8,4] el que otorgó mejor porcentaje precisión en el conjunto de entrenamiento y en el de prueba (Tabla 23), comparativamente se probó con el clasificador XGB y el clasificador árbol de decisiones.

4. Se implementó el modelo de red neuronal de perceptron multicapa con capa de salida softmax, antes mencionado (Capitulo Resultados, numeral 3, página 69), con el cual se puede identificar tanto a los estudiantes con alto riesgo de abandono que no lograran culminar y obtener el diploma correspondiente, con una precisión de 81.73% en el conjunto de prueba y 98.97% en el conjunto de entrenamiento, por encima del clasificador XGB y el clasificador árbol de decisiones comprobados con el mismo dataset.

5. Se determinó que las variables que mayor influencia ofrecían a la predicción eran las variables académicas número de semestres y promedios, y de las demográficas fueron el internet y el celular (Capitulo Resultados, numeral 4, página 76).
6. La evaluación del modelo predictivo se dio a través del comparativo con diseños de red neuronal de perceptron multicapa con capa de salida softmax que partieron con 02 capas hasta 07 capas y con combinaciones de cantidades de neuronas por capas, a través del porcentaje de precisión en la predicción en el conjunto de entrenamiento como en el de prueba, adicionalmente se evaluó contra los resultados del Clasificador XGB y el clasificador árbol de decisiones, obteniéndose los mejores resultados con el modelo aprendizaje profundo, la red neuronal artificial de 6 capas con la secuencia de nodos [56,42,28,14,8,4] como el seleccionado con 98.97% de precisión en el conjunto de entrenamiento y 81.73% de precisión en el conjunto de prueba (Capitulo Resultados, numeral 5, página 77).

RECOMENDACIONES

1. Se debe implementar en cada escuela el presente modelo de predicción, para poder prevenir y orientar a los estudiantes que se detecte el perfil de no conseguir egresar, graduarse o titularse.
2. Se debe acompañar lo anterior con el mejoramiento de los reglamentos existentes e implementar atendiendo las buenas prácticas (Valenzuela & Pérez, 2012) sistemas de seguimiento del desempeño estudiantil, para ello el Vicerrectorado Académico debe implementar una Oficina Técnica en DEDA que coordine con La Oficina de Tecnologías y las Escuelas profesionales.
3. Dada la dificultad de tener poquísimos atributos o características para poder implementar el modelo de predicción se debe mejorar la adquisición de datos de los estudiantes, agregando data del entorno familiar, colegio de procedencia, estados sentimentales y compromisos familiares, para verificar si contextualmente si tienen la influencia debida que es predicada por los diversos autores.

4. Desde ya nos estamos planteando como futuro trabajo escalar el presente modelo de predicción con modelos secuenciales a través de redes neuronales recurrentes que hagan uso de la data histórica del promedio de notas semestrales que en muchos casos se han alargado hasta 34 semestres, para poder tener mayor certeza de predicción con los estudiantes que recién inician la carrera y emplear las matrices de confusión que enriquecerán las predicciones.

REFERENCIAS BIBLIOGRÁFICAS

Abarca Rodríguez, A., & Sánchez Vindas, A. (2005). La deserción estudiantil en la educación superior: El caso de la Universidad de Costa Rica. *Revista Electrónica Actualidades Investigativas En Educación*, *5*. https://doi.org/10.15517/aie.v5i4.9186

Anuradha, C., & T, V. (2015). A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance. *Indian Journal of Science and Technology*, *8*, 974–6846. https://doi.org/10.17485/ijst/2015/v8i15/74555

Arco-Tirado, J. L., Fernández-Martín, F. D., & Fernández-Balboa, J.-M. (2011). The impact of a peer-tutoring program on quality standards in higher education. *Higher Education*, *62*(6), 773–788. https://doi.org/10.1007/s10734-011-9419-x

Bean, J. P. (1985). Interaction Effects Based on Class Level in an Explanatory Model of College Student Dropout Syndrome. *American Educational Research Journal*, *22*(1), 35–64. https://doi.org/10.3102/00028312022001035

Bendangnuksung, & Prabu, D. (2018). *Students ' Performance Prediction Using Deep Neural Network*. 1171–1176.

Berea, A. (2017). Predictive Analytics. In L. A. Schintler & C. L. McNeely (Eds.), *Encyclopedia of Big Data* (pp. 1–5). https://doi.org/10.1007/978-3-319-32001-4_170-1

Brownlee, J. (2018). *XGBoost With Python* (1.1). Jason Brownlee.

Calua Torres, J. (2016). *Potencia Predictiva de Variables Académicas en el Rendimiento Académico de Estudiantes Universitarios del Primer Ciclo-2015-1. Caso de la Universidad Privada del Norte-Cajamarca* (Universidad Nacional de Cajamarca). Retrieved from http://repositorio.unc.edu.pe/handle/UNC/1348

Castañeda Castañeda, R. S. (2013). *Factores asociados a la deserción de*

estudiantes universitarios (Universidad San Martín de Porras). Retrieved from https://hdl.handle.net/20.500.12727/1172

Castaño, E., Gallón, S., Gómez, K., & Vásquez, J. (2004). Deserción estudiantil universitaria: una aplicación de modelos de duración. *Lecturas de Economía*, (60), 39–65. Retrieved from https://www.redalyc.org/articulo.oa?id=155217798002

Chanlekha, H., & Niramitranon, J. (2018). Student Performance Prediction Model for Early-Identification of at-Risk Students in Traditional Classroom Settings. *Proceedings of the 10th International Conference on Management of Digital EcoSystems*, 239–245. https://doi.org/10.1145/3281375.3281403

Chollet, F. (2017). *Deep Learning with Python* (1st ed.). USA: Manning Publications Co.

CLABES. (2019). Noveno Congreso Latinoamericano sobre el abandono en la educación superior. Retrieved from https://www.urosario.edu.co/CLABES/inicio/

Colvin, J. W. (2015). *Peer Mentoring and Tutoring in Higher Education BT - Exploring Learning & Teaching in Higher Education* (M. Li & Y. Zhao, Eds.). https://doi.org/10.1007/978-3-642-55352-3_9

Databricks. (2020). Databricks Supported Instance Types. Retrieved from https://databricks.com/product/aws-pricing/instance-types

DataRobot. (2020). DataRobot Review. Retrieved from https://reviews.financesonline.com/p/datarobot/

Donoso Díaz, S., & Arias R., Ó. (2010). *Retención de estudiantes y éxito académico en la educación superior análisis de buenas prácticas*. Retrieved from https://www.google.com.pe/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwipy6ipgovwAhX7TjABHeJoBnUQFjABegQIBhAD&url=https%3A%2F%2Fwww.cned.cl%2Ffile%2F1927%2Fdownload%3Ftoken%3Dug9_b9uA&usg=AOvVaw2DCa_n1O4ht95ylUD6Jo_M

Donoso, S., & Schiefelbein, E. (2007). Análisis de los modelos explicativos de retención de estudiantes en la universidad una visión desde la desigualdad social. *Estudios Pedagógicos (Valdivia)*, *33*, 7–27. Retrieved from https://scielo.conicyt.cl/scielo.php?script=sci_arttext&pid=S0718-07052007000100001&nrm=iso

Dresel, M., & Rindermann, H. (2011). Counseling University Instructors Based on Student Evaluations of Their Teaching Effectiveness: A Multilevel Test of its Effectiveness Under Consideration of Bias and Unfairness Variables. *Research in Higher Education*, *52*(7), 717–737. https://doi.org/10.1007/s11162-011-9214-7

Ethington, C. A. (1990). A psychological model of student persistence. *Research in Higher Education*, *31*(3), 279–293. https://doi.org/10.1007/BF00992313

Fernández, S. G. (2018). Rendimiento Académico en Educación Superior: Desafíos para el Docente y Compromiso del Estudiante. *Revista Científica de La UCSA*, *5*, 55–63. Retrieved from http://scielo.iics.una.py/scielo.php?script=sci_arttext&pid=S2409-87522018000300055&nrm=iso

Ferreyra, M. M., Avitabile, C., Álvarez, J., Paz, F. H., & Urzúa, S. (2017). *Momento Decisivo: La Educación Superior en América Latina y el Caribe*.

Feurer, M., & Hutter, F. (2019). Hyperparameter Optimization. In F. Hutter, L. Kotthoff, & J. Vanschoren (Eds.), *Automated Machine Learning* (pp. 3–38). https://doi.org/https://doi.org/10.1007/978-3-030-05318-5_1

Fishbein, M. A., & Ajzen, I. (1975). *Belief, attitude, intention and behaviour: An introduction to theory and research* (Vol. 27).

Forero Zea, L. D., Piñeros Reina, Y. F., & Rodríguez Molano, J. I. (2019). *Machine Learning for the Identification of Students at Risk of Academic Desertion BT - Learning Technology for Education Challenges* (L. Uden, D. Liberona, G. Sanchez, & S. Rodríguez-González, Eds.). Cham: Springer International Publishing.

Gartner. (2021). Magic Quadrant for Data Science and Machine Learning

Platforms. Retrieved from Data Science and Machine Learning (ML) Platforms Reviews and Ratings website: https://www.gartner.com/reviews/market/data-science-machine-learning-platforms

González Nespereira, C., Elhariri, E., El-Bendary, N., Fernández Vilas, A., & Díaz Redondo, R. P. (2016). *Machine Learning Based Classification Approach for Predicting Students Performance in Blended Learning BT - The 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), November 28-30, 2015, Beni Suef, Egypt* (T. Gaber, A. E. Hassanien, N. El-Bendary, & N. Dey, Eds.). Cham: Springer International Publishing.

Google Cloud. (2020). Acelera tu transformación con Google Cloud. Retrieved from https://cloud.google.com/?hl=es

H2O.ai. (2020). H2O.ai is the Open Source Leader in AI and ML. Retrieved from https://www.google.com.pe/search?q=H2O.ai&biw=1920&bih=938&ei=iqx_YK_NGaayggfu3aTwDw&oq=H2O.ai&gs_lcp=Cgdnd3Mtd2l6EAMyAggAMgIIADICCAAyAggAMgQIABAeMgQIABAeMgQIABAeMgQIABAeMgQIABAeUOKdBVjinQVg6aEFaAFwAngAgAGhAYgBrQKSAQMwLjKYAQCgAQGgAQKqAQdnd3Mtd2l6s

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V, Gutica, M., Hynninen, T., … Liao, S. N. (2018). Predicting Academic Performance: A Systematic Literature Review. *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, 175–199. https://doi.org/10.1145/3293881.3295783

Hernández Herrera, C. A. (2016). Diagnóstico del rendimiento académico de estudiantes de una escuela de educación superior en México. *Revista Complutense de Educación*, *27*(3 SE-Artículos). https://doi.org/10.5209/rev_RCED.2016.v27.n3.48551

Jia, J.-W., & Mareboyana, M. (2014). *Predictive Models for Undergraduate Student Retention Using Machine Learning Algorithms BT - Transactions on Engineering Technologies* (H. K. Kim, S.-I. Ao, & M. A. Amouzegar, Eds.). Dordrecht: Springer Netherlands.

Kamath, U., Liu, J., & Whitaker, J. (2019). Basics of Deep Learning. In U. Kamath, J. Liu, & J. Whitaker (Eds.), *Deep Learning for NLP and Speech Recognition* (pp. 141–201). https://doi.org/10.1007/978-3-030-14596-5

Kingma, D. P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations, ICLR 2015*.

KNIME. (2020). KNIME Analytics Platform for AWS. Retrieved from https://aws.amazon.com/marketplace/pp/KNIME-KNIME-Analytics-Platform-for-AWS/B071ZNNLC6

Kostopoulos, G., Lipitakis, A.-D., Kotsiantis, S., & Gravvanis, G. (2017). *Predicting Student Performance in Distance Higher Education Using Active Learning BT - Engineering Applications of Neural Networks* (G. Boracchi, L. Iliadis, C. Jayne, & A. Likas, Eds.). Cham: Springer International Publishing.

Kučak, D., Juricic, V., & Đambić, G. (2018). *Machine Learning in Education - a Survey of Current Research Trends*. https://doi.org/10.2507/29th.daaam.proceedings.059

MathWorks. (2020). Trial gratis MATLAB. Retrieved from https://la.mathworks.com/campaigns/products/trials.html

Microsoft Azure. (2020). Configure productos de Azure y calcule sus costos. Retrieved from https://azure.microsoft.com/es-es/pricing/calculator/

Ministerio de Educación Nacional Colombia. (2015). *Guía para la implementación del Modelo de Gestión de permanencia y graduación estudiantil en Instituciones de Educación Superior*. Imprenta Nacional de Colombia.

Mori Sánchez, M. del P. (2012). DESERCIÓN UNIVERSITARIA EN ESTUDIANTES DE UNA UNIVERSIDAD PRIVADA DE IQUITOS. *Revista Digital de Investigación En Docencia Universitaria*, *6*(1 SE-Artículos de investigación). https://doi.org/10.19083/ridu.6.42

Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python*. O'Reilly Media, Inc.

Musso, M. F., Hernández, C. F. R., & Cascallar, E. C. (2020). Predicting key

educational outcomes in academic trajectories: a machine-learning approach. *Higher Education (00181560)*, *80*(5), 875–894. https://doi.org/10.1007/s10734-020-00520-7

Nandeshwar, A., Menzies, T., & Nelson, A. (2011). Learning patterns of university student retention. *Expert Systems with Applications*, *38*(12), 14984–14996. https://doi.org/https://doi.org/10.1016/j.eswa.2011.05.048

Oancea, B., Dragoescu, R., & Ciucu, S. (2013, April 26). *Predicting students' results in higher education using neural networks*. 190–193. Jelgava, Latvia.

Ocaña Fernández, Y. (2011). Variables académicas que influyen en el rendimiento académico de los estudiantes universitarios. *Investigación Educativa*, *15*(27), 165–179. Retrieved from https://revistasinvestigacion.unmsm.edu.pe/index.php/educa/article/view/6473

Ojha, T., Heileman, G., Martinez-Ramon, M., & Slim, A. (2017). *Prediction of graduation delay based on student performance*. 3454–3460. https://doi.org/10.1109/IJCNN.2017.7966290

Paredes Esparza, R. (Universidad A. B., Aguirre Larrain, F. (Universidad A. B., & Quense Abarzúa, M. de los Á. (Universidad A. B. (2017). *Modelo de retención universitaria: desafíos y oportunidades en su diseño e implementación.* Retrieved from https://revistas.utp.ac.pa/index.php/clabes/article/view/1545

Pittman, K. (2008). *Comparison of Data Mining Techniques used to Predict Student Retention*.

Puchi, R., Moraga, A., & Villagran, W. (2016). *Modelo De Seguimiento De La Retención Y El Rendimiento De Los Estudiantes De Pregrado De Primer Año En La Universidad De La Frontera*.

Quinto, B. (2020). *Next-Generation Machine Learning with Spark: Covers XGBoost, LightGBM, Spark NLP, Distributed Deep Learning with Keras, and More*. https://doi.org/10.1007/978-1-4842-5669-5

RapidMiner. (2020). RapidMiner Enterprise Pricing. Retrieved from https://rapidminer.com/pricing/

Real Academia Española. (2019). Diccionario de la lengua española. Retrieved from https://dle.rae.es/

Rossmann, J. E., & Kirk, B. A. (1970). Factors related to persistence and withdrawal among university students. *Journal of Counseling Psychology*, *17*(1), 56–62. https://doi.org/10.1037/h0028636

Ruiz Palacios, M. A. (2018). Factores que influyen en la deserción de los alumnos del primer ciclo de educación a distancia en la Escuela de Administración de la Universidad Señor de Sipán. Períodos académicos 2011-1 al 2013-1: lineamientos para disminuir la deserción. *Educación*, *27*(52), 160–173. https://doi.org/10.18800/educacion.201801.009

Sabharwal, N., Barua, S., Anand, N., & Aggarwal, P. (2020). *Developing Cognitive Bots Using the IBM Watson Engine, Practical, Hands-on Guide to Developing Complex Cognitive Bots Using the IBM Watson Platform*. https://doi.org/10.1007/978-1-4842-5555-1

Salvador Blanco, L., & Garcia-Valcarcel Muñoz-Repiso, A. (1989). *El rendimiento académico en la universidad de Cantabria: abandono y retraso en los estudios* (M. de E. y Ciencia, Ed.). Retrieved from https://sede.educacion.gob.es/publiventa/el-rendimiento-academico-en-la-universidad-de-cantabria-abandono-y-retraso-en-los-estudios/investigacion-educativa/1300

SAS. (2020). Visual Data Mining and Machine Learning. Retrieved from https://www.sas.com/es_pe/software/visual-data-mining-machine-learning.html

Scott, G., Shah, M., Grebennikov, L., & Singh, H. (2008). Improving Student Retention A University of Western Sydney Case Study. *Journal of Institutional Research*, *14*(1), 9–23.

Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Computer Science*, *72*, 414–422. https://doi.org/https://doi.org/10.1016/j.procs.2015.12.157

Spady, W. G. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange*, *1*(1), 64–85. https://doi.org/10.1007/BF02214313

SUNEDU I. (2018). *Informe Bienal sobre la realidad universitaria peruana*. Retrieved from https://www.gob.pe/institucion/sunedu/informes-publicaciones/606251-informe-bienal-sobre-la-realidad-universitaria-2018

SUNEDU II. (2020). *II informe bienal sobre la realidad universitaria en el Perú*. Retrieved from https://www.gob.pe/institucion/sunedu/informes-publicaciones/1093280-ii-informe-bienal-sobre-la-realidad-universitaria-en-el-peru

TIBCO. (2020). Scale Data Science across your Enterprise and Act in Real Time. Retrieved from https://www.tibco.com/data-science-and-streaming

Tinto, V. (1975). Drop-Outs From Higher Education: A Theoretical Synthesis of Recent Research. *Review of Educational Research*, *45*, 89–125. https://doi.org/10.2307/1170024

Torres Guevara, L. E. (2012). *Retención estudiantil en la educación superior. Revisión de la literatura y elementos de un modelo para el contexto colombiano* (Editorial Pontificia Universidad Javeriana, Ed.). Bógota.

Universidad Estatal de Sonora. (2019). *Programa de Apoyo y Seguimiento Académico* (p. 21). p. 21. Retrieved from https://www.google.com.pe/search?q=Programa+de+Apoyo+y+Seguimiento+Académico+-+UES+MX&ei=wtB9YM_eEKWEwbkP2cudqAc&oq=Programa+de+Apoyo+y+Seguimiento+Académico+-+UES+MX&gs_lcp=Cgdnd3Mtd2l6EAM6BwgAEEcQsAM6BQghEKABUOniAVjY7gFg3vEBaABwA3gAgAHsAYgBzgSSAQUwLjIuM

Valenzuela, C., & Pérez, S. (2012). Diseño e implementación de Sistema de Seguimiento de Estudiantes y Titulados de la Universidad Diego Portales. *Calidad En La Educación*, *0*(37), 223. https://doi.org/10.31619/caledu.n37.91

Vijayalakshmi, V., & Vengatachalapathy, K. (2019). Deep Neural Network for Multi-Class Prediction of Student Performance in Educational Data.

International Journal of Recent Technology and Engineering, *8*, 5073–5081. https://doi.org/10.35940/ijrte.B2155.078219

Vijayalakshmi, V., & Venkatachalapathy, K. (2019). Comparison of Predicting Student's Performance using Machine Learning Algorithms. *International Journal of Intelligent Systems & Applications*, *11*(12), 34–45. https://doi.org/10.5815/ijisa.2019.12.04

Whitlock, J. L. (2018). *Using Data Science and Predictive Analytics to Understand 4-Year University Student Churn*.

Yalçın, O. G. (2021). *Applied Neural Networks with TensorFlow 2: API Oriented Deep Learning with Python*. https://doi.org/10.1007/978-1-4842-6513-0

Zaccone, G., & Karim, M. R. (2018). *Deep Learning with TensorFlow - Second Edition*.

Mag. Hugo Caselli Gismondi

Doctorando Ingeniería de Sistemas e Informática

ANEXO 1 Código Pandas

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import numpy as np
import matplotlib.pyplot as plt

path = 'C:/Users/HUGO/python/ETL-Doctorado/'
pdf=pd.read_csv(path + '2004-2018 SocioEco Master.csv', encoding = 'utf-8')

pdf["Luz"] = pdf["Servicios"].str.contains("Luz")
pdf["Agua"] = pdf["Servicios"].str.contains("Agua")
pdf["Desague"] = pdf["Servicios"].str.contains("Desague")
pdf["Telefono"] = pdf["Servicios"].str.contains("Telefono")
pdf["Cable"] = pdf["Servicios"].str.contains("Cable")
pdf["Internet"] = pdf["Servicios"].str.contains("Internet")
pdf["Luz"] = pdf["Luz"].replace({True: 1,  False : 0})
pdf["Agua"] = pdf["Agua"].replace({True: 1,  False : 0})
pdf["Desague"] = pdf["Desague"].replace({True: 1,  False : 0})
pdf["Telefono"] = pdf["Telefono"].replace({True: 1,  False : 0})
pdf["Cable"] = pdf["Cable"].replace({True: 1,  False : 0})
pdf["Internet"] = pdf["Internet"].replace({True: 1,  False : 0})

pdf=pdf.drop("Servicios", axis=1)
pdf=pdf.drop("Fecha de Nacimiento", axis = 1)
pdf=pdf.drop("Alumno Trabaja", axis=1)
pdf=pdf.drop("Lugar de Nacimiento", axis = 1)
pdf=pdf.drop("Estado Civil", axis=1)
pdf=pdf.drop("Tipo Vivienda", axis=1)
pdf=pdf.drop("Material Vivienda", axis=1)
pdf=pdf.drop("Tipo de Colegio", axis=1)
pdf=pdf.drop("Veces que Postulo", axis=1)
```

```
pdf["Sexo"] = pdf["Sexo"].replace({"F":0,"M":1})
pdf["Celular"] = pdf["Celular"].replace({"NO":0,"SI":1})

pdf["Dependencia Economica"] = pdf["Dependencia Economica"].replace({"Del Padre":1,"De Ambos Padres":2,"De la Madre":3,"Autosostenimiento":4,"De Familiares/Parientes":5,"De Los Hermanos":6,"Apoderado":7,"Padrastro":8,"Esposo(a)":9})
pdf["Dependencia Economica"]=pdf["Dependencia Economica"].fillna(method='pad')
pdf["Dependencia Economica"]=pdf["Dependencia Economica"].astype(np.int64)

pdf["Condicion Trabajo Responsable Familia"] = pdf["Condicion Trabajo Responsable Familia"].replace({"Trabajo Dependiente":1,"Trabajo Independiente":2,"Trabajo Eventual":3,"Jubilado / Cesante":4,"NINGUNO":5})
pdf["Condicion Trabajo Responsable Familia"]=pdf["Condicion Trabajo Responsable Familia"].fillna(method='pad')
pdf["Condicion Trabajo Responsable Familia"]=pdf["Condicion Trabajo Responsable Familia"].astype(np.int64)

bins=[0.0,450.0,850.0,1500.0,2800.0,6700.0,10000.0]
names = ["1", "2", "3", "4", "5", "6"]
pdf["Ingreso Total Familiar"]=pd.cut(pdf["Ingreso Total Familiar"], bins, labels=names)
pdf["Ingreso Total Familiar"]=pdf["Ingreso Total Familiar"].fillna(method='pad')

pdf["Lugar de Procedencia"] = pdf["Lugar de Procedencia"].replace({"NINGUNO":0,"CHIMBOTE":1,"NUEVO CHIMBOTE":2,"CASMA":3,"SANTA":4,"COISHCO":5})
pdf["Lugar de Procedencia"] = pdf["Lugar de Procedencia"].replace({"SAMANCO":6,"NEPEnA":7,"GUADALUPITO":8,"TRUJILLO":9,"SAN JUAN DE LURIGANCHO":9})
```

```
pdf["Lugar de Procedencia"] = pdf["Lugar de Procedencia"].replace({"MORO":9,"CONCHUCOS":9,"JESUS MARIA":9,"LIMA":9,"SAN JOSE":9,"MOCHUMI":9,"YAUTAN":9})
pdf["Lugar de Procedencia"] = pdf["Lugar de Procedencia"].replace({"SANTA ROSA":9,"HUARMEY":9,"CACERES DEL PERU":9,"CARAZ":9,"VILLA EL SALVADOR":9})

pdf=pd.merge(pdf, gdf, how="left", on=['Codigo'])

pdf=pd.merge(pdf, tdf, how="left", on=['Codigo'])

pdf_train = pdf.iloc[1:395, 0:18]
pdf_test = pdf.iloc[395:, 0:18]

sns.heatmap(pdf_train.isnull(),yticklabels=False,cbar=False,cmap='viridis')
plt.title('Detecta Valores Nulos/Faltantes')
plt.savefig('valores_nulos_01.jpg')

sns.catplot(x="Sexo", kind="count", palette="ch:.25", data=pdf_train)

sns.catplot(x="Grado", kind="count", palette="ch:.25", data=pdf_train)

sns.catplot(x="Titulo", kind="count", palette="ch:.25", data=pdf_train)
```

ANEXO 2 Predicción con data real al azar

Ejemplo Predicción Real

```
prediction_label=["Estudiante", "Egresado", "Bachiller", "Titulado"]
```

```
path = 'C:/Users/HUGO/python/Doctorado/Experimento 04 FI/'
datosr=pd.read_csv(path + 'zz_real_00.csv', encoding = 'utf-8')
```

Dato real: Estudiante

```
xrdf = datosr.iloc[0:, 0:14]
```

```
xrdf
```

	Sexo	Celular	Dependencia Economica	Condicion Trabajo Responsable Familia	Ingreso Total Familiar	Lugar de Procedencia	Luz	Agua	Desague	Telefono	Cable	Internet	Promedios	NumSemestres
0	1	0	0	0.25	0.2	0.111111	1	1	1	0	0	0	0.546848	0.411765

```
xrdf=xrdf.T
```

```
my_prediction = predict(xrdf, parameters)
```

```
PREDICCION=np.squeeze(my_prediction)[()]
PREDICCION
```

```
0
```

```
print("El algoritmo predice: y = " + prediction_label[PREDICCION])
```

```
El algoritmo predice: y = Estudiante
```

```
datosr=pd.read_csv(path + 'zz_real_01.csv', encoding = 'utf-8')
```

Dato real: Egresado

	Sexo	Celular	Dependencia Economica	Condicion Trabajo Responsable Familia	Ingreso Total Familiar	Lugar de Procedencia	Luz	Agua	Desague	Telefono	Cable	Internet	Promedios	NumSemestres
0	1	0	0	0	0.2	0.111111	1	1	1	0	0	0	0.735352	0.352941

```
PREDICCION=np.squeeze(my_prediction)[()]
PREDICCION
```

```
1
```

```
print("El algoritmo predice: y = " + prediction_label[PREDICCION])
```

```
El algoritmo predice: y = Egresado
```

```
datosr=pd.read_csv(path + 'zz_real_02.csv', encoding = 'utf-8')
```

Dato real: Bachiller

	Sexo	Celular	Dependencia Economica	Condicion Trabajo Responsable Familia	Ingreso Total Familiar	Lugar de Procedencia	Luz	Agua	Desague	Telefono	Cable	Internet	Promedios	NumSemestres
0	1	0	0	0	0	0.111111	1	1	1	0	0	0	0.788697	0.294118

```
PREDICCION=np.squeeze(my_prediction)[()]
PREDICCION
```

```
2
```

```
print("El algoritmo predice: y = " + prediction_label[PREDICCION])
```

```
El algoritmo predice: y = Bachiller
```

```
datosr=pd.read_csv(path + 'zz_real_03.csv', encoding = 'utf-8')
```

Dato real: Titulado

	Sexo	Celular	Dependencia Economica	Condicion Trabajo Responsable Familia	Ingreso Total Familiar	Lugar de Procedencia	Luz	Agua	Desague	Telefono	Cable	Internet	Promedios	NumSemestres
0	1	0	0	0.75	0.4	0.111111	1	1	1	1	0	0	0.681662	0.352941

```
PREDICCION=np.squeeze(my_prediction)[()]
PREDICCION
```

```
3
```

```
print("El algoritmo predice: y = " + prediction_label[PREDICCION])
```

```
El algoritmo predice: y = Titulado
```

```
datosr=pd.read_csv(path + 'zz_real_04.csv', encoding = 'utf-8')
```

Dato real: Estudiante

	Sexo	Celular	Dependencia Economica	Condicion Trabajo Responsable Familia	Ingreso Total Familiar	Lugar de Procedencia	Luz	Agua	Desague	Telefono	Cable	Internet	Promedios	NumSemestres
0	1	0	0.125	0.5	0.2	0.222222	1	1	1	0	0	0	0.66922	0.764706

```
PREDICCION=np.squeeze(my_prediction)[()]
PREDICCION
```

```
1
```

```
print("El algoritmo predice: y = " + prediction_label[PREDICCION])
```

```
El algoritmo predice: y = Egresado
```

```
datosr=pd.read_csv(path + 'zz_real_06.csv', encoding = 'utf-8')
```

Dato real: Titulado

	Sexo	Celular	Dependencia Economica	Condicion Trabajo Responsable Familia	Ingreso Total Familiar	Lugar de Procedencia	Luz	Agua	Desague	Telefono	Cable	Internet	Promedios	NumSemestres
0	0	1	0.25	0.5	0.2	0.111111	0	0	0	0	0	0	0.872724	0.264706

```
PREDICCION=np.squeeze(my_prediction)[()]
PREDICCION
```

```
0
```

```
print("El algoritmo predice: y = " + prediction_label[PREDICCION])
```

```
El algoritmo predice: y = Estudiante
```

Informe Tesis Doctoral H. Caselli G.

INFORME DE ORIGINALIDAD

10%	9%	2%	4%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	repositorio.uns.edu.pe Fuente de Internet	2%
2	www.uns.edu.pe Fuente de Internet	<1%
3	revistas.utp.ac.pa Fuente de Internet	<1%
4	bibliotecadigital.univalle.edu.co Fuente de Internet	<1%
5	repositorio.utp.edu.pe Fuente de Internet	<1%
6	Submitted to Universidad Distrital FJDC Trabajo del estudiante	<1%
7	Submitted to Institución Tecnológica Metropolitana de Medellín Trabajo del estudiante	<1%
8	Submitted to Universidad Nacional del Santa Trabajo del estudiante	<1%
9	1library.co Fuente de Internet	

		<1 %
10	core.ac.uk Fuente de Internet	<1 %
11	repositorio.cuc.edu.co Fuente de Internet	<1 %
12	Submitted to Universidad Nacional del Centro del Peru Trabajo del estudiante	<1 %
13	scoif.com Fuente de Internet	<1 %
14	contextoseducativosinteractivos.files.wordpress.com Fuente de Internet	<1 %
15	ridda2.utp.ac.pa Fuente de Internet	<1 %
16	www.scielo.cl Fuente de Internet	<1 %
17	I. Papastefanou, D. Wright, K. H. Nicolaides. "Competing - risks model for prediction of small - for - gestational - age neonate from maternal characteristics and medical history", Ultrasound in Obstetrics & Gynecology, 2020 Publicación	<1 %
18	clabes-alfaguia.org Fuente de Internet	<1 %

19	Submitted to Universidad Alas Peruanas Trabajo del estudiante	<1 %
20	Submitted to Instituto Europeo de Posgrado Trabajo del estudiante	<1 %
21	Submitted to Universidad Santo Tomas Trabajo del estudiante	<1 %
22	ridum.umanizales.edu.co:8080 Fuente de Internet	<1 %
23	tesis.ucsm.edu.pe Fuente de Internet	<1 %
24	repository.usta.edu.co Fuente de Internet	<1 %
25	repositorio.unsm.edu.pe Fuente de Internet	<1 %
26	www.mecs-press.org Fuente de Internet	<1 %
27	digibug.ugr.es Fuente de Internet	<1 %
28	repositorio.usanpedro.edu.pe Fuente de Internet	<1 %
29	www.scielo.org.mx Fuente de Internet	<1 %
30	www.alfaguia.org Fuente de Internet	<1 %

31	Submitted to UNILIBRE Trabajo del estudiante	<1 %
32	Submitted to Universidad Internacional de la Rioja Trabajo del estudiante	<1 %
33	Submitted to Universitat Politècnica de València Trabajo del estudiante	<1 %
34	hdl.handle.net Fuente de Internet	<1 %
35	biblioteca.uniatlantico.edu.co Fuente de Internet	<1 %
36	www.go-mono.com Fuente de Internet	<1 %
37	www.graficacorrientes.com.ar Fuente de Internet	<1 %
38	W. E. Luera Peña, L. A. Minim. "APLICACIÓN DE REDES NEURONALES ARTIFICIALES EN LA MODELIZACIÓN DEL TRATAMIENTO TÉRMICO DE ALIMENTOS APPLICATION OF NEURAL NETWORKS IN THE MODELLING OF THE THERMAL TREATMENT OF FOOD APLICACIÓN DE REDES NEURONAIS ARTIFICIAIS NA MODELIZACIÓN DO TRATAMENTO TÉRMICO DOS ALIMENTOS", Ciencia y Tecnologia Alimentaria, 2001 Publicación	<1 %

39	www.scribd.com Fuente de Internet	<1 %
40	Submitted to Politécnico Colombiano Jaime Isaza Cadavid Trabajo del estudiante	<1 %
41	cybertesis.unmsm.edu.pe Fuente de Internet	<1 %
42	cybertesis.urp.edu.pe Fuente de Internet	<1 %
43	docplayer.es Fuente de Internet	<1 %
44	doctorpenguin.com Fuente de Internet	<1 %
45	heroica.upt.edu.pe Fuente de Internet	<1 %
46	pt.scribd.com Fuente de Internet	<1 %
47	repositorio.ual.es:8080 Fuente de Internet	<1 %
48	repositorio.uniandes.edu.co Fuente de Internet	<1 %
49	repositorio.uwiener.edu.pe Fuente de Internet	<1 %
	www.cned.cl	

50	Fuente de Internet	<1 %
51	www.dgae.unam.mx Fuente de Internet	<1 %
52	www.slideshare.net Fuente de Internet	<1 %
53	Submitted to Universidad Cesar Vallejo Trabajo del estudiante	<1 %
54	bonga.unisimon.edu.co Fuente de Internet	<1 %
55	fam.eluniversal.com Fuente de Internet	<1 %
56	link.springer.com Fuente de Internet	<1 %
57	rcientificas.uninorte.edu.co Fuente de Internet	<1 %
58	repositorio.uchile.cl Fuente de Internet	<1 %
59	repositorio.unab.cl Fuente de Internet	<1 %
60	repositorio.unfv.edu.pe Fuente de Internet	<1 %
61	reunir.unir.net Fuente de Internet	<1 %

62	revistas.libertadores.edu.co Fuente de Internet	<1 %
63	sedici.unlp.edu.ar Fuente de Internet	<1 %
64	tecnoysoc.com Fuente de Internet	<1 %
65	wikizero.com Fuente de Internet	<1 %
66	www.ivis.org Fuente de Internet	<1 %
67	www.usmp.edu.pe Fuente de Internet	<1 %
68	ansotophd.blogspot.com Fuente de Internet	<1 %
69	ccc.inaoep.mx Fuente de Internet	<1 %
70	es.news.yahoo.com Fuente de Internet	<1 %
71	future.inese.es Fuente de Internet	<1 %
72	issuu.com Fuente de Internet	<1 %
73	oa.upm.es Fuente de Internet	<1 %

74	posgrado.lapaz.tecnm.mx Fuente de Internet	<1 %
75	repositorio.unicartagena.edu.co Fuente de Internet	<1 %
76	revistas.uladech.edu.pe Fuente de Internet	<1 %
77	ridum.umanizales.edu.co Fuente de Internet	<1 %
78	ww2.ufps.edu.co Fuente de Internet	<1 %
79	www.dednet.org Fuente de Internet	<1 %
80	www.dirinfo.unsl.edu.ar Fuente de Internet	<1 %
81	www.medicina.usmp.edu.pe Fuente de Internet	<1 %
82	www.researchgate.net Fuente de Internet	<1 %
83	repository.unad.edu.co Fuente de Internet	<1 %
84	"Intelligent Human Systems Integration 2021", Springer Science and Business Media LLC, 2021 Publicación	<1 %

85 Orhan Gazi Yalçın. "Applied Neural Networks with TensorFlow 2", Springer Science and Business Media LLC, 2021 <1%

Publicación

Excluir citas	Activo	Excluir coincidencias	Apagado
Excluir bibliografía	Activo		