UNIVERSIDAD NACIONAL DE SAN CRISTÓBAL DE HUAMANGA

FACULTAD DE INGENIERÍA DE MINAS, GEOLOGÍA Y CIVIL

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

PREDICCIÓN DEL RENDIMIENTO ACADÉMICO BASADO EN MACHINE LEARNING, ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS, AYACUCHO 2021.

TESIS PARA OPTAR EL TÍTULO PROFESIONAL DE:

INGENIERO DE SISTEMAS

PRESENTADO POR:

Bach. Ever Aronés Ayala

ASESOR:

Dr. Ing. Manuel Avelino Lagos Barzola

Ayacucho - Perú

2021

ACTA DE SUSTENTACIÓN DE TESIS N° 075-2021-FIMGC

En la ciudad de Ayacucho, en cumplimiento a la **Resolución Decanal N° 752-2021-FIMGC-D**, siendo los veintinueve días del mes de diciembre del 2021, a horas 2:00 p.m.; se reunieron los jurados del acto de sustentación, en el Auditórium virtual google meet del Campus Universitario de la Universidad Nacional de San Cristóbal de Huamanga.

Siendo el Jurado de la sustentación de tesis compuesto por el Presidente el **Dr. Ing. Efraín Elías PORRAS FLORES,** Jurado el **Mg. Ing. Eloy VILA HUAMAN,** Jurado el **Mg. Ing. Christian LEZAMA CUELLAR,** Jurado – Asesor el **Dr. Ing. Manuel Avelino LAGOS BARZOLA** y Secretario del proceso el **Ing. José Antonio GUERRERO HINOSTROZA,** con el objetivo de recepcionar la sustentación de la tesis denominada **"PREDICCIÓN DEL RENDIMIENTO ACADÉMICO BASADO EN MACHINE LEARNING, ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS, AYACUCHO 2021"**, del Bachiller **Ever ARONÈS AYALA,** sustentado por el Señor **Ever ARONÈS AYALA,** Bachiller en **Ingeniería de Sistemas.**

El Jurado luego de haber recepcionado la sustentación de la tesis y realizado las preguntas, el sustentante al haber dado respuesta a las preguntas, y el Jurado haber deliberado; califica con la nota aprobatoria de **16 (dieciséis).**

En fe de lo cual, se firma la presente acta, por los miembros integrantes del proceso de sustentación.

UNIVERSIDAD NACIONAL DE SAN CRISTÓBAL DE HUAMANGA
FACULTAD DE INGENIERÍA DE MINAS GEOLOGÍA Y CIVIL
Dr. Efraín Elías Porras Flores
DECANO

Firmado digitalmente por Dr. Ing. Efrain Elias Porras Flores
Fecha: 2022.01.10 18:18:20 -05'00'

Dr. Ing. Efraín Elías PORRAS FLORES
Presidente

Mg. Ing. Eloy VILA HUAMÁN
Jurado

Firmado digitalmente por LEZAMA CUELLAR CHRISTIAN

Mg. Ing. Christian LEZAMA CUELLAR
Jurado

Dr. Ing. Manuel Avelino LAGOS BARZOLA
Jurado - Asesor

Ing. José Antonio GUERRERO HINOSTROZA
Secretario del Proceso

c.c.:
Bach. Ever ARONÉS AYALA
Jurados (4)
Archivo

CONSTANCIA DE ORIGINALIDAD DE TRABAJO DE INVESTIGACIÓN

El que suscribe; responsable verificador de originalidad de trabajos de tesis de pregrado en segunda instancia para las **Escuelas Profesionales** de la **Facultad de Ingeniería de Minas, Geología y Civil**; en cumplimiento a la Resolución de Consejo Universitario N° 039-2021-UNSCH-CU, Reglamento de Originalidad de Trabajos de Investigación de la UNSCH y Resolución Decanal N° 158-2021-FIMGC-UNSCH-D, deja constancia que Sr./Srta.

Apellidos y Nombres	:	ARONÉS AYALA, Ever
Escuela Profesional	:	INGENIERÍA DE SISTEMAS
Título de la Tesis	:	"PREDICCIÓN DEL RENDIMIENTO ACADÉMICO BASADO EN MACHINE LEARNING, ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS, AYACUCHO 2021"
Evaluación de la Originalidad	:	**4 %** Índice de Similitud
Identificador de la entrega	:	1734008433

Por tanto, según los Artículos 12, 13 y 17 del Reglamento de Originalidad de Trabajos de Investigación, es **PROCEDENTE** otorgar la **Constancia de Originalidad** para los fines que crea conveniente.

Ayacucho, 20 de diciembre del 2021

Firmado digitalmente por LEZAMA CUELLAR CHRISTIAN

Mg. Ing. Christian LEZAMA CUELLAR
Verificador de Originalidad de Trabajos de Tesis de Pregrado de la FIMGC

Numero de constancia: **172**-2021-FIMGC.

(X) Con depósito para Sustentación y Tramite de Titulo

FACULTAD DE INGENIERÍA
DE MINAS, GEOLOGÍA Y CIVIL
Av. Independencia S/N
Ciudad Universitaria
Central Tel 066 312510
Anexo 151

“PREDICCIÓN DEL RENDIMIENTO ACADÉMICO BASADO EN MACHINE LEARNING, ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS, AYACUCHO 2021”

por Ever Aronés Ayala

Fecha de entrega: 19-dic-2021 11:47p.m. (UTC-0500)
Identificador de la entrega: 1734008433
Nombre del archivo: Tesis_Ever_Aron_s_Ayala_EPIS.pdf (1.93M)
Total de palabras: 22426
Total de caracteres: 127348

“PREDICCIÓN DEL RENDIMIENTO ACADÉMICO BASADO EN MACHINE LEARNING, ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS, AYACUCHO 2021”

INFORME DE ORIGINALIDAD

4%	5%	2%	3%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	repositorio.uncp.edu.pe Fuente de Internet	1%
2	repositorio.unsch.edu.pe Fuente de Internet	1%
3	cybertesis.unmsm.edu.pe Fuente de Internet	1%
4	Submitted to Universidad Nacional de San Cristóbal de Huamanga Trabajo del estudiante	<1%
5	repositorioacademico.upc.edu.pe Fuente de Internet	<1%
6	repositorio.unab.cl Fuente de Internet	<1%
7	hdl.handle.net Fuente de Internet	<1%
8	repositorio.unac.edu.pe Fuente de Internet	<1%

9	docslide.us Fuente de Internet	<1 %
10	upcommons.upc.edu Fuente de Internet	<1 %
11	www.clubensayos.com Fuente de Internet	<1 %

Excluir citas	Activo	Excluir coincidencias	< 30 words
Excluir bibliografía	Activo		

DEDICATORIA

A mis estimados padres, pues gracias a su incondicional apoyo pude avanzar en las diferentes etapas de mi vida, siempre ayudando y aconsejándome con la mejor de las intenciones para hacer de mí una persona de buenos valores.

A mis queridos hermanos, por ser un ejemplo de ética profesional, por lo que representan para mí y por ser parte importante de una hermosa familia unida.

AGRADECIMIENTO

A Dios, porque gracias a la fe pude encontrar calma, tranquilidad y un respaldo fuerte para afrontar los diferentes dilemas que en la vida existen.

A mi alma mater, la Universidad Nacional San Cristóbal de Huamanga por ser la institución formadora de excelentes profesionales, y donde pude concretar mi meta de superación académica.

A todos los docentes que en algún momento de mi vida universitaria estuvieron ahí brindándome sus conocimientos y apoyo para contribuir en mi formación profesional.

Al Dr. Ing. Manuel A. Lagos Barzola, mi asesor de tesis, por orientarme en el proceso de mi investigación, guiándome y brindándome sus conocimientos y experiencias.

RESUMEN

Las técnicas del aprendizaje automático (Machine Learning) son usadas ampliamente en los altos niveles de una organización ganando cada vez más trascendencia, pues, dependiendo de su correcto uso, se obtienen datos, patrones, tendencias y/o probabilidades que son de gran importancia y relevancia cuando se toman decisiones; dado que la fuente con la que trabaja el aprendizaje automático son los datos históricos, los cuales toda empresa la tiene y son reunidos hasta inconscientemente, es sólo cuestión de tiempo para que el uso de las técnicas del machine learning obtengan prioridad como herramienta eficaz en el tratamiento de la información.

En este trabajo, el objetivo es diseñar un modelo de machine learning que tiene como finalidad la predicción del rendimiento académico de los alumnos de la Escuela Profesional de Ingeniería de Sistemas, el nivel de investigación es descriptivo y como tipo de investigación: transversal, observacional y retrospectivo.

Se diseñó un modelo de machine learning para predecir el rendimiento académico de los alumnos de la Escuela Profesional de Ingeniería de Sistemas, aplicando la metodología fundamental para la ciencia de datos, pues es una metodología muy fácil de entender ya que dentro de sus etapas se describe todo lo que se necesita saber, desde conocer el negocio, pasando por la recolección de datos, su tratamiento, análisis y limpieza, continuando con el modelado y evaluación de los algoritmos, para finalmente terminar con la implementación y retroalimentación del modelo creado; se utilizó como lenguaje de programación a Python, como librerías de tratamiento de datos a pandas, numpy y matplotlib, como interfaz de desarrollo a Jupyter Notebook perteneciente a la suite Anaconda y como algoritmos predictivos a la regresión logística y random forest.

PALABRAS CLAVE
Data Science, Machine Learning, rendimiento académico, Python, metodología fundamental para la ciencia de datos, modelo predictivo, aprendizaje supervisado.

ABSTRACT

Machine learning techniques are widely used at the high levels of an organization gaining more and more importance, because, depending on their correct use, data, patterns, trends and / or probabilities are obtained that are of great importance and relevance when making decisions; Given that the source with which machine learning works is historical data, which every company has and is even unconsciously collected, it is only a matter of time before the use of machine learning techniques obtain priority as an effective tool in the information processing.

In this work, the objective is to design a machine learning model that aims to predict the academic performance of the students of the Professional School of Systems Engineering, the research level is descriptive and as a type of research: cross-sectional, observational and retrospective.

A machine learning model was designed to predict the academic performance of the students of the Professional School of Systems Engineering, applying the fundamental methodology for data science, since it is a very easy to understand methodology since within its stages it is described everything you need to know, from knowing the business, going through data collection, treatment, analysis and cleaning, continuing with the modeling and evaluation of the algorithms, to finally finish with the implementation and feedback of the created model; Python was used as a programming language, as pandas, numpy and matplotlib data processing libraries, as a development interface to Jupyter Notebook belonging to the Anaconda suite, and as predictive algorithms for logistic regression and random forest.

KEYWORDS:
Data Science, Machine Learning, academic performance, Python, fundamental methodology for data science, predictive model, supervised learning.

ÍNDICE

Pág.

Dedicatoria ii
Agradecimiento iii
Resumen iv
Abstract v
Índice vi
Lista de tablas ix
Lista de figuras x
Introducción 1

CAPÍTULO I
PLANTEAMIENTO DEL PROBLEMA

1.1. Diagnóstico y enunciado del problema 3
1.2. Formulación del problema 5
1.3. Objetivos de la investigación 5
1.4. Hipótesis de la investigación 6
1.5. Justificación y delimitación de la investigación 6
1.5.1. Importancia del tema 6
1.5.2. Delimitación 7
1.5.3. Justificación 7

CAPÍTULO II
MARCO TEÓRICO

2.1. Antecedentes de la investigación 9
2.2. Marco teórico 10
2.2.1. Ciencia de datos 10
2.2.2. Machine learning 12
2.2.3. Metodología fundamental para la ciencia de datos 24
2.2.4. Rendimiento académico 33
2.2.5. Python 34

2.2.6. Lenguaje de programación orientada a objetos 35
2.2.7. Población.....36
2.2.8. Muestra 36
2.2.9. Muestreo por conveniencia 36

CAPÍTULO III
MATERIALES Y MÉTODOS

3.1. Tipo y nivel de la investigación.....37
3.1.1. Tipo de investigación.....37
3.1.2. Nivel de investigación.....38
3.2. Diseño de la investigación.....39
3.3. Población y muestra.....39
3.4. Variables e indicadores 39
3.4.1. Definición conceptual de las variables.....39
3.4.2. Definición operacional de las variables 40
3.5. Técnicas e instrumentos para recolectar información 40
3.5.1. Técnicas para recolectar información.....40
3.5.2. Instrumentos para recolectar información 40
3.5.3. Herramientas para el tratamiento de datos e información.....41
3.5.4. Técnica para aplicar la metodología fundamental para la ciencia de datos.....43

CAPÍTULO IV
RESULTADOS Y DISCUSIÓN

4.1. Resultados de la metodología fundamental para la ciencia de datos...46
4.1.1. Comprensión del negocio.....46
4.1.2. Objetivo del negocio.....48
4.1.3. Enfoque analítico.....48
4.1.4. Requisito de datos.....48
4.1.5. Recopilación de datos 49
4.1.6. Comprensión de datos 51
4.1.7. Preparación de datos 61

4.1.8. Modelado.....65
4.1.9. Evaluación.....80
4.1.10. Implementación.....81
4.1.11. Retroalimentación.....82
4.2. Discusión.....82

CAPÍTULO V

CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones.....86
5.2. Recomendaciones.....87

REFERENCIAS BIBLIOGRÁFICAS.....88

LISTA DE TABLAS

Nº. TÍTULO DE LA TABLA PAG.

Tabla 1: Resumen académico 2019..... 4
Tabla 2: Herramientas tecnológicas.....41
Tabla 3: Etapas de la metodología fundamental para ciencia de datos.....44
Tabla 4: Datos obtenidos de la Oficina General de Informática y Sistemas. 49
Tabla 5: Resultados del modelo de regresión logística. 80
Tabla 6: Resultados del modelo de k vecinos cercanos.....80
Tabla 7: Resultados del modelo de máquina de vector de soporte. 80
Tabla 8: Resultados del modelo de árbol de decisión. 80
Tabla 9: Resultados del modelo de random forest. 81

LISTA DE FIGURAS

Nº. TÍTULO DE LA FIGURA PAG.

Figura 1: Ciclo de vida del dato 11
Figura 2: Flujo de datos en el modelo POSMAD 12
Figura 3: Componentes del Machine Learning 13
Figura 4: Ubicación jerárquica del Machine Learning 13
Figura 5: Forma de trabajo del aprendizaje supervisado 14
Figura 6: Algoritmo de clasificación 15
Figura 7: Gráfica de una función logística 17
Figura 8: Gráfica k-vecinos cercanos 18
Figura 9: Figura máquina de soporte vectorial.....18
Figura 10: Figura Árbol de decisión.....19
Figura 11: Procesos del aprendizaje no supervisado 21
Figura 12: Metodología fundamental para la ciencia de datos.....25
Figura 13: Matriz de confusión 31
Figura 14: Curva ROC.....32
Figura 15: Esquema k-fold validación cruzada 33
Figura 16: Registro inicial de alumnos. 50
Figura 17: Registro filtrado con alumnos de la EPIS.....50
Figura 18: Registro final de alumnos de la EPIS. 51
Figura 19: Visualización de los 5 primeros registros del Dataframe.....52
Figura 20: Distribución general del rendimiento académico.....53
Figura 21: Distribución del rendimiento académico según el género. 54
Figura 22: Distribución del rendimiento académico según el semestre académico.....55
Figura 23: Distribución del rendimiento académico según la modalidad del curso.....56
Figura 24: Distribución del rendimiento académico según los créditos. 57
Figura 25: Distribución del rendimiento académico según los cursos. 58
Figura 26: Distribución del rendimiento académico según docentes. 59
Figura 27: Distribución del rendimiento académico según el colegio de egreso. 61
Figura 28: Columnas a usar para el modelo. 62
Figura 29: Tipo inicial de datos que tiene el dataframe.....63
Figura 30: Tipo final de datos que tiene el dataframe. 64
Figura 31: Dataframe final. 64
Figura 32: Conjunto X: variables independientes. 65
Figura 33: Conjunto Y: variable dependiente.....65

Figura 34: Matriz de confusión del algoritmo de regresión logística. 68
Figura 35: Curva ROC del modelo de regresión logística. 69
Figura 36: Matriz de confusión del algoritmo de k vecinos cercanos. 71
Figura 37: Curva ROC del modelo de k vecinos cercanos 72
Figura 38: Matriz de confusión del algoritmo máquina de vector de soporte 73
Figura 39: Curva ROC del modelo de máquina de vector de soporte. 74
Figura 40: Matriz de confusión del algoritmo árbol de decisión 76
Figura 41: Curva ROC del modelo de árbol de decisión. 77
Figura 42: Matriz de confusión del algoritmo random forest. 78
Figura 43: Curva ROC del modelo de random forest. 79

INTRODUCCIÓN

Todas las organizaciones generan una gran cantidad de datos, tal es el caso de la Escuela de Formación Profesional de Ingeniería de Sistemas; por tanto, es prudente usarlos adecuadamente empleando las mejores herramientas con las que se cuenta en la actualidad. El Machine Learning es una disciplina científica que en la actualidad brinda muchas ventajas y facilidades en cuanto se trata del análisis de datos, dando como resultado muchos modelos que brindan información valiosa para cualquier organización que la implemente.

Mi motivación para desarrollar un modelo predictivo es para poder contar con una herramienta que ayude a los directivos de la Escuela Profesional de Ingeniería de Sistemas para que tomen medidas preventivas en cuanto al control de la población estudiantil y sobre todo ayudar al mismo estudiante en identificar sus probabilidades de mejora en el rendimiento académico.

Actualmente existen muchos modelos basados en Machine Learning que producen y extraen información valiosa, pero por diversos motivos la implementación de dichos modelos no sucede, lo que causa que la información generada por las organizaciones simplemente se guarde, desaprovechando una fuente muy valiosa y que definitivamente puede generar un valor adicional.

Como objetivos específicos están: a) Analizar y determinar la técnica predictiva más adecuada, con la finalidad de predecir el rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas, Ayacucho 2021. b) Evaluar y determinar el algoritmo predictivo más preciso, con la finalidad de predecir el rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas, Ayacucho 2021. c) Evaluar y determinar la métrica de validez más adecuada para el modelo de Machine Learning, con la finalidad de predecir el

rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas, Ayacucho 2021.

CAPÍTULO I
PLANTEAMIENTO DEL PROBLEMA

1.1 DIAGNÓSTICO Y ENUNCIADO DEL PROBLEMA

Uno de los ejes fundamentales para el progreso de un país es la educación, pues es la herramienta con la que cuentan las personas para superarse en el ámbito profesional, por tanto, cualquier tema que ayude en la mejora educativa del estudiante siempre es bienvenida.

La palabra educación es definida en el Diccionario de la Lengua Española como: “Crianza, enseñanza y doctrina que se da a los niños y a los jóvenes” es decir que está presente en buena parte de la vida de una persona. Si nos centramos en la formación de profesionales, actualmente uno de los problemas de la sociedad son los resultados de la educación universitaria, ya que son los egresados quienes se enfrentarán a los problemas de la realidad y es lógico la preocupación por la calidad y demás detalles que se generan a lo largo de su vida universitaria; siendo uno de esos detalles los resultados académicos de los estudiantes, más precisamente su rendimiento académico.

El bajo rendimiento académico de los estudiantes universitarios es un problema constante, pero a diferencia de antes, ya no puede ser ignorado, pues el crecimiento de las ciudades y su población hace que dicho problema genere otros como es la sobrepoblación de estudiantes, pocas aulas, creación de varios grupos en los cursos, y que la admisión de estudiantes sea más rigurosa pues los egresados muchas veces demoran más de 5 años en acabar su carrera.

El rendimiento académico debe ser analizado de manera cuantitativa, pues se cuenta con datos históricos registrados y es su estudio y análisis los que nos brindarán información sólida que ayudarán en la toma de decisiones en cualquier nivel;

actualmente la Escuela Profesional de Ingeniería de Sistemas cuenta con dichos datos, pero simplemente se encuentran guardados, esto hace que la implementación de un método de análisis de datos sea necesario para aprovechar esa vital información con que se cuenta.

Gracias al proceso de globalización, la cantidad de datos que se genera es descomunal, por tanto, se deben emplear diversos métodos de análisis de datos, de los cuales la Data Science y Big Data son la nueva realidad, y de estos es el concepto de Machine Learning el más interesante; pues, es un método de análisis de datos que automatiza la construcción de modelos analíticos empleando algoritmos que aprenden de los datos y encuentra patrones o datos interesantes, muchas veces sin estar programado para dicha búsqueda.

Tabla 1
Resumen académico 2019.

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS				
N.º		**SEMESTRE**		**TOTAL**
		2019-I	**2019-II**	
1	Ingresantes	42	53	**95**
2	Matriculados	560	511	**1071**
3	Egresados	10	34	**44**

Nota. Fuente: Boletín estadístico UNSCH - 2019

Como se puede observar en la tabla N.º 1, en el año 2019 el total de matriculados fue de 1071, una cantidad alta la cual debe ser discutida y analizada del por qué hay muchos alumnos, de la misma tabla se puede deducir una de las razones, pues la cantidad de alumnos que egresaron (44) es menos de la mitad que la cantidad de ingresantes (95), indicándonos claramente que el flujo de salida es menor al de entrada, lo que incrementa la población estudiantil dentro de la escuela, que por ahora está por encima de los 500; esto trae consigo sus propios problemas como son la creación de nuevos grupos, retraso en iniciar las clases para dichos grupos, insuficiencia de salones, docentes y laboratorios; pues el atender a tal cantidad de matriculados requiere de estrategias para cumplir con el derecho de estudiar que tiene el alumno universitario.

A lo anteriormente descrito se debe añadir que, como una razón del crecimiento de la población estudiantil es el bajo rendimiento académico que tienen los alumnos, por

tanto, el usar las herramientas tecnológicas de análisis de datos como los modelos de machine learning es una propuesta viable y que tiene una buena fuente de alimentación como son los registros que año tras año la Escuela Profesional de Ingeniería de Sistemas guarda de los estudiantes que se matriculan.

1.2 FORMULACIÓN DEL PROBLEMA

PROBLEMA PRINCIPAL

¿De qué manera diseñar un modelo de Machine Learning para predecir el rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas, Ayacucho 2021?

PROBLEMAS SECUNDARIOS

a. ¿Cómo determinar la técnica predictiva más adecuada para la predicción del rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas, Ayacucho 2021?
b. ¿Cuál es el algoritmo predictivo más preciso en la predicción del rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas, Ayacucho 2021?
c. ¿Cómo determinar la métrica de validez para el modelo de Machine Learning que permite la predicción del rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas, Ayacucho 2021?

1.3 OBJETIVOS DE LA INVESTIGACIÓN

OBJETIVO GENERAL

Diseñar un modelo de Machine Learning mediante técnicas e instrumentos, metodología fundamental para la ciencia de datos, un lenguaje de programación orientada a objetos y tecnologías de análisis de datos, con la finalidad de predecir el rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas, Ayacucho 2021.

OBJETIVOS ESPECÍFICOS

a. Analizar y determinar la técnica predictiva más adecuada, con la finalidad de predecir el rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas, Ayacucho 2021.
b. Evaluar y determinar el algoritmo predictivo más preciso, con la finalidad de predecir el rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas, Ayacucho 2021.

c. Evaluar y determinar la métrica de validez más adecuada para el modelo de Machine Learning, con la finalidad de predecir el rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas, Ayacucho 2021.

1.4 HIPÓTESIS DE LA INVESTIGACIÓN

Según (Supo, 2014) “las investigaciones de nivel predictivo desde el punto de vista estadístico, no se trata de poner a prueba hipótesis, sino de construir modelos predictivos; para ello se aplican técnicas específicas, como las ecuaciones estructurales, las series de tiempo y el análisis de supervivencia, así como la minería de datos”.

De la misma manera (Hernández, Fernández y Baptista,2014) indican:

> Que no en todas las investigaciones cuantitativas se plantean hipótesis. El hecho de que formulemos o no hipótesis depende de un factor esencial: el alcance inicial del estudio. Las investigaciones cuantitativas que formulan hipótesis son aquellas cuyo planteamiento define que su alcance será correlacional o explicativo. (p.104)

(Supo, 2014) afirma que:

> Los estudios no llevan hipótesis porque no es la intencionalidad del investigador, el propósito del estudio no expresa la necesidad de afirmar o negar. En la investigación cuantitativa, los estudios que no poseen hipótesis por lo general buscan la cuantificación de la relación entre las variables. (p.141)

Por tanto, al tratarse de una investigación de nivel predictivo y dado que el objetivo es el diseño de un modelo predictivo empleando las técnicas del machine learning sin buscar la afirmación o negación de un hecho futuro, se determinó no plantear alguna hipótesis.

1.5 JUSTIFICACIÓN Y DELIMITACIÓN DE LA INVESTIGACIÓN

1.5.1 IMPORTANCIA DEL TEMA

IMPORTANCIA TÉCNICA

Al contar con un modelo predictivo de Machine Learning se podrá tener una herramienta para ser usada en la predicción del rendimiento académico, ayudando al personal profesional de la Escuela Profesional de Ingeniería de Sistemas a identificar oportunamente las falencias en los alumnos y así tomar las medidas adecuadas.

Asimismo, los estudiantes tendrán a su disponibilidad un método para analizar datos, el cual los incentivará, primero en su habilidad de recopilar, procesar y extraer el valor de un conjunto de datos y segundo en la creación de nuevos datos, tendencias y probabilidades que se generan de dichos datos.

IMPORTANCIA ECONÓMICA

El diseñar este modelo predictivo basado en Machine Learning logrará predecir qué alumnos contarán con bajo rendimiento académico y en qué cursos ayudándoles a tomar las precauciones necesarias para no incurrir en un gasto innecesario si es que se desaprueba el semestre; además la Escuela Profesional de Ingeniería de Sistemas también reduciría sus costos pues no se contratarían a más docentes necesarios cuando la cantidad de alumnos en los cursos obliga a la división y formación de grupos; se tendría un régimen más tranquilo y fluido con los horarios, grupos y laboratorios pues los alumnos no se estancarían y no se aglomerarían en unos pocos cursos.

1.5.2 JUSTIFICACIÓN

La gran variedad existente de métodos y técnicas que permiten analizar datos, pertenecen generalmente al Big Data y Machine Learning; por tanto, es importante su implementación, pues dada la cantidad de datos que se posee, es vital procesarlos para extraer información valiosa lo que supondría una ventaja competitiva y, si son tratados de manera eficiente, se podrá lograr las predicciones de alto valor, los que permitirán tomar mejores decisiones y desarrollar mejores estrategias.

Por lo tanto, es necesario contar con un modelo predictivo de Machine Learning basado en algoritmos que aprendan de manera automática y además que puedan detectar patrones de comportamiento en la base de datos histórica de los alumnos de la Escuela Profesional de Ingeniería de Sistemas, esto para predecir el rendimiento académico de los mismos.

1.5.3 DELIMITACIÓN

El presente trabajo limitará las dimensiones que engloba el rendimiento académico y sólo se centrará en el promedio final del estudiante reduciendo los valores a aprobado y desaprobado, para ello tomará como base de estudio a los datos históricos de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas, el modelo de Machine Learning analizará cinco algoritmos de aprendizaje para determinar cuál es el más adecuado; el sistema operativo que se usará será Windows 10, la suite de Data

Science será Anaconda, el entorno de desarrollo será Jupyter Notebook, el lenguaje de programación usado será Python, los paquetes empleados en el análisis de datos serán Numpy, Matplotlib y Pandas, los datos se obtuvieron en el año 2021.

CAPÍTULO II
MARCO TEÓRICO

2.1 ANTECEDENTES DE LA INVESTIGACIÓN

(Pojon, 2017), en su investigación titulada “Using machine learning to predict student performance”, indica que los modelos usados, en dos grupos de datos diferentes, fueron los árboles de decisión, la regresión lineal y Naive Bayes, concluyendo que las características de las variables pueden mejorar la predicción, obteniendo un 98% de exactitud en la predicción del primer grupo con el algoritmo de Naive Bayes y un 78% de exactitud en la predicción del segundo grupo con el algoritmo del árbol de decisión.

(Lizares, 2017), en su investigación titulada “Comparación de modelos de clasificación: regresión logística y árboles de clasificación para evaluar el rendimiento académico”, indica que los árboles de decisión tienen una mayor preponderancia al momento de clasificar y predecir; además concluyó que: primero, lo que más influyó fueron el tipo de colegio, las horas de estudio empleadas y el interés que tienen hacia la asignatura; segundo, lo que menos tuvo influencia fueron el sexo y la formación profesional de los padres; finalmente, menciona que obtuvo un 92.34% de probabilidad de reprobar una asignatura en aquellos alumnos que, provienen de instituciones educativas nacionales, que no le dedican ni tres horas de estudio y que no les interesa el curso.

(Zevallos, 2017), en su investigación titulada “Predicción del rendimiento académico mediante redes neuronales”, concluye que de toda la red neuronal el 84% válido y que los factores que determinan el rendimiento académico del estudiante de nivel primario dependen de su grado de conocimiento.

(Menacho, 2017), en su apartado científico denominado “Predicción del rendimiento académico aplicando técnicas de minería de datos”, indica que para su modelo probabilístico usó las siguientes técnicas: las redes bayesianas, los árboles de

decisión, las redes neuronales y la regresión logística sobre los registros de alumnos del curso de Estadística General; concluyendo que el modelo de Naive Bayes es el que tiene una mayor precisión, con 71% de similitud con los resultados reales.

(Camborda, 2014), en su investigación titulada “Aplicación de árboles de decisión para la predicción del rendimiento académico de los estudiantes de los primeros ciclos de la carrera de Ingeniería Civil de la Universidad Continental”, indica que se centró en las variables de los factores que influyen en mayor grado al estudiante, concluyendo que las variables académicas son las que tienen más relevancia al momento de predecir.

2.2 MARCO TEÓRICO

2.2.1 CIENCIA DE DATOS

Según (Ozdemir, 2016), la ciencia de datos o Data Science es la habilidad para obtener conocimiento empleando y/o manipulando la información almacenada (datos); dependiendo en gran medida en cómo manejan dicha información, es decir, que los datos resultantes de dicho aprendizaje dependen directamente en cómo utilizan los datos que poseen.

El conocimiento acerca de la Data Science es aún nuevo, los límites todavía no están definidos por tanto puede crecer o acortarse, pero las ciencias informáticas tienen un largo periodo estudiándola. Abarca muchos elementos con los que trabaja, como son: el aprendizaje automático o machine learning, distintos modelos estadísticos, la matemática del álgebra lineal, emplea el lenguaje computacional para sus visualizaciones, por tanto, se pueden realizar el análisis de diferentes gráficos que pueden ayudar a la inteligencia de negocios o business inteligence. (Massaron y Boschetti, 2018).

Para (Churpeck, 2018), engloba a la ciencia de datos o data science como un conglomerado de reglas básicas que respaldan y orientan la obtención y tratamiento de datos, fundamentados en conocimientos y principios de manipulación de información.

A. DATOS

(DAMA, 2017), define como dato a cualquier representación de entidades o hechos que se haga mediante textos, números, gráficos, imágenes, sonido o video.

B. TIPO DE DATO

Ya desde (Cardelli, 1985), se tiene una definición acerca de este tema; indica, en síntesis, que el tipo no es más que una característica, un atributo informático, que posee cualquier dato, pero es importante pues le indica a la máquina, programa, algoritmo y/o programador información del dato con el que trabaja; en un lenguaje operativo, el tipo de dato es un espacio reservado en la memoria que cuenta con restricciones (inherentes al tipo de dato) y que puede ser manipulado con una finalidad.

C. CICLO DE VIDA DEL DATO

(DAMA, 2017), presenta el siguiente modelo lineal del ciclo de vida de los datos (véase la figura 1), afirmando que primero se crean o se adquieren los datos, luego se almacenan, se mantienen y se usan, para finalmente destruirse; en la mayoría de veces, los datos permanecen almacenados un buen tiempo, pero al final todos los datos terminan eliminándose, ya sea porque fueron reemplazados por unos nuevos, porque no aportan información valiosa o porque simplemente ya no son usados.

Figura 1: Ciclo de vida del dato. (DAMA, 2017)

(McGilvray, 2008), propone el modelo POSMAD (véase la figura 2) que contiene las siguientes etapas:

- Plan (Planificar): Se prepara los datos que se deben utilizar.
- Obtain (Obtener): Se obtienen los datos que se tienen que utilizar.
- Store and Share (Almacenar y Compartir): Se mantiene una copia de los datos física o electrónicamente, y se hacen disponibles a los usuarios a través de un método de distribución.
- Maintain (Mantener): Se asegura que los recursos de los datos estén disponibles.
- Apply (Aplicar): Se usan los datos para alcanzar el objetivo.
- Dispose (Eliminar): Se descartan los datos que ya no se van a usar más.

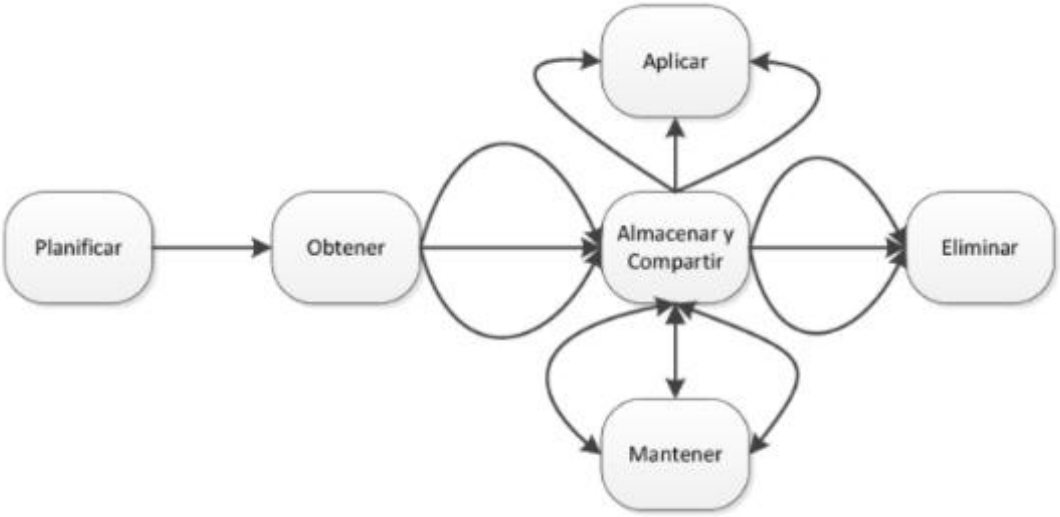


Figura 2: Flujo de datos en el modelo POSMAD. (McGilvray, 2008)

D. DATOS ESTRUCTURADOS

Para (Ozdemir, 2016), los datos estructurados son aquellos que pueden ser organizados en una tabla, pues al ser presentados en filas y columnas, es mucho más fácil para las personas realizar observaciones y análisis, además que recalca la sencillez del manejo y trabajo con un dato estructurado; podemos añadir que la mayor parte de la información es almacenada en datos no estructurados, guardados sin ningún criterio, convirtiéndoles en datos inservibles a menos que implementen una estrategia de transformación hacia datos estructurados.

2.2.2 MACHINE LEARNING

(Samuel, 1959), lo define como el campo de estudio que le da a las computadoras la habilidad para aprender sin ser explícitamente programadas.

(Mathews, 2019), centra al Machine Learning como el equipamiento de programas con una base histórica de datos con los que pueda trabajar de manera autónoma, teniendo como propósito la solución de problemas; además que dicho programa pueda retroalimentarse, aprendiendo y mejorando su eficiencia a través del tiempo; hace hincapié que el programa debe ser independiente y que aprenda por sí mismo.

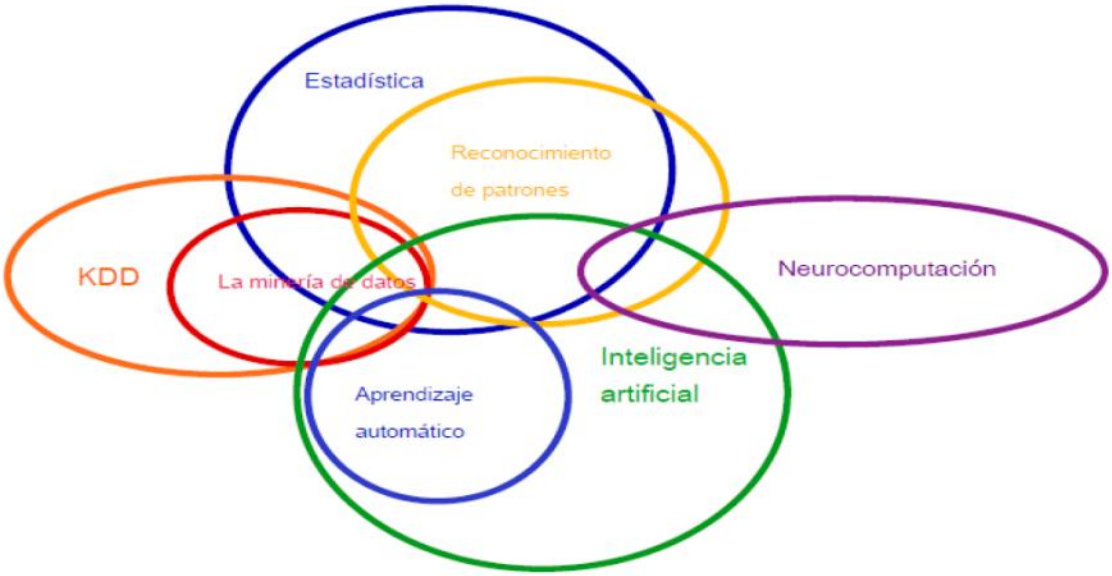


Figura 3: Componentes del Machine Learning. (Mathews, 2019)

2.2.2.1 ESTRUCTURA

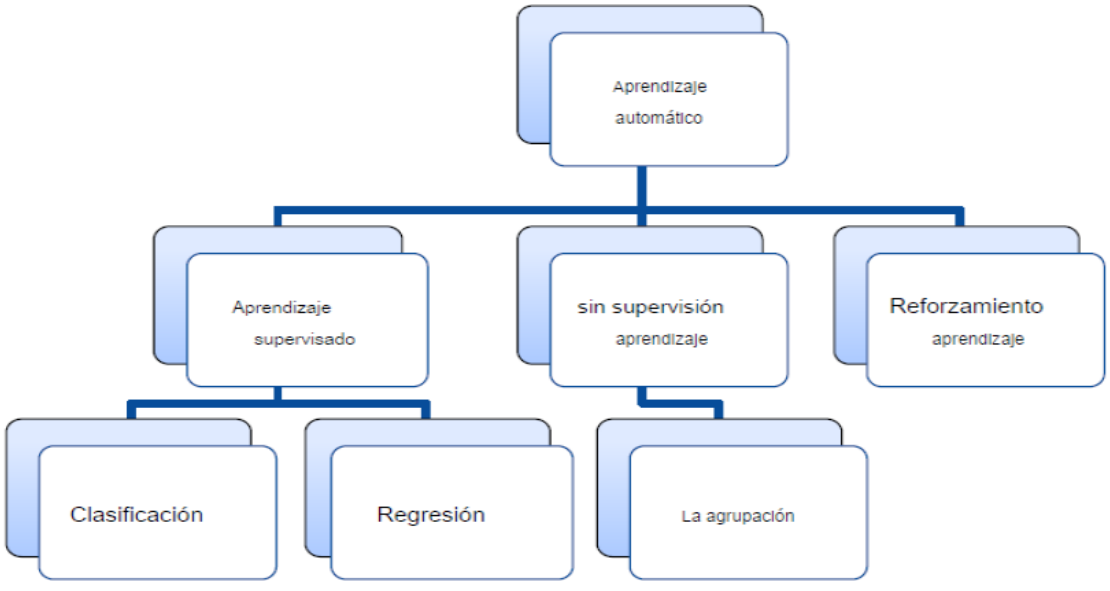


Figura 4: Ubicación jerárquica del Machine Learning. (Huddleston y Brown, 2018)

A. APRENDIZAJE SUPERVISADO

(McGlade y Scott-Hayward, 2019), indican que el aprendizaje supervisado es aquel que realiza su predicción empleando datos registrados o etiquetados para obtener un resultado que también se irá acumulando; es decir que se trabaja con datos de entrada que producirán un resultado saliente.

(Ozdemir, 2016), detalla un poco más, pues indica que el trabajo se realiza con dos tipos: los datos o variables predictoras y la variable respuesta, que en datos estructurados vendría a ser la columna objetivo; eso implica que, en un determinado momento del aprendizaje supervisado, el modelo sólo trabajará con las variables predictoras para obtener un resultado; aquí se puede observar una desventaja, pues se requiere que los base de datos tengan la variable a predecir dentro de sus columnas; de manera general, la forma de trabajo es empelar a las variables predictoras como información de entrenamiento y preparación, para luego realizar una predicción con información nueva.

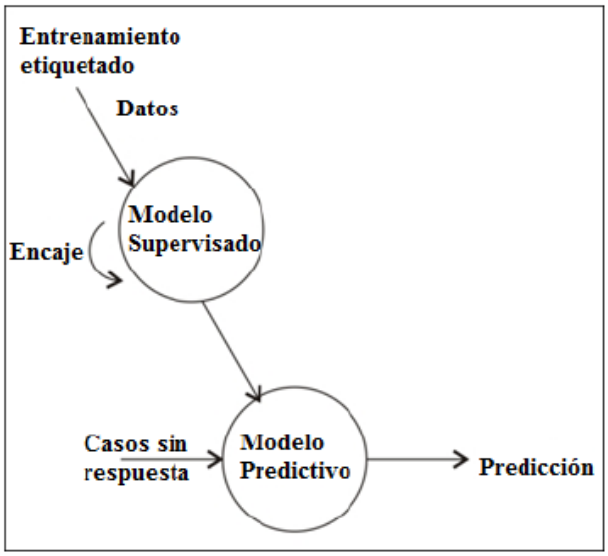


Figura 5: Forma de trabajo del aprendizaje supervisado. (Ozdemir, 2016)

ALGORITMOS DE CLASIFICACIÓN

(Huddleston y Brown, 2018), señalan que absolutamente todos los datos pueden convertirse en datos estructurados, es decir que agruparlos, clasificarlos y etiquetarlos no es algo imposible; es en este grupo de datos en donde los algoritmos de clasificación son utilizados para crear modelos predictivos los cuales tienen múltiples aplicaciones, por ejemplo, reconocer si un correo electrónico es de tipo no deseado.

Los algoritmos de clasificación forman parte del aprendizaje supervisado, teniendo como finalidad la predicción de nuevas etiquetas, resultantes de datos pasados o que fueron tomados con anterioridad, dichas etiquetas tienen valor del tipo discreto además que estar desordenados. (Raschka & Mirjalili, 2017, p.41)

Según (González, 2018), “los algoritmos de clasificación se usan cuando el resultado deseado es una etiqueta discreta, en otras palabras, son útiles cuando la respuesta al problema cae dentro de un conjunto finito de resultados posibles”.

Estos algoritmos son empelados para solucionar problemas donde se conoce de antemano los datos y que estos deben estar agrupados por clases para que el reconocimiento de patrones sea eficiente; en consecuencia, la clasificación o mejor dicho el análisis de nuevos datos depende mucho del conocimiento que haya obtenido de un grupo previo de datos, es decir que el resultado dependerá de los datos almacenados y del criterio que se tuvo al momento de tratarlos. (ECURED, 2016).

Para (Sierra, 2008), estos algoritmos deben contar con una información previa que sirva de conocimiento inicial, y que estos se encuentren clasificados es decir que sean datos estructurados agrupados en categorías, o por características comunes.

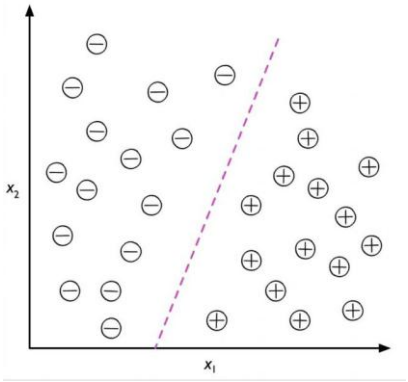


Figura 6: Algoritmo de clasificación (Raschka & Mirjalili, 2017)

REGRESIÓN LOGÍSTICA

(McGlade y Scott-Hayward, 2019), señalan que una regresión lineal es la asociación de dos variables que van a tomar el papel de ser estudiadas, para obtener una predicción de un problema.

(Igual y Seguí, 2017), indican que este tipo de algoritmo, emplean datos cuantitativos para realizar sus predicciones, el resultado es una probabilidad perteneciente a un tipo o clase.

Para (González, 2019), es un procedimiento que la estadística posee para la predicción de una clase binaria, dando como variable resultante un resultado dicotómico, es decir que sólo puede tener dos posibles valores; este algoritmo de

aprendizaje es usado en modelos para predecir una enfermedad, o para determinar si un hecho o evento va a suceder.

(Raschka y Mirjalili, 2017), nos explican matemáticamente las fórmulas de la regresión logística; sea razón de probabilidades:

$$\frac{P}{(1-P)}$$

Dónde la variable "P" representa la probabilidad del acontecimiento que sí va a suceder o un evento positivo, un ejemplo de este tipo de suceso, sería si una persona se enferma; podemos pensar en el evento positivo como etiqueta de clase Y=1. Entonces podemos definir la función logística, que es simplemente el logaritmo de la razón de posibilidades:

$$logist(P) = log\frac{p}{1-p}$$

Donde "log" es el logaritmo natural; se debe indicar que la función logist toma al 0 y 1 como únicos valores o datos de entrada y los transforma en valores en todo el rango de números reales, que podemos usar para expresar una relación lineal entre los valores de las características y las probabilidades de registro:

$$logist\left(P\left(Y=\frac{1}{X}\right)\right) = \sum_{i=0}^{m} w_i X_i = w^t X$$

Donde la probabilidad condicional es la que se encuentra dentro del paréntesis:

$$P(Y=\frac{1}{X})$$

Esta probabilidad indica la posibilidad que una muestra en especial sea parte de la clase o tipo 1 dadas sus características X, entonces la probabilidad que una determinada sea incluida en un tipo en particular es la función logística sigmoidea, que simplemente es una versión inversa de la función logist:

$$\text{ø}(\text{z}) = \frac{1}{1+e^{-z}}$$

Donde z es la entrada pura; es decir es la combinación lineal de los pesos (w) con las características de muestra (X), que al resolverlo se despliega una sumatoria:

$$z = w^t X = w_0 x_0 + w_1 x_1 + \cdots + w_m x_m$$

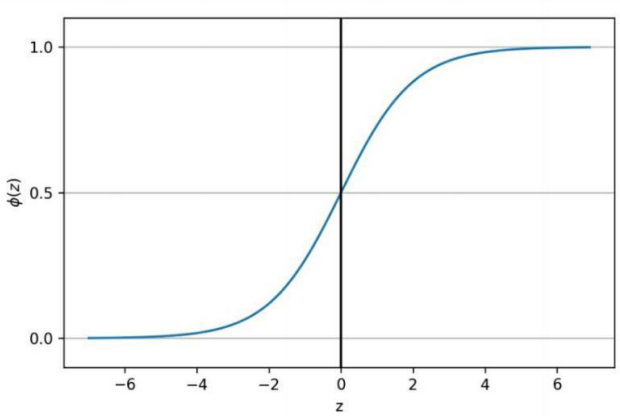


Figura 7: Gráfica de una función logística. (Raschka & Mirjalili, 2017)

La función sigmoide toma valores de números reales como entrada y los transforma en valores en el rango [0, 1] con una intersección en ø(z)=0.5, entonces la probabilidad pronosticada se puede convertir simplemente en un binario, resultado a través de una función umbral $y = \begin{cases} 1, si\ \text{ø}(\text{z}) \geq 0.5 \\ 0, si\ \text{ø}(\text{z}) < 0.5 \end{cases}$ (Raschka & Mirjalili, 2017, p.121).

K VECINOS CERCANOS

Según (González, 2019), “es un algoritmo de aprendizaje no paramétrico, esto significa que no hace suposiciones para la distribución de datos subyacentes. En otras palabras, la estructura del modelo es determinada a partir del conjunto de datos. Esto es muy útil en la práctica donde la mayoría de los conjuntos de datos del mundo real no siguen suposiciones teóricas matemática”.

Para (Raschka y Mirjalili, 2017), “Es un ejemplo típico de un estudiante perezoso, se llama perezoso no por su aparente simplicidad, sino porque no aprende una función discriminatoria de los datos de entrenamiento, pero memoriza el conjunto de datos de entrenamiento en su lugar, el algoritmo KNN encuentra las k muestras en el conjunto de datos de entrenamiento que son más cercanos (más similares) al punto que queremos clasificar, la etiqueta de clase del nuevo punto de datos se determina por un voto mayoritario entre sus k vecinos más cercanos”.

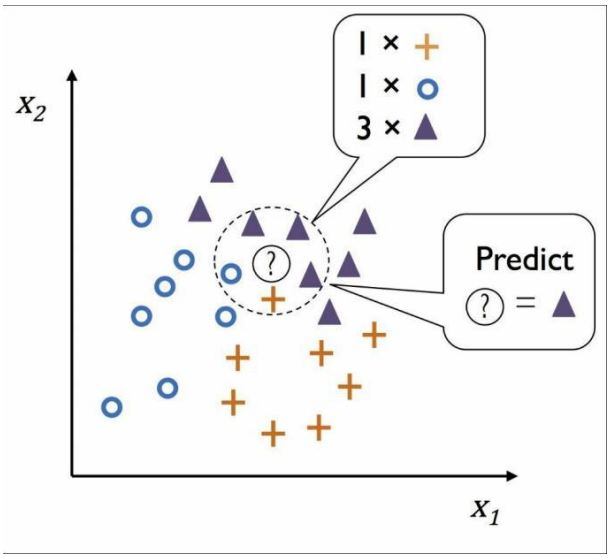


Figura 8: Gráfica k-vecinos cercanos. (Raschka & Mirjalili, 2017)

MÁQUINA DE SOPORTE VECTORIAL

Según (González, 2019), “buscan la línea que mejor separa dos clases. Las características de datos que están más cerca de la línea que mejor separa las clases se denominan vectores de soporte e influyen en la ubicación de la línea. De particular importancia es el uso de diferentes funciones del kernel a través del parámetro del kernel”.

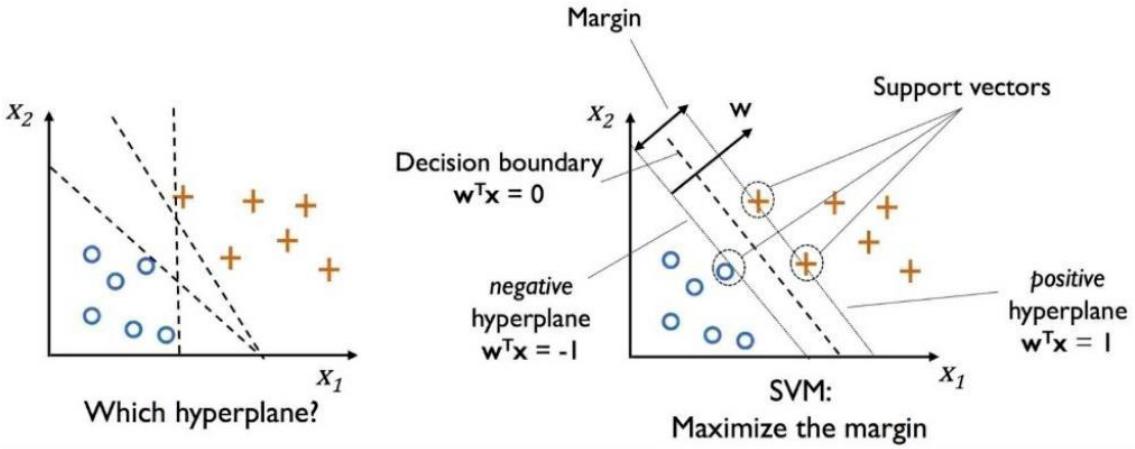


Figura 9: Figura máquina de soporte vectorial. (Raschka & Mirjalili, 2017)

Según Raschka & Mirjalili (2017), en máquina de soporte vectorial nuestro objetivo de optimización es maximizar el margen. El margen se define como la distancia entre el hiperplano de separación (límite de decisión) y el entrenamiento muestras que están más cerca de este hiperplano, que son los llamados vectores de soporte, La razón de tener límites de decisión con grandes márgenes es que tienden tener un error de

generalización más bajo, mientras que los modelos con márgenes pequeños son más propensos al overfitting".

ÁRBOL DE DECISIÓN

Según (González, 2019), "es un tipo de algoritmo de aprendizaje supervisado que se utiliza principalmente en problemas de clasificación, aunque funciona para variables de entrada y salida categóricas como continuas. El árbol de decisión identifica la variable más significativa y su valor que proporciona los mejores conjuntos homogéneos de población, todas las variables de entrada y todos los puntos de división posibles se evalúan y se elige la que tenga mejor resultado".

"Podemos pensar en este modelo como desglosando nuestros datos al tomar una decisión basada en una serie de preguntas" (Raschka & Mirjalili, 2017, p.155).

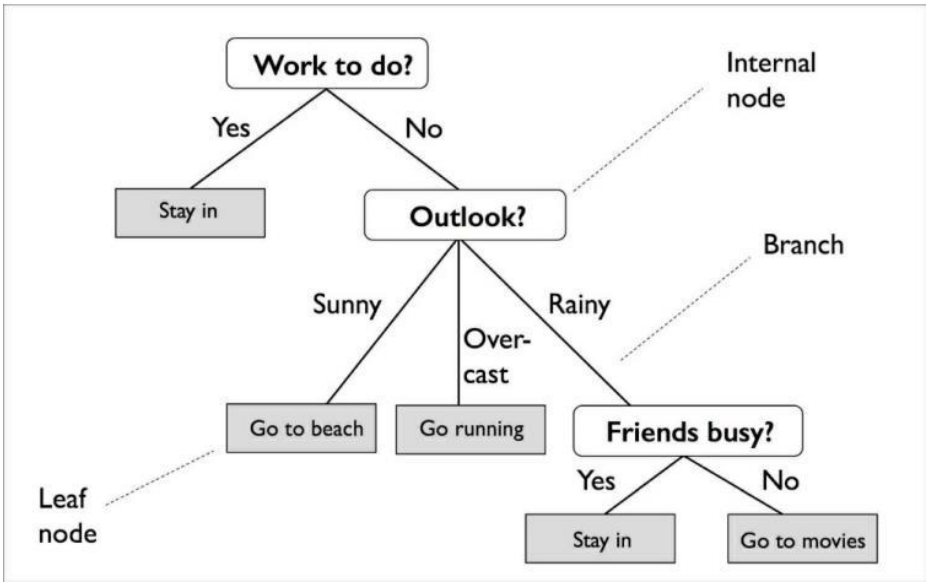


Figura 10: Figura Árbol de decisión. (Raschaka & Mirjalili, 2017)

BOSQUES ALEATORIOS

Según (González, 2019), es la técnica del Machine Learning con buena versatilidad; ya que puede realizar múltiples métodos como son: la reducción dimensional, el tratamiento de datos perdidos, el manejo de datos atípicos o raros, y otras funciones más que son básicas en la exploración de información, realizando un buen trabajo. Los bosques aleatorios o random forest realiza su aprendizaje mediante grupos o conjuntos, siendo cada uno de estos conformado por modelos básicos o débiles pero que al combinarlos se conforma un modelo más eficiente.

(Igual y Seguí, 2017), también hablan acerca del Random Forest, y los definen como un método de conjunto, emplea estrategias o técnicas basadas en la agregación para combinar a los distintos clasificadores; el problema del sobreajuste es resuelto con ayuda de las propiedades del random forest.

Un bosque aleatorio puede considerarse como un conjunto de árboles de decisión. La idea detrás de un bosque aleatorio es promediar múltiples árboles de decisión que sufren individualmente una alta varianza, para construir una más robusta modelo que tiene un mejor rendimiento de generalización y es menos susceptible a overfitting. (Raschka & Mirjalili, 2017, p.172).

(Massaron y Boschetti, 2018), analizaron e indicaron que la idea de ensamblar los diferentes árboles de decisión en grupos pequeños y estos que pertenezcan a uno cada vez más grande, vendría a ser el secreto de los resultados satisfactorios que se obtienen; la idea de que cada árbol de decisión ofrezca un resultado y al final combinar dichas respuestas genera un beneficio a largo plazo, pues el aprendizaje es equivalente al número de árboles que se posea y el tiempo que lleve alimentándose de nueva información.

B. APRENDIZAJE SIN SUPERVISACIÓN

(Ozdemir, 2016), indica que a diferencia del aprendizaje supervisado en el no supervisado los modelos, que son llamados sin supervisión, trabajan con datos o variables desconocidas, por tanto, no es usado en solucionar problemas de predicción sino de estimación; pero, sí puede encontrar patrones y también puede agrupar los datos con que trabaja empleando un método de condensación.

(Igual y Seguí, 2017), añaden que el inconveniente que presenta este tipo de aprendizaje es el obvio trabajo con variables que no conoce y, que, a diferencia de trabajar con datos estructurados, la manipulación y el tratamiento para hallar un patrón o comportamiento oculto es más complicado, por tanto, los algoritmos de aprendizaje tendrán una clara desventaja al momento de aprender de este tipo de información.

En la imagen N.º 2.8, nos muestra la secuencia del proceso de un aprendizaje no supervisado, nos indica que del grupo de datos que se tiene, primero se debe seleccionar el grupo de datos que van a ser los predictores o de entrenamiento, luego los modelos emplearán dos tipos de técnicas, una es de los clústeres y la otra es la reducción de las dimensiones de los datos. Pero en todo el proceso lo que se está

realizando son sugerencias no acciones concretas, sino que hace estimaciones para encontrar similitudes y/o diferencias, por tanto, el criterio de un intérprete es necesaria, para que el modelo pueda continuar de manera correcta con su aprendizaje.

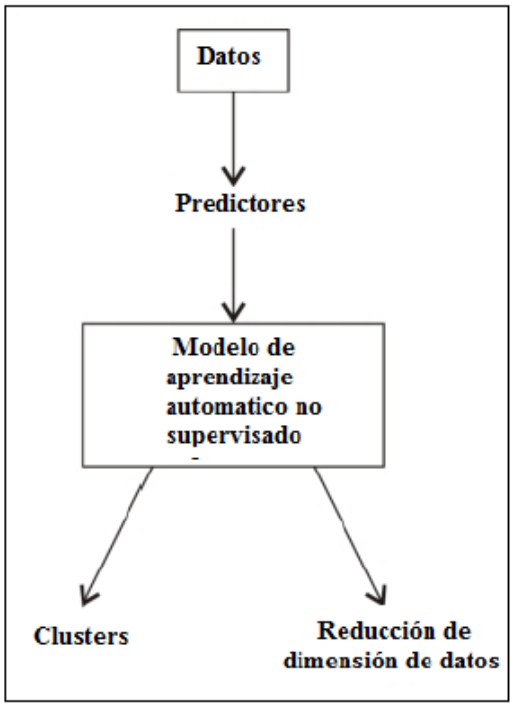


Figura 11: Procesos del aprendizaje no supervisado. (Ozdemir, 2016)

C. REFORZAMIENTO DE APRENDIZAJE

(Huddleston y Brown, 2018), explican que el reforzamiento es cuando el modelo, que ya es independiente, se comunica con su entorno para reafirmar sus conceptos tomando decisiones de acuerdo a sus algoritmos que lo conforman.

2.2.2.2 ALGORITMOS DE APRENDIZAJE

(Stephen, 2014), clasifica a los algoritmos de aprendizaje de acuerdo a su finalidad, por ejemplo, hay algoritmos que están destinados para solucionar problemas de aprendizaje supervisado, por tanto, esos son llamados algoritmos supervisados, normalmente los algoritmos pertenecen a un sólo tipo de aprendizaje, pero existen algunos que pueden ser usados en ambos tipos de aprendizaje.

2.2.2.3 MODELO DE MACHINE LEARNING

Según (Hurwitz y Kirsch, 2018), definen que el modelo Machine Learning como una secuencia de acciones que se deben realizar para generar una salida cuando se entrena a su algoritmo de machine learning con datos. También los centran como parte de la Inteligencia Artificial, ya que admite el aprendizaje con datos, pero dicho

aprendizaje no es sencillo como simplemente seguir y seguir entrenando a un algoritmo, sino que el criterio para cambiar las variables, ayuda a mejorar la precisión del modelo.

Según (González, 2019), lo define como un proceso, cuya finalidad es encontrar el patrón de comportamiento en una base de datos; indica que una vez hallado dichos patrones, es cuando recién el modelo puede hacer una predicción, esta vez empleando nueva información.

(Bagnato, 2017), agrupa al Machine Learning dentro de la IA (Inteligencia artificial) como un subgrupo, y que dentro se busca la solución del aprendizaje automático con la capacidad de retroalimentarse, dicho problema debe resolverse mediante programas; esto implica que el modelo resultante no debería necesitar de un programador explícito, pues las reglas ya deben estar fijadas y se debe seguir los pasos establecidos.

A. TÉCNICA PREDICTIVA

Para (Rouse, 2017), esta técnica es avanzada pues su análisis es complejo, requiriendo de datos históricos para aprender y empleando información nueva para su predicción; dentro de su estructura de análisis están las técnicas estadísticas y los algoritmos de aprendizaje, empleados para hallar un comportamiento, patrón o tendencia, todo para que nos dé un resultado probable sobre un suceso futuro.

Para (Calabrese, 2012), es el procedimiento o conjunto de procedimientos que tienen como objetivo obtener un resultado predictivo. También podemos decir que se trata de el o los procedimientos puestos en práctica al realizar una actividad (construir algo, efectuar una medición o un análisis), así también como la pericia o capacidad que se ponen de manifiesto cuando se realiza una actividad.

(Pérez y Merino, 2008), la definen como un conjunto de pasos definidos los cuales tienen como objetivo encaminarse hacia un objetivo común, y que en futuras situaciones se repita el procedimiento siguiendo dichos procedimientos pautados, produciendo el mismo resultado; en síntesis, se resume en la repetición reglamentada, siguiendo las acciones predefinidas para realizar una correcta ejecución ordenada.

En (Management Solutions, 2018), maneja el concepto de técnica como una agrupación de normas que tienen la capacidad de localizar un patrón de manera automática, la utilización de dichos patrones es para realizar una predicción.

(Hurwitz y Kirsch, 2018), indican que la técnica es una ejecución sistemática cuyo fin es optimizar los recursos empleados, reduciendo la cantidad consumida, pero logrando el mismo objetivo; recalca que la técnica alcanza las metas establecidas, pero con la menor cantidad de costo, añade que el testeo es fundamental para corregirse continuamente, pues la prueba y el error son los que van a determinar si los procedimientos continúan establecidos o si se deben cambiar por otros más efectivos

B. ALGORITMO PREDICTIVO

(IBM, 2015), indica que el algoritmo es estudiado por la algoritmia; lo define como un grupo finito y ordenado de pasos reescritos pero definidos y bien establecidos cuyo fin es la realización de una actividad mediante la secuencia de dichos pasos que no ocasiones ninguna duda al ejecutante sobre su ejecución sino más bien lo esclarezca; los algoritmos poseen dos estados, uno es el inicial donde los datos de entrada siguen la secuencia establecida a fin de llegar al segundo estado que es el de salida, el cual dependerá del recorrido del algoritmo.

Para (Juan, 2020), el algoritmo es una agrupación limitada de instrucciones estructuradas con un sentido lógico que solucionan un problema determinado; se trata de llegar a la respuesta siguiendo una serie de pasos o reglas secuenciales las cuales fueron analizadas y establecidas con anterioridad.

Según (Liu, 2019), se trata de una serie de pasos secuenciales, las cuales cumplen la función de realizar procesos determinados para dar respuesta a la interrogante formulada, es decir es un grupo ordenado y limitado de instrucciones que permiten solucionar diferentes problemas o llegar a una conclusión.

(Aritmetrics, 2015), afirma: “secuencia de pasos basado en estadísticas inferenciales, que se utiliza para predecir la respuesta a una promoción de marketing o a una determinada inversión”.

C. VALIDEZ DEL MODELO

Según (García y Molina, 2012), la validez se mide de acuerdo al nivel de confianza que el modelo tiene al momento de cumplir el fin por el que fue realizado; el resultado

de dicha validez no es una “demostración de que sí funciona” sino que por medio de diferentes resultados obtenidos con una muestra de prueba es la que nos proporcionará esa “confianza”, por ende, los testeos son importantes pues nos visualizará lo negativo y lo positivo del modelo para ejecutar las modificaciones debidas.

Según (Torres, 2013), “se determina si el modelo de simulación construido refleja correctamente el modelo conceptual diseñado. Es decir, después de haber terminado la construcción del modelo es necesaria la comprobación, la confirmación de que el modelo trabaja correctamente”.

(Roman, 2019), inicia con la frase “la exactitud es la ausencia del error” para incidir que la exactitud demostrada siempre va a mostrar error ya que todos los modelos creados por el hombre son creaciones incompletas, son representaciones que imitan la situación real, así que una exactitud del 100% no es posible pero aún si los modelos no tienen esa exactitud, igual son muy útiles Se puede encontrar información general sobre el trabajo con modelos matemáticos en diversos libros de texto teóricos y aplicados.

Según (Kumar, 2017), la cuantificación del nivel de incierto e incertidumbre es lo que la validación realiza, el proceso es comparar las predicciones obtenidas del modelo con la información real para encontrar el nivel de aciertos que obtuvo, para esto, es claro que los datos de testeo nunca deben ser los mismos que entrenaron a dicho modelo, pues se caería en la mediocridad; la validez depende de la pureza de los datos pero casi siempre son los mismos datos los que tienen errores, por tanto es suficiente demostrar pequeñas diferencias entre los datos de muestra y los del modelo.

2.2.3 METODOLOGÍA FUNDAMENTAL PARA LA CIENCIA DE DATOS

La metodología fundamental para la ciencia de datos sirve como “estrategia de guía para resolver problemas y esta es independiente de tecnologías o herramientas particulares, que proporciona un marco para continuar con los métodos y procesos que se utilizarán para obtener respuestas y resultados” (Romero, 2019).

Para (IBM, 2015), la metodología fundamental es “capaz de proporcionar una estrategia de guía, independientemente de las tecnologías, los volúmenes de datos o los enfoques involucrados. Esta metodología tiene algunas similitudes con las reconocidas metodologías para la extracción de datos, pero enfatiza varias de las

nuevas prácticas en la ciencia de datos, como el uso de grandes volúmenes de datos, la incorporación de análisis de texto en el modelado predictivo y la automatización de algunos procesos. La metodología consta de 10 etapas que forman un proceso iterativo para el uso de datos y cada etapa juega un papel vital en el contexto de la metodología general".

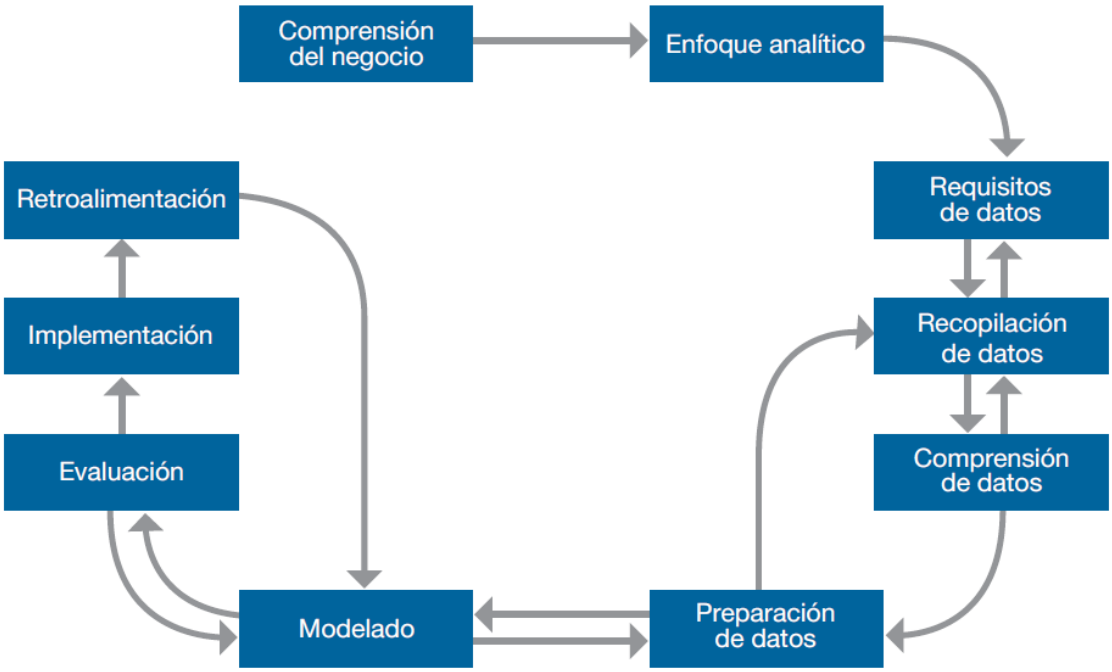


Figura 12: Metodología fundamental para la ciencia de datos. (IBM, 2015)

2.2.3.1 ETAPAS

A. COMPRENSIÓN DEL NEGOCIO

Es la primera etapa de la metodología, es donde se define "el problema, los objetivos del proyecto y los requisitos de la solución desde una perspectiva empresarial". Esta información es necesaria para que la solución del problema sea exitosa. "Para ayudar a garantizar el éxito del proyecto, los promotores deben participar mientras dure el proyecto para proporcionar experiencia en el dominio, revisar los hallazgos intermedios y garantizar que el trabajo siga su curso para generar la solución deseada" (IBM, 2015, p. 03).

(Chapman et al. 2000), se refiere como:

La comprensión a fondo, "desde una perspectiva de negocio, lo que el cliente realmente quiere lograr. A menudo el cliente tiene muchos objetivos y restricciones que compiten que deben ser correctamente equilibrados. El objetivo del analista debe destapar factores importantes en el principio del proyecto esto puede influir en el

resultado final. Una consecuencia probable de descuidar este paso debe ser a expensas de un dar un gran esfuerzo de producir las respuestas correctas a las preguntas incorrectas".

"Es el punto en que las soluciones se especifican. Como resultado de sus investigaciones y reuniones, debe crear un objetivo principal concreto acordando con los patrocinadores del proyecto y otras unidades comerciales que se vean afectadas por los resultados" (IBM, 2012).

B. ENFOQUE ANALÍTICO

(IBM, 2015) específica que en esta etapa implica:

"Expresar el problema bajo el contexto de las técnicas estadísticas y de aprendizaje automático, para que la organización pueda identificar las más adecuadas para el resultado deseado". Está relacionado con la pregunta ¿Cómo puedes usar los datos para responder la pregunta?, teniendo así un enfoque analítico descriptivo, de diagnóstico, predictivo o prescriptivo. Una vez establecida la directriz se utilizará herramientas estadísticas y de Machine Learning (identifica relaciones y tendencias en datos que de otra manera no serían accesibles o identificados). (p. 03)

C. REQUISITOS DE DATOS

"El enfoque analítico elegido determina los requisitos de datos. Más concretamente, los métodos analíticos a utilizar requieren de determinados contenidos de datos, formatos y representaciones, orientados por el conocimiento en el dominio" (IBM, 2015, p. 04).

Según (IBM, 2012):

"La mayoría de datos en la cantidad y calidad de los datos; la cantidad de datos disponible y el estado de los datos". Algunas características para describir datos: "

- Cantidad de datos. En la mayoría de técnicas de modelado, los tamaños de datos tienen un equilibrio relacionado. Los grandes conjuntos de datos pueden producir modelos más precisos, pero también pueden aumentar el tiempo de procesamiento. Considerar el uso de un subconjunto de datos. Incluyendo estadísticos de tamaños para todos los conjuntos de datos y tener en cuenta tanto el número de registros como los campos (atributos) cuando se describa los datos.

- Tipos de valores. Los datos pueden incluir una variedad de formatos, como numérico, categórico (cadena) o Booleano (verdadero/falso).
- Esquemas de codificación. Con frecuencia, los valores de la base de datos son representaciones de características como género o tipo de producto. Registrar los esquemas incoherentes en el informe de datos."

D. RECOPILACIÓN DE DATOS

Para (IBM, 2015) "en esta etapa inicial de recopilación de datos, los científicos de datos identifican y reúnen los recursos de datos disponibles (estructurados, no estructurados y semi estructurados) y relevantes para el dominio del problema" (p. 04).

"Se adquieren en el proyecto los datos, listados en los recursos del proyecto. Esta colección inicial incluye carga de datos, si es necesario para la comprensión de los datos" (Chapman et al., 2000).

Según (IBM, 2012), las fuentes de origen de los datos pueden ser: "

- Datos existentes: Incluye una amplia variedad de datos, como los datos transaccionales, datos de encuesta, registros Web, etc. Tener en cuenta si los datos existentes son suficientes para adaptarse a sus necesidades.
- Datos adquiridos: Si no se utiliza los datos adicionales, considere lo necesario.
- Datos adicionales: Si las fuentes anteriores no satisfacen las necesidades, es posible realizar encuestas o realizar seguimientos adicionales para servir de complemento a los almacenes de datos actuales."

E. COMPRENSIÓN DE DATOS

(IBM, 2015) indicó que, una vez recolectada la información inicial necesaria, "los científicos de datos suelen utilizar estadísticas descriptivas y técnicas de visualización para comprender el contenido de los datos, evaluar su calidad y descubrir insights iniciales sobre ellos" (p. 04).

Según (Chapman et al. 2000), hace referencia como:
"El entendimiento de datos y todas aquellas actividades relacionadas con la limpieza de datos, identificación de problemas vinculados con la toma de datos, procedimientos para determinar la calidad de datos y todo lo tendiente a facilitar la familiarización con los datos. A partir de esta etapa se determinan los primeros subconjuntos de datos que pueden contener la información que se está buscando".

F. PREPARACIÓN DE LOS DATOS

Esta etapa abarca “todas las actividades para construir el conjunto de datos que se utilizará en la subsiguiente etapa de modelado. Entre las actividades de preparación de datos están la limpieza de datos (tratar con valores no válidos o que faltan, eliminar duplicados y dar un formato adecuado), combinar datos de múltiples fuentes (archivos, tablas y plataformas) y transformar los datos en variables más útiles” (IBM, 2015, p. 04).

“La preparación de los datos es uno de los aspectos más importantes de machine learning Ya que se tiene que preparar, empaquetar y clasificar los datos para el modelado” (IBM, 2012).

“La limpieza de los datos implica observar más de cerca los problemas en los datos que ha sido seleccionado para los análisis” (IBM, 2012).

“Esta etapa también incluye la construcción de operaciones de preparación de datos tales como la producción de atributos derivados, completar registros nuevos, o transformar valores para atributos existentes” (Chapman et al., 2000).

“Transformar formateando se refiere principalmente a modificaciones sintácticas hechas a los datos que no cambian su significado, pero podría ser requerido por la herramienta de modelado” (Chapman et al., 2000).

En la presente etapa también incluye la división de datos en donde se fragmenta “el conjunto total de datos en tres partes: conjuntos de entrenamiento, validación y pruebas”. “Entrenamos el modelo con datos de entrenamiento, se comprueba el aprendizaje con datos de validación, y finalmente, una vez está listo, se valida con el conjunto de datos de prueba”. La creación del modelo conlleva un ajuste que vendría a ser la selección de algunos datos que serán los “hiper parámetros”, los cuales no son iguales a los parámetros con los que ya cuenta el modelo; dicho “ajuste” se ejecuta con el aprendizaje obtenida de la data de validación, y tiene como finalidad que “el modelo pueda generalizar bien con datos no procesados aún, en otras palabras, predecir con exactitud resultados con datos nuevos, basados en sus parámetros internos ajustados mientras el modelo fue entrenado y validado” (Roman, 2019).

G. MODELADO

Para (IBM, 2015) la etapa de modelado “utiliza la primera versión del conjunto de datos preparado y se enfoca en desarrollar modelos predictivos o descriptivos según el enfoque analítico previamente definido. En los modelos predictivos, los científicos de datos utilizan un conjunto de capacitación (datos históricos en los que se conoce el resultado de interés) para construir el modelo” (p. 04).

Según (IBM, 2012), se debe tener “la preparación suficiente para construir el modelo, considerar el tiempo necesario para experimentar con diferentes modelos antes de llegar a conclusiones definitivas. En el análisis se suelen generar varios modelos y comparar los resultados antes de aplicarlos”. Durante el proceso de creación de cualquier modelo, se administrará o manejará tres diferentes tipos de datos, los cuales afectarán las futuras decisiones: “

- Configuración de parámetros incluye las notas que ha tomado sobre los parámetros que producen los mejores resultados.
- Los modelos reales producidos.
- Descripciones de resultados de modelos, incluyendo problemas de datos y rendimiento que hayan ocurrido durante la ejecución del modelo y exploración de los resultados.”

“En el modelado se suele ejecutar en múltiples iteraciones. Normalmente, los analistas de datos ejecutan varios modelos utilizando los parámetros por defecto y ajustan los parámetros o vuelven a la fase de preparación de datos para las manipulaciones necesarias por su modelo” (IBM, 2012).

H. EVALUACIÓN

“El modelo debería ser evaluado para asegurar que se encontró criterio de éxito de la minería de datos y aprobar los criterios de prueba deseados. Este es una evaluación puramente técnica basada en el resultado de las tareas de modelado” (Chapman et al., 2000).

(IBM, 2015) hace referencia lo siguiente:

“Que durante el desarrollo del modelo y antes de su implementación, el científico de datos evalúa el modelo para comprender su calidad y garantizar que aborda el problema empresarial de manera adecuada y completa. La evaluación del modelo implica el cálculo de varias medidas de diagnóstico y de otros resultados, como tablas

y gráficos, lo que permite al científico de datos interpretar la calidad y la eficacia del modelo en la resolución del problema" (p. 05).

Según (Chapman et al. 2000), en esta etapa se evalúa lo siguiente:
"El grado al que el modelo encuentra los objetivos del proyecto, y procura determinar si hay alguna razón por el cual este modelo es deficiente. Los resultados del modelo cubren los modelos que están relacionados con los objetivos originales de negocio y todas las demás conclusiones".

"Al tener un conjunto de modelos iniciales, se debe observar detenidamente para determinar cuáles son los más precisos o eficaces para considerar al final". En este punto se debe utilizar gráficos de evaluación o nodos de análisis para analizar la efectividad de los resultados. (IBM, 2012).

"La evaluación trata con factores como la exactitud y la generalidad del modelo. Este paso evalúa el grado en el que el modelo encuentra los objetivos de negocio, y procura determinar si hay alguna razón de negocio por la que este modelo sea deficiente. Se comparan los resultados con los criterios de evaluación definidos en el proyecto" (Chapman et al., 2000).

I. IMPLEMENTACIÓN

"Cuando el modelo satisfactorio ha sido desarrollado y aprobado por los promotores del negocio, se implementa en el entorno de producción o en un entorno de pruebas comparable. Por lo general, se implementa de forma limitada hasta que su rendimiento se haya evaluado completamente" (IBM, 2015, p. 05).

J. RETROALIMENTACIÓN

"Al recopilar los resultados del modelo implementado, la organización obtiene retroalimentación sobre el rendimiento del modelo y su impacto en el entorno en el que se implementó" (IBM, 2015, p. 05).

2.2.3.2 MÉTRICAS DE EVALUACIÓN DEL MODELO

A. MATRIZ DE CONFUSIÓN

Barrero (2019) la matriz de confusión:
"Es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las

instancias en la clase real. Uno de los beneficios de las matrices de confusión es que facilitan ver si el sistema está confundiendo dos clases. Si en los datos de entrada el número de muestras de clases diferentes cambia mucho la tasa de error del clasificador no es representativa de lo bien que realiza la tarea el clasificador. Si por ejemplo hay 990 muestras de la clase 1 y sólo 10 de la clase 2, el clasificador puede tener fácilmente un sesgo hacia la clase 1."

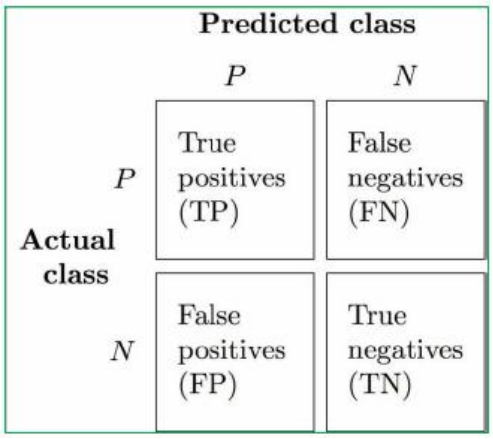


Figura 13: Matriz de confusión. (Raschaka & Mirjalili, 2017)

Barrios (2019) indica que "a partir de las cuatro variables se construyen otro conjunto de métricas muy ligadas a la matriz de confusión", estas son:

EXACTITUD: "Se refiere a lo cerca que está el resultado de una medición del valor verdadero. En términos estadísticos, la exactitud está relacionada con el sesgo de una estimación. También se conoce como Verdadero Positivo."

$$\frac{(VP + VN)}{(VP + FP + FN + VN)}$$

PRECISIÓN: "Se refiere a la dispersión del conjunto de valores obtenidos a partir de mediciones repetidas de una magnitud. Cuanto menor es la dispersión mayor la precisión. Se representa por la proporción entre el número de predicciones correctas (tanto positivas como negativas) y el total de predicciones."

$$\frac{VP}{(VP + FP)}$$

SENSIBILIDAD (RECALL): "También se conoce como Tasa de Verdaderos Positivos (True Positive Rate) ó TP. Es la proporción de casos positivos que fueron correctamente identificadas por el algoritmo."

$$\frac{VP}{(VP + FN)}$$

ESPECIFICIDAD: “También conocida como la Tasa de Verdaderos Negativos o TN. Se trata de los casos negativos que el algoritmo ha clasificado correctamente. Expresa cuan bien puede el modelo detectar esa clase.”

$$\frac{VN}{(VN + FP)}$$

B. CURVA ROC

Roc es el acrónimo de Receiver Operating Characteristic o Característica operativa del receptor, su gráfica cruza la sensibilidad y la especificidad. “El propósito es permitir que el espectador evalúe la precisión de la prueba M para cualquier posible valor de corte c. Esto ayuda a decidir qué corte usar en la práctica, comparando diferentes pruebas para la misma cosa y para evaluar la precisión general. El análisis de ROC proporciona herramientas para seleccionar modelos posiblemente óptimos y descartar modelos subóptimos independientemente de (y antes de especificar) el contexto de costos o la distribución de clases. El análisis de ROC se relaciona de manera directa y natural con el análisis de costo / beneficio de la toma de decisiones de diagnóstico” (Barrero, 2019).

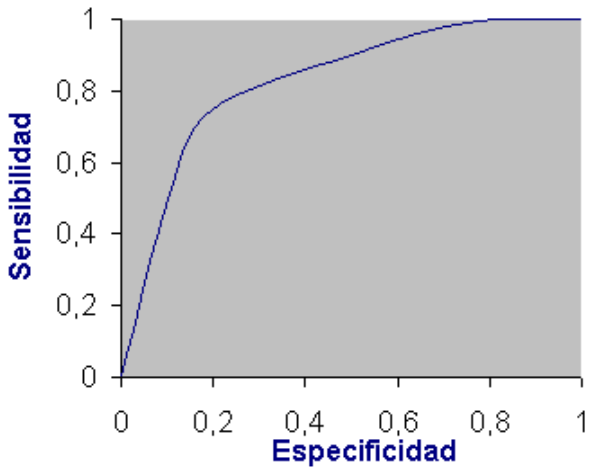


Figura 14: Curva ROC. (Barrero, 2019)

C. VALIDACIÓN CRUZADA

Su funcionamiento es dividir el total de información en dos conjuntos de k partes o pliegues, dicho valor de k es dado a criterio del analista; el primer grupo es formado por los k-1 partes o pliegues y los usa para entrenar al modelo, y con el último pliegue realiza la evaluación del desempeño. Este procedimiento se repite k veces para que

obtengamos k modelos y estimaciones de rendimiento. (Raschka & Mirjalili, 2017, p.157).

Figura 15: Esquema k-fold validación cruzada. (Raschka & Mirjalili, 2017)

2.2.4 RENDIMIENTO ACADÉMICO

(Hernández, 2015) nos lo define como el nivel que un alumno obtiene en su proceso de aprendizaje, el cual es medido mediante las calificaciones que se tiene en exámenes u otro ítem calificativo, siempre se realiza esta medición al final del ciclo, periodo u semestre académico cuando se tiene a la mano todas las calificaciones finales en todos los cursos, obteniendo así el rendimiento académico del estudiante; por tanto, es justo decir que el rendimiento es secuela del grado de motivación que posea el alumno, pero no es el único factor, pues la capacidad de aprender, la influencia del entorno, son algunos factores que también intervienen en el desenvolvimiento del estudiante.

(Álvaro, 1990) considera que el rendimiento académico es un constructor y que su delimitación conceptual es complicada, pues no sólo se trata de un factor sencillo, sino que se habla de una variable más compleja y multidimensional, pero en ciertas condiciones basta con tener sólo unos factores para determinarlo, pero en muchos casos se trata de una mezcla de muchos factores que incluso, algunos de ellos, no parecen influir; la idea es hallar los aspectos determinantes que definirán el rendimiento, puede que en algunos casos sólo sean individuales, en otros externos y también hay casos que son una combinación de ambos; pero se resume a no simplificarlo a un solo factor.

2.2.5 PYTHON

“Se trata de un lenguaje de programación multiparadigma, ya que soporta orientación a objetos, programación imperativa y, en menor medida, programación funcional. Es un lenguaje interpretado, dinámico y multiplataforma” (Fernández, 2009).

“Python es utilizado por más y más científicos de datos y desarrolladores para la construcción y análisis de modelos. Además, es un éxito entre los principiantes que son nuevos en Machine Learning” (González, 2019, p.8).

A. NUMPY

(González, 2019), nos dice que Numpy o Numerical Python es una de las librerías que posee Python, la cual es usada en la informática científica, ofrece la capacidad de vectorizar las diferentes operaciones que la matemática tiene y las convierte en matrices, para, de esta manera, ayudar al rendimiento y, por ende, mejorar la ejecución del sistema.

Las matrices creadas son multidimensionales, pero también crea rutinas las cuales ayudan a los especialistas en análisis de datos en crear las funciones estadísticas que necesiten, pero sin la necesidad de crear código innecesario.

B. PANDAS

(González, 2019), Pandas es otra librería que ofrece Python, cuya especialidad es el trabajo con información etiquetada, es decir datos estructurados; Pandas es eficaz en el tratamiento total de este tipo de datos, nos permite manipular, agregar, eliminar y visualizar datos de manera rápida y sencilla, su uso se extiende para todas las áreas posibles, como finanzas, ingeniería y otro que requiera el análisis estadístico.

C. MATPLOTLIB

(González, 2019), otra librería que ofrece Python para el estudio en ciencia de datos es Matplotlib, usada para la creación de gráficos y diagramas, el inconveniente es que se trata de una librería de bajo nivel, es decir que para su generación gráfica se requiere escribir más comandos, pero lo recompensa con la flexibilidad que posee, pues otras librerías gráficas de más alto nivel son estancadas al momento de “improvisar” gráficos.

D. SCIKIT-LEARN

Esta librería es popular, pues posee una inmensa cantidad de propiedades que pueden ser usadas en el data mining y en la ciencia de datos, por tanto analizar datos con esta librería es lo ideal; tiene como base de construcción las otras librerías populares como son Scipy, Numpy y Matplotlib, por tanto el usarlo no ofrece una experiencia distinta o nueva para el analista o programador, como información técnica de la librería se debe indicar que su código es de calidad, posee documentación, es fácil de usar además que ofrece un rendimiento alto, por último, scikit-learn ya es un estándar en Python al momento de implementar machine learning. (González, 2019).

E. OVERLIFTTING

Para (Raschka y Mirjalili, 2017), “es un problema común en el aprendizaje automático, donde un modelo funciona bien en datos de entrenamiento, pero no se generaliza bien a datos no vistos. Si un modelo sufre de sobreajuste, también decimos que el modelo tiene una alta varianza, que puede ser causado por tener demasiados parámetros que conducen a un modelo que es demasiado complejo dados los datos subyacentes, del mismo modo, nuestro modelo también puede sufrir de falta de equipamiento (alto sesgo), lo que significa que nuestro modelo no es lo suficientemente complejo como para capturar el patrón en los datos de entrenamiento bien y por lo tanto también sufre de bajo rendimiento en datos no vistos”.

2.2.6 LENGUAJE DE PROGRAMACIÓN ORIENTADA A OBJETOS

(Weitzenfeld, s.f.), afirma que estos lenguajes tienen diferencias entre sí, tanto en su estructura como en el flujo interno, esto explica que, a pesar de ser creados y diseñados para un mismo objetivo de programación, siempre se muestran detalles, características o funcionalidades que algunos lenguajes ofrecen y otros no

“Es un conjunto de símbolos, palabras y reglas que permiten implementar un algoritmo en una computadora. Dicha implementación se conoce como programa y se escribe como una secuencia de frases del lenguaje de programación” (Osorio, s.f.).

Las ventajas que ofrecen estos lenguajes de programación son: la cualidad de su código por ser reutilizable, son menos costosos, ya no son requeridas hacer muchas pruebas y son rápidas en su implementación, los programadores deben realizar estrategias de acoplado para emplear varios lenguajes en el desarrollo de un solo sistema. (Stair y Reynolds, 1999).

(Craing, 2002), define a estos lenguajes como una agrupación de propiedades y que la medida de satisfacción en la que el mismo lenguaje tiene sobre sus propiedades es la que le da el grado de aceptación como lenguaje orientado a objetos.

2.2.7 POBLACIÓN

(Chávez, 2007), dice que la población "es el universo de estudio de la investigación, sobre el cual se pretende generalizar los resultados, constituida por características o estratos que le permiten distinguir los sujetos unos de otros".

Para (Tamayo y Tamayo, 1997), globaliza al conjunto donde sucede el fenómeno, hecho o evento a estudiar cómo la población, donde todos los componentes de dicho conjunto tienen por lo menos una característica en común, y son estos, los que generan la información inicial para la investigación.

2.2.8 MUESTRA

Según (Tamayo y Tamayo, 1997), la muestra, son los integrantes de la población, la cual es separada, por algún criterio en especial, para estudiar estadísticamente en ellos la influencia que tiene un problema o un fenómeno.

A este respecto, Bavaresco (2006), refiere que "cuando se hace difícil el estudio de toda la población, es necesario extraer una muestra, la cual no es más que un subconjunto de la población con la que se va a trabajar".

2.2.9 MUESTREO POR CONVENIENCIA

A este tipo de muestreo también se le conoce como sesgado, es sencilla su selección, pues depende del criterio del investigador para crear una muestra así, pero exige responsabilidad y madurez al momento de crear dicho criterio de separación. (Tamayo y Tamayo, 1997).

Para (Bavaresco, 2006), la selección de los integrantes que van a pertenecer a este tipo de muestra, es porque su reclutamiento es sencillo y están a disposición inmediata; además que el investigador no es estricto con las características que hacen que dichos sujetos sean buenos representantes de la población.

CAPÍTULO III
MATERIALES Y MÉTODOS

3.1 TIPO Y NIVEL DE LA INVESTIGACIÓN

3.1.1 TIPO DE INVESTIGACIÓN

Según (Salinas, 2010), define la investigación observacional como: “Aquella investigación que se basa en la observación de los fenómenos, características, situaciones, variaciones, etc. del asunto que se quiere investigar. Solo se observa los fenómenos en su ambiente natural para analizarlos, sin manipular, cambiar o variar nada”.

Según (Supo, 2014), la investigación observacional “es cuando no existe intervención del investigador los datos reflejan la evolución natural de los eventos ajena a la voluntad del investigador”.

De acuerdo a las definiciones anteriores, esta investigación es del tipo observacional, ya que no hay ninguna intención de intervenir en el flujo de los datos, bajo ninguna manera y en ninguna de las etapas; simplemente se observará, tratará y analizará como los datos obtenidos de la Escuela Profesional de Ingeniería de Sistemas tienen una relevancia al momento de realizar una predicción del rendimiento académico.

Según (Supo, 2014), define que “los estudios retrospectivos utilizan datos que se obtienen de registros preexistentes, datos que provienen de mediciones en donde el investigador no tuvo participación alguna. A este tipo de información se le suele llamar datos secundarios”

Para (Supo J., 2015), “los estudios retrospectivos los datos se recogen de registros en donde el investigador no tuvo participación, aquí el investigador no realizó las

mediciones, los datos se obtuvieron anteriormente con fines ajenos al proyecto de investigación, por lo que se les denomina datos secundarios".

De acuerdo a las definiciones anteriores, esta investigación es también del tipo retrospectivo, ya que se tomará como base de información los datos de los estudiantes que ya fueron recolectados a lo largo de los años por la Escuela de Formación profesional de ingeniería de Sistemas, simplemente limitaremos el rango de dicha información y la tomaremos desde los periodos 2016-II hasta el 2019-I.

Según (Supo, 2014), "la investigación transversal es cuando todas las variables son medidas en una sola ocasión; por ello de realizar comparaciones, se trata de muestras independientes".

Según (Hernández, Fernández, & Baptista, 2006), "el tipo de investigación transaccional o transversal recolecta datos en un solo momento, en un tiempo único. Su propósito es describir variables, y analizar su incidencia e interrelación en un momento dado".

De acuerdo a las definiciones anteriores, esta investigación es también del tipo transaccional, ya que los datos fueron recolectados en un determinado momento, que, es al inicio de cada semestre donde se definen la cantidad de alumnos, los cursos y los docentes; y es terminada al final del semestre cuando se completa con las notas obtenidas.

3.1.2 NIVEL DE INVESTIGACIÓN

Según Supo (2013), en el nivel de la investigación predictiva:
"Los estudios que se desarrollan en este nivel investigativo tienen como finalidad calcular la probabilidad de ocurrencia de un suceso, como, por ejemplo, un problema o una enfermedad; también tiene la finalidad de calcular el tiempo medio en que ocurriría" (p.49).

Monje (2011) indica que el estudio predictivo:
"Busca anticipar acontecimientos futuros mediante la predicción, el investigador puede definir acciones prácticas que se orienten a hacer que tales eventos sucedan o no, obteniéndose de esta forma la capacidad de solucionar los problemas que se producen en su objeto de conocimiento. La predicción equivale a aquello que el investigador espera que ocurra a partir de las acciones que defina soluciones" (p.96).

De acuerdo a las definiciones anteriores, esta investigación es de nivel predictivo, ya que contamos con información pasada (histórica) que nos ayuda a conocer la realidad del rendimiento académico, y a partir de ellos, podremos responder a las preguntas ¿Qué sucederá en el futuro? y ¿En qué condiciones o circunstancias ocurrirá?, dado que el futuro para nuestra investigación es si un alumno aprobará o desaprobará y que las condiciones o circunstancias en las ocurrirá dicha predicción dependerán de nuestras variables predictoras.

3.2 DISEÑO DE LA INVESTIGACIÓN

Según (Supo, 2014), define “el diseño de la investigación es una estrategia metodológica y estadística puntual para el desarrollo de un trabajo de investigación”.

Según Hernández et al. (2014), el diseño no experimental es “aquella que se realiza sin manipular deliberadamente variables y no se hace variar en forma intencional las variables independientes para ver su efecto sobre otras variables. Solo se observan los fenómenos tal como se dan en su contexto natural para analizarlos”.

De acuerdo a las definiciones anteriores, esta investigación se ubica como diseño no experimental, pues no se harán variaciones en las variables sino simplemente serán sometidas al tratamiento respectivo que ofrece el machine learning para obtener un resultado que ya se tiene registrado anteriormente junto con la base de datos inicial.

3.3 POBLACIÓN Y MUESTRA

POBLACIÓN

La población de estudio analizada está compuesta por los registros de información de los alumnos de la Escuela Profesional de Ingeniería de Sistemas desde los semestres académicos 2016-II hasta el 2019-I.

MUESTRA

La muestra estará conformada por los mismos elementos de la población.

3.4 VARIABLES E INDICADORES

3.4.1 DEFINICIÓN CONCEPTUAL DE LAS VARIABLES

VARIABLE DE INTERÉS

Modelo de Machine Learning: Es un prototipo que genera una salida deseada de información cuando el algoritmo de Machine Learning que tiene implementado es alimentado con volúmenes de datos.

VARIABLES DESCRIPTIVAS

Técnica predictiva: Es un método avanzado de análisis el cual usa datos históricos o nuevos con la finalidad de generar una tendencia o una predicción de algún acontecimiento.

Algoritmo predictivo: Es la agrupación de reglas, pasos y procedimientos establecidos, organizados y finitos que están destinadas a la realización de una tarea predictiva, es decir que siguiendo los pasos sucesivos se llegará a la predicción deseada sin generar dudas.

Métrica de validez: Es el proceso de comprobación de la eficiencia del modelo, una vez terminado el desarrollo de construcción del modelo, se debe someter a la comprobación del correcto funcionamiento del mismo.

3.4.2 DEFINICIÓN OPERACIONAL DE LAS VARIABLES

VARIABLE DE INTERÉS

X: Modelo de Machine Learning

VARIABLES DESCRIPTIVAS

X1: Técnica predictiva

X2: Algoritmo predictivo.

X3: Métrica de validez.

OPERACIONALIZACIÓN DE LAS VARIABLES

Se muestra en el Anexo A

3.5 TÉCNICAS E INSTRUMENTOS PARA RECOLECTAR INFORMACIÓN

3.5.1 TÉCNICAS PARA RECOLECTAR INFORMACIÓN

La técnica para recolectar información fue:

Análisis documental

Esta técnica de recolección fue considerada para recolectar la información necesaria tanto de los estudiantes como de los modelos de machine learning, que nos permitirá predecir el rendimiento académico.

3.5.2 INSTRUMENTOS PARA RECOLECTAR INFORMACIÓN

El instrumento para el recojo de información fue:

Ficha de registro

Con este instrumento se obtuvo los registros de los estudiantes, colegio de egreso, cursos matriculados, semestre académico, modalidad de matrícula docente y nota final; información necesaria para la construcción del modelo predictivo, dicho instrumento se muestra en el anexo B.

Ficha bibliográfica

Con este instrumento se obtuvo la información de las técnicas predictivas, de los algoritmos predictivos y de las métricas de validación, necesarias para la construcción del modelo predictivo, dicho instrumento se muestra en el anexo C.

3.5.3 HERRAMIENTAS PARA EL TRATAMIENTO DE DATOS E INFORMACIÓN

Considerando los detalles que se requieren para el desarrollo de un modelo de machine learning como son la programación, el tratamiento de datos, su visualización y métricas, se necesitarán las siguientes herramientas tecnológicas:

Tabla 2
Herramientas tecnológicas.

NOMBRE	FABRICANTE	SERVICIO
Windows 10	Microsoft Corporation	Software insignia de la compañía Microsoft, ofrece una interfaz amigable y fácil de usar, además el manejo de otras aplicaciones o programas no impide el correcto funcionamiento, más bien cumple la función de ser una buena base sobre la cual trabajar.
Anaconda Distribution	Anaconda, Inc.	Es una suite open source muy completa diseñada para ser usada en la ciencia de datos, cumple la función de ser el gestor de entorno y de paquetes. En su instalación se crean 4 grupos o sectores que son: Anaconda Navigator, Anaconda Project, las librerías de Data Science y Conda.

Jupyter Notebook	Proyecto Jupyter	Es una aplicación open source, muy usada en la creación de documentos que además tengan código, se resumen como documentos web empleados para el análisis de datos; permiten realizar depuración de datos, modelización estadística, creación y entrenamiento de modelos de machine learning, finalmente permite visualizar los datos.
Python 3.7	Python Software Foundation	Es el lenguaje de programación multiparadigna que ofrece muchas ventajas como: la curva de aprendizaje es más fácil, su lenguaje es dinámico, la programación se organiza por celdas, las aplicaciones creadas no solo se centran al desarrollo web o de escritorio, sino que también se implementa en el área de Data Science, machine learning y demás; las características más relevantes de Python son su sencillez, legibilidad y exactitud en su sintaxis.
Scikit-learn	Google Summer of Code	Es una biblioteca de aprendizaje automático de software libre para el lenguaje de programación Python, cuenta con varios algoritmos de clasificación, regresión y agrupación.
Pandas	Wes McKinney	Es la herramienta básica para el data science en Python, siendo una de sus librerías, empelada para el tratamiento,

		manipulación y posterior análisis de datos, lo realiza creando dataframes que cuentan con indexación integrada pues es más fácil y eficaz trabajar cuando los datos están organizados de esta manera.
NumPy	Python Software Foundation	Es un paquete o librería de Python, permite crear matrices para acceder más rápido a los datos, dicha función de vectorización acelera en gran medida las funciones de aprendizaje que pueden tener los modelos de machine learning pues las matrices multidimensionales garantizan cálculos eficientes.
Matplotlib	Python Software Foundation	Es la librería de bajo nivel usado para la visualización gráfica de datos u otra característica que se desee graficar, es necesaria escribir más código si se desea ver gráficos mejor elaborados, pero eso también le da más versatilidad a matplotlib.

Nota. Fuente: Elaboración propia.

3.5.4 TÉCNICA PARA APLICAR LA METODOLOGÍA FUNDAMENTAL PARA LA CIENCIA DE DATOS

De acuerdo a los conceptos presentados en el capítulo II, sección 2.2.3, correspondiente al marco teórico; donde se explica la metodología general de ciencia de datos detallando todas las fases que tiene dicha metodología, a continuación, se presenta un resumen:

Tabla 3
Etapas de la metodología fundamental para ciencia de datos.

ETAPAS	DESCRIPCIÓN	RESULTADO
01. Comprensión del negocio.	Se trata de profundizar y conocer más acerca de los intereses y del problema que se presenta y del cual se tratará de resolver.	Información del negocio.
02. Objetivos del negocio.	Se enlista los objetivos que desea alcanzar el negocio.	Objetivos principales.
03. Enfoque analítico.	Se plantea la variable o el dato a predecir y las posibles técnicas a empelar.	Propuesta de posibles enfoques analíticos.
04. Requisitos de los datos.	Se enlista los requisitos que deben poseer los datos para que sean admitidos como una fuente informativa.	Lista de requisitos.
05. Recopilación de los datos.	Recojo de los datos históricos, los cuales la empresa ha recolectado con anterioridad.	Registros de datos.
06. Comprensión de los datos.	Se observa, analiza e interpreta los datos mediante gráficos, haciendo los cruces necesarios entre los datos con la variable predictora.	Gráficos estadísticos.
07. Preparación de los datos.	Se identifican los datos necesarios para el modelo, se eliminan los datos innecesarios, finalmente, se cuantifican las variables.	Dataframe con los datos adecuados para el uso en el modelo de machine learning.
08. Modelado.	Se identifican las variables a predecir y las predictoras, se separan en grupos,	Modelos predictivos separados por el tipo de algoritmo empelado.

	finalmente, se aplican los algoritmos seleccionados de aprendizaje automático.	
09. Evaluación.	Aplicación de las métricas seleccionadas para la validez del modelo anteriormente creado.	Modelo predictivo validado con la mayor probabilidad de predicción.
10. Implementación.	Se exporta el algoritmo entrenado en el modelo, para llevarlo a otra estación para continuar su entrenamiento, mejorando la precisión.	Archivos con extensión pkl conteniendo el algoritmo entrenado.
11. Retroalimentación.	Se enlista las maneras de cómo continuar entrenando el algoritmo y /o el modelo.	Lista de recomendaciones.

Nota. Fuente: Elaboración propia.

CAPÍTULO IV
RESULTADOS Y DISCUSIÓN

4.1 RESULTADOS DE LA METODOLOGÍA FUNDAMENTAL PARA LA CIENCIA DE DATOS

Como ya se mencionó antes, las etapas de la metodología fundamental de ciencia de datos ya fue explicada en el capítulo II y también se presentó una tabla que resumen dichas etapas en el capítulo III, por tanto a continuación se mostrarán los resultados etapa por etapa, empezando con la comprensión del negocio, y es a partir de ahí que se debe encaminar el desarrollo hacia los objetivos propuestos, siempre teniendo en cuenta las limitaciones con la finalidad de garantizar el éxito del proyecto:

4.1.1 COMPRENSIÓN DEL NEGOCIO

La Universidad Nacional San Cristóbal de Huamanga (UNSCH) tiene 28 Escuelas Profesionales, siendo una de ellas la de Ingeniería de Sistemas; sus egresados pueden realizar diferentes funciones como, por ejemplo, se encargan del diseño, programación, implementación y el mantenimiento de los sistemas de información de una organización.

Esta carrera requiere de muchos conocimientos para implementar sus diseños en la vida real, no solamente de saberes tecnológicos sino también conocimientos matemáticos que ayudan en la realización de una aplicación tecnológica. El Ingeniero de Sistemas posee la capacidad de recolectar, almacenar, recuperar, procesar y comunicar datos e información, para así lograr que las operaciones de una organización sean gestionadas eficientemente.

Su creación no fue planeada desde un inicio, sino que en el año 2025 se cambió la denominación de la anterior Escuela de Formación Profesional de Ingeniería Informática por la de Escuela de Formación Profesional de Ingeniería de Sistemas bajo la resolución N.º 001-2005-UNSCH-AU, aprobada en asamblea universitaria.

La Escuela Profesional de Ingeniería de Sistemas recibe mucha información por cada alumno, en cada proceso de admisión como también en cada semestre académico; dicha información es registrada y almacenada mediante el Sistema de Matrícula (SIMA), llegando a concentrar una gran cantidad de datos tanto de los alumnos actuales como de los egresados.

PERFIL PROFESIONAL

- Diseña modelos y desarrolla sistemas con capacidad de retroalimentación (feedback).
- Crea softwares a medida dependiendo de los requerimientos del negocio, implementando las arquitecturas de software más eficientes para tal propósito.
- Se hace cargo del área de soporte de cualquier empresa, dando soluciones inmediatas, eficaces y duraderas.
- Su contribución para el área de toma de decisiones es vital, facilitando acceso a datos, mostrando los reportes requeridos, creando soluciones informáticas para el manejo de datos.
- Puede mejorar la integración de negocios creando sistemas multiempresas, respetando los planes estratégicos y mejorando la calidad en tecnologías de información.
- Desempeñarse en la investigación científica.
- Ayudar al desarrollo social implementando sistemas que proporcionen mejoras en la vida cotidiana.
- Puede ejercer la docencia superior, como también ayudar en la mejora del conocimiento profesional.

MISIÓN

Desarrollar, emplear y difundir los conocimientos propios de la carrera universitaria, las nuevas tecnologías y su tendencia, así como también la cultura como labor social, instruyendo a los estudiantes para forjar profesionales muy competitivos, que con sus valores, conocimientos y aptitudes ayuden en el progreso de la sociedad.

VISIÓN

Ser una líder como escuela profesional, cuyas investigaciones sean de utilidad tanto en la ciencia como en la tecnología, también que sus profesionales egresados sean reconocidos cumpliendo una función vital en la sociedad, oteando los siguientes 20 años.

4.1.2 OBJETIVO DEL NEGOCIO

“Formar profesionales con fundamentos en la teoría general de sistemas, en la tecnología informática y en la gestión de negocios; capaces de desempeñarse con criterio científico-técnico y humanístico en funciones de análisis, diseño, construcción y ejecución, evaluación y control, mantenimiento, dirección e integración de proyectos interdisciplinarios, con el propósito de liderar con una visión de futuro el desarrollo de la región y del país”. (UNSCH, 2021)

4.1.3 ENFOQUE ANALÍTICO

De acuerdo al objetivo planteado en el presente proyecto, se espera predecir el rendimiento académico, es decir si un alumno aprobó o desaprobó, por tanto, el resultado obtenido tendrá sólo dos valores posibles, es decir una variable dicotómica y por ende la probabilidad es binomial.

La técnica que se va usar depende de los objetivos y del tipo de datos que se poseen, para esta investigación los datos obtenidos son de tipo discreto, además que la finalidad es predecir una variable dicotómica, por tanto, los algoritmos de agrupación son los más recomendados, lo que define que la técnica usada será la de aprendizaje supervisado.

4.1.4 REQUISITO DE DATOS

En la ciencia de datos se deben usar datos que cumplan ciertos requisitos, esto se entiende pues al tratarse de técnicas y métodos estadísticos tanto en el procesamiento, modelamiento y visualización de datos se requiere el cumplimiento de restricciones, normas y estándares, los cuales se listan a continuación:

- No deben existir filas, columnas o campos vacíos, esto para ayudar al modelo predictivo; se debe eliminar toda la columna si hay muchos alumnos que no tienen ese dato, o la fila de un registro si el alumno tiene muchos campos vacíos; pues al momento de procesarlos le generará confusión al modelo.
- No debe existir datos duplicados; es decir que todos los registros deben ser únicos y que tengan una estructura definida,
- Los datos deben ser cuantitativos, específicamente variables Dummy, pues los modelos de machine learning no pueden procesar variables cualitativas, por tanto, es necesario transformarlas.
- Todos los datos deben estar agrupados en un solo registro, de formato abierto y sencillo; la ciencia de datos ya tiene su estándar pues procesa los archivos de formato CSV (Valores Separados por Comas).

4.1.5 RECOPILACIÓN DE DATOS

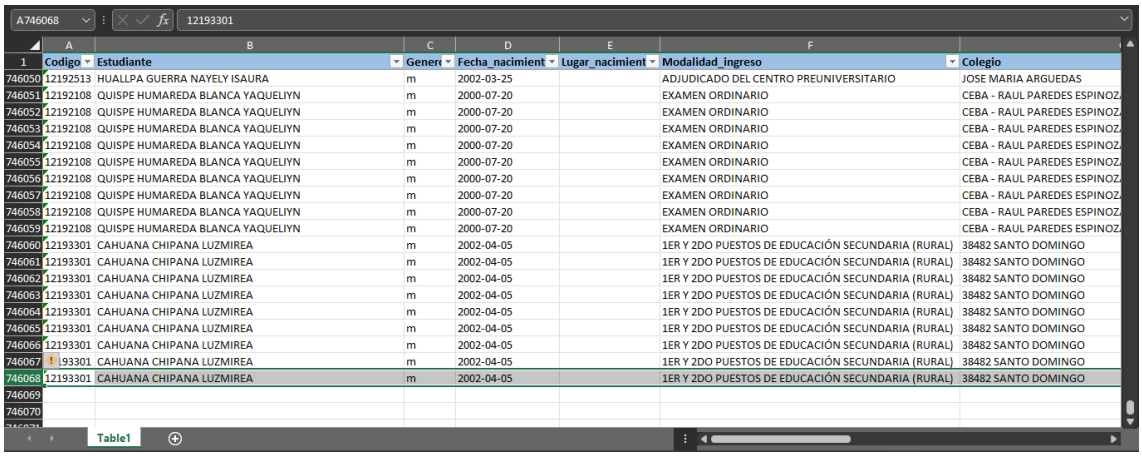
En esta etapa de recopilación de datos, se tuvo que pedirla a la Oficina General de Informática y Sistemas, por tanto, se presentó una solicitud (ver anexo D), pidiendo la información de los estudiantes para esta investigación; a continuación, se detalla la información obtenida:

Tabla 4
Datos obtenidos de la Oficina General de Informática y Sistemas.

DATOS	DESCRIPCIÓN	TIPO	MEDICIÓN
Dato de interés	Promedio final del curso	Cuantitativa	Ordinal
Datos personales	Código del estudiante	Cuantitativa	Ordinal
	Nombres y apellidos	Cualitativa	Nominal
	Género	Cualitativa	Nominal
	Fecha de nacimiento	Cualitativa	Nominal
	Modalidad de ingreso	Cualitativa	Nominal
Datos académicos	Colegio de egreso	Cualitativa	Nominal
	Ubicación del colegio	Cualitativa	Nominal
	Semestre académico	Cualitativa	Nominal
	Código del curso	Cualitativa	Nominal
	Nombre del curso	Cualitativa	Nominal
	Plan de estudios	Cualitativa	Nominal
	Créditos del curso	Cuantitativa	Ordinal
	Modalidad del curso	Cualitativa	Nominal
	Docente del curso	Cualitativa	Nominal

Nota. Fuente: Elaboración propia.

El documento obtenido fue un archivo en formato .xlsx con un total de 746 067 registros, pero conteniendo información de los estudiantes de todas las escuelas de la universidad, eso incluye las 2 sedes (Huamanga y Pichari).

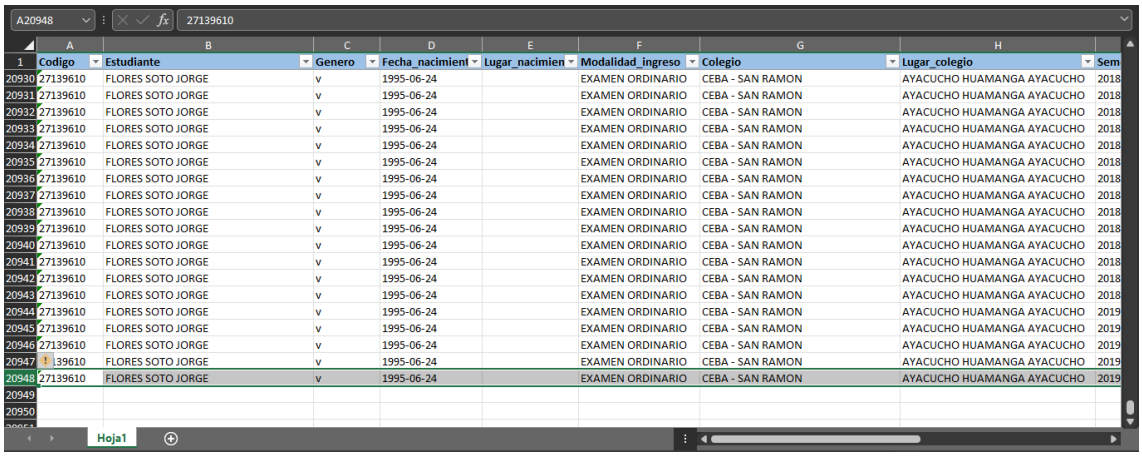


	A	B	C	D	E	F	
1	Codigo	Estudiante	Generc	Fecha_nacimient	Lugar_nacimient	Modalidad_ingreso	Colegio
746050	12192513	HUALLPA GUERRA NAYELY ISAURA	m	2002-03-25		ADJUDICADO DEL CENTRO PREUNIVERSITARIO	JOSE MARIA ARGUEDAS
746051	12192108	QUISPE HUMAREDA BLANCA YAQUELIYN	m	2000-07-20		EXAMEN ORDINARIO	CEBA - RAUL PAREDES ESPINOZ
746052	12192108	QUISPE HUMAREDA BLANCA YAQUELIYN	m	2000-07-20		EXAMEN ORDINARIO	CEBA - RAUL PAREDES ESPINOZ
746053	12192108	QUISPE HUMAREDA BLANCA YAQUELIYN	m	2000-07-20		EXAMEN ORDINARIO	CEBA - RAUL PAREDES ESPINOZ
746054	12192108	QUISPE HUMAREDA BLANCA YAQUELIYN	m	2000-07-20		EXAMEN ORDINARIO	CEBA - RAUL PAREDES ESPINOZ
746055	12192108	QUISPE HUMAREDA BLANCA YAQUELIYN	m	2000-07-20		EXAMEN ORDINARIO	CEBA - RAUL PAREDES ESPINOZ
746056	12192108	QUISPE HUMAREDA BLANCA YAQUELIYN	m	2000-07-20		EXAMEN ORDINARIO	CEBA - RAUL PAREDES ESPINOZ
746057	12192108	QUISPE HUMAREDA BLANCA YAQUELIYN	m	2000-07-20		EXAMEN ORDINARIO	CEBA - RAUL PAREDES ESPINOZ
746058	12192108	QUISPE HUMAREDA BLANCA YAQUELIYN	m	2000-07-20		EXAMEN ORDINARIO	CEBA - RAUL PAREDES ESPINOZ
746059	12192108	QUISPE HUMAREDA BLANCA YAQUELIYN	m	2000-07-20		EXAMEN ORDINARIO	CEBA - RAUL PAREDES ESPINOZ
746060	12193301	CAHUANA CHIPANA LUZMIREA	m	2002-04-05		1ER Y 2DO PUESTOS DE EDUCACIÓN SECUNDARIA (RURAL)	38482 SANTO DOMINGO
746061	12193301	CAHUANA CHIPANA LUZMIREA	m	2002-04-05		1ER Y 2DO PUESTOS DE EDUCACIÓN SECUNDARIA (RURAL)	38482 SANTO DOMINGO
746062	12193301	CAHUANA CHIPANA LUZMIREA	m	2002-04-05		1ER Y 2DO PUESTOS DE EDUCACIÓN SECUNDARIA (RURAL)	38482 SANTO DOMINGO
746063	12193301	CAHUANA CHIPANA LUZMIREA	m	2002-04-05		1ER Y 2DO PUESTOS DE EDUCACIÓN SECUNDARIA (RURAL)	38482 SANTO DOMINGO
746064	12193301	CAHUANA CHIPANA LUZMIREA	m	2002-04-05		1ER Y 2DO PUESTOS DE EDUCACIÓN SECUNDARIA (RURAL)	38482 SANTO DOMINGO
746065	12193301	CAHUANA CHIPANA LUZMIREA	m	2002-04-05		1ER Y 2DO PUESTOS DE EDUCACIÓN SECUNDARIA (RURAL)	38482 SANTO DOMINGO
746066	12193301	CAHUANA CHIPANA LUZMIREA	m	2002-04-05		1ER Y 2DO PUESTOS DE EDUCACIÓN SECUNDARIA (RURAL)	38482 SANTO DOMINGO
746067	193301	CAHUANA CHIPANA LUZMIREA	m	2002-04-05		1ER Y 2DO PUESTOS DE EDUCACIÓN SECUNDARIA (RURAL)	38482 SANTO DOMINGO
746068	12193301	CAHUANA CHIPANA LUZMIREA	m	2002-04-05		1ER Y 2DO PUESTOS DE EDUCACIÓN SECUNDARIA (RURAL)	38482 SANTO DOMINGO
746069							
746070							

Table1

Figura 16: Registro inicial de alumnos.

La información obtenida es de los semestres académicos 2016-II hasta el 2019-I, pero para esta investigación la población solamente son los estudiantes de la escuela profesional de ingeniería de sistema, por tanto, se procedió a realizar una primera filtración de datos, obteniendo 20 947 registros.



	A	B	C	D	E	F	G	H	
1	Codigo	Estudiante	Genero	Fecha_nacimient	Lugar_nacimien	Modalidad_ingreso	Colegio	Lugar_colegio	Sem
20930	27139610	FLORES SOTO JORGE	v	1995-06-24		EXAMEN ORDINARIO	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2018
20931	27139610	FLORES SOTO JORGE	v	1995-06-24		EXAMEN ORDINARIO	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2018
20932	27139610	FLORES SOTO JORGE	v	1995-06-24		EXAMEN ORDINARIO	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2018
20933	27139610	FLORES SOTO JORGE	v	1995-06-24		EXAMEN ORDINARIO	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2018
20934	27139610	FLORES SOTO JORGE	v	1995-06-24		EXAMEN ORDINARIO	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2018
20935	27139610	FLORES SOTO JORGE	v	1995-06-24		EXAMEN ORDINARIO	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2018
20936	27139610	FLORES SOTO JORGE	v	1995-06-24		EXAMEN ORDINARIO	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2018
20937	27139610	FLORES SOTO JORGE	v	1995-06-24		EXAMEN ORDINARIO	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2018
20938	27139610	FLORES SOTO JORGE	v	1995-06-24		EXAMEN ORDINARIO	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2018
20939	27139610	FLORES SOTO JORGE	v	1995-06-24		EXAMEN ORDINARIO	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2018
20940	27139610	FLORES SOTO JORGE	v	1995-06-24		EXAMEN ORDINARIO	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2018
20941	27139610	FLORES SOTO JORGE	v	1995-06-24		EXAMEN ORDINARIO	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2018
20942	27139610	FLORES SOTO JORGE	v	1995-06-24		EXAMEN ORDINARIO	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2018
20943	27139610	FLORES SOTO JORGE	v	1995-06-24		EXAMEN ORDINARIO	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2018
20944	27139610	FLORES SOTO JORGE	v	1995-06-24		EXAMEN ORDINARIO	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2019
20945	27139610	FLORES SOTO JORGE	v	1995-06-24		EXAMEN ORDINARIO	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2019
20946	27139610	FLORES SOTO JORGE	v	1995-06-24		EXAMEN ORDINARIO	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2019
20947	139610	FLORES SOTO JORGE	v	1995-06-24		EXAMEN ORDINARIO	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2019
20948	27139610	FLORES SOTO JORGE	v	1995-06-24		EXAMEN ORDINARIO	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2019
20949									
20950									

Hoja1

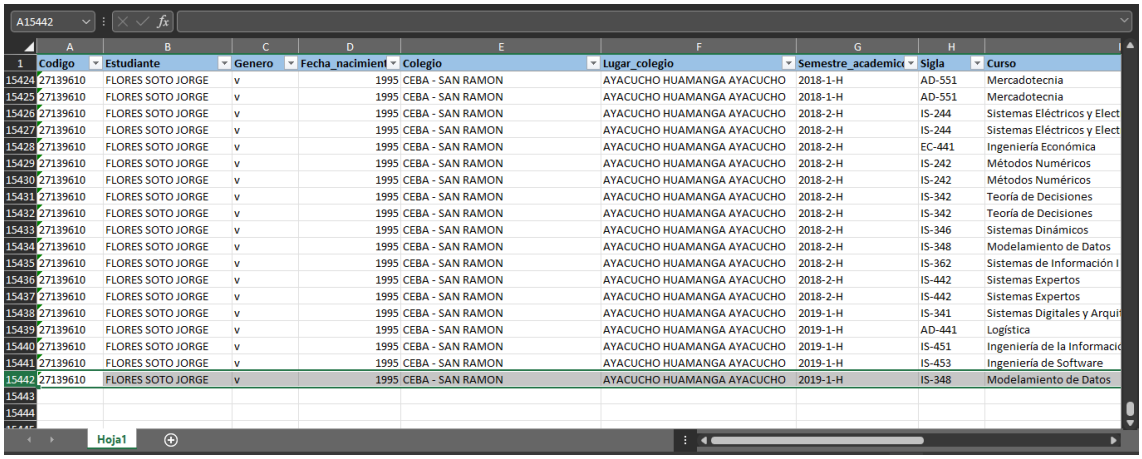
Figura 17: Registro filtrado con alumnos de la EPIS.

Observando los datos se identificó dos grandes problemas: la falta de información y los registros repetidos; en cuanto al primer problema se realizó lo siguiente:

- Se eliminó la columna “Lugar_nacimiento” pues está completamente vacía
- En las columnas “Colegio” y “Lugar_colegio” los espacios vacíos fueron llenados con el dato “No especificado”, para poder analizarlo después.
- En la columna “Género” los espacios vacíos fueron llenados con los valores respectivos de varón o mujer, pues eran muy pocos registros.
- Se encontraron 26 registros vacíos en la columna “Fecha_nacimiento”; se procedió a eliminar todas las filas a las que pertenecían estos campos, pues no había manera de usar esos registros.

En cuanto al segundo problema, se realizó lo siguiente:

- Se identificó a la columna “Tipo_matrícula” como innecesaria y problemática pues analizando sus valores se observa una incoherencia con la columna “Modalidad_curso”.
- Con la herramienta Filtro Avanzado, se procedió a filtrar toda la data para que sólo contenga registros únicos, obteniendo ahora un archivo con 15 441 registros.



	Codigo	Estudiante	Genero	Fecha_nacimient	Colegio	Lugar_colegio	Semestre_academic	Sigla	Curso
15424	27139610	FLORES SOTO JORGE	v	1995	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2018-1-H	AD-551	Mercadotecnia
15425	27139610	FLORES SOTO JORGE	v	1995	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2018-1-H	AD-551	Mercadotecnia
15426	27139610	FLORES SOTO JORGE	v	1995	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2018-2-H	IS-244	Sistemas Eléctricos y Elect
15427	27139610	FLORES SOTO JORGE	v	1995	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2018-2-H	IS-244	Sistemas Eléctricos y Elect
15428	27139610	FLORES SOTO JORGE	v	1995	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2018-2-H	EC-441	Ingeniería Económica
15429	27139610	FLORES SOTO JORGE	v	1995	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2018-2-H	IS-242	Métodos Numéricos
15430	27139610	FLORES SOTO JORGE	v	1995	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2018-2-H	IS-242	Métodos Numéricos
15431	27139610	FLORES SOTO JORGE	v	1995	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2018-2-H	IS-342	Teoría de Decisiones
15432	27139610	FLORES SOTO JORGE	v	1995	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2018-2-H	IS-342	Teoría de Decisiones
15433	27139610	FLORES SOTO JORGE	v	1995	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2018-2-H	IS-346	Sistemas Dinámicos
15434	27139610	FLORES SOTO JORGE	v	1995	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2018-2-H	IS-348	Modelamiento de Datos
15435	27139610	FLORES SOTO JORGE	v	1995	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2018-2-H	IS-362	Sistemas de Información I
15436	27139610	FLORES SOTO JORGE	v	1995	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2018-2-H	IS-442	Sistemas Expertos
15437	27139610	FLORES SOTO JORGE	v	1995	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2018-2-H	IS-442	Sistemas Expertos
15438	27139610	FLORES SOTO JORGE	v	1995	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2019-1-H	IS-341	Sistemas Digitales y Arquit
15439	27139610	FLORES SOTO JORGE	v	1995	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2019-1-H	AD-441	Logística
15440	27139610	FLORES SOTO JORGE	v	1995	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2019-1-H	IS-451	Ingeniería de la Informaci
15441	27139610	FLORES SOTO JORGE	v	1995	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2019-1-H	IS-453	Ingeniería de Software
15442	27139610	FLORES SOTO JORGE	v	1995	CEBA - SAN RAMON	AYACUCHO HUAMANGA AYACUCHO	2019-1-H	IS-348	Modelamiento de Datos
15443									
15444									

Figura 18: Registro final de alumnos de la EPIS.

Como el formato de una hoja de cálculo .xlsx, no es recomendable para trabajar en ciencia de datos se procede a guardar el archivo en formato .csv, aprovechando que MS Excel nos brinda esa facilidad, quedando finalmente el archivo “SistemasFinal.csv” para continuar con el trabajo en las siguientes etapas.

4.1.6 COMPRENSIÓN DE DATOS

Con la información anterior, ahora se procederá a usar las herramientas tecnológicas de ciencia de datos que nos ofrece Python, y como ya se indicó, se usará del Anaconda Navigator el IDE de análisis de datos Jupyter Notebook; pues en esta etapa se deben comprender mejor los datos con los que vamos a trabajar y la mejor manera es analizarlos mediante gráficos.

Lo primero es la importación de las librerías adecuadas para el trabajo, después se debe cargar el archivo “SistemasFinal.csv” en un dataframe para poder trabajar adecuadamente en la elaboración de las estadísticas; el siguiente bloque de código nos permite realizar dichas operaciones:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

```
data = pd.read_csv('SistemasFinal.csv', encoding='latin1')
```

```
data.head()
```

	Codigo	Estudiante	Genero	Fecha_nacimiento	Colegio	Lugar_colegio	Semestre_academico	Sigla	Curso	Creditos	Modalidad_curso	Promedio_c
0	27021206	APERRIGUE CABRERA RENAN CARLOS	v	1984	SAN AGUSTIN	AYACUCHO HUAMANGA AYACUCHO	2016-2-H	IS-442	Sistemas Expertos	4	REGULAR	
1	27021206	APERRIGUE CABRERA RENAN CARLOS	v	1984	SAN AGUSTIN	AYACUCHO HUAMANGA AYACUCHO	2016-2-H	AD-442	Gestión Financiera	3	REGULAR	
2	27021206	APERRIGUE CABRERA RENAN CARLOS	v	1984	SAN AGUSTIN	AYACUCHO HUAMANGA AYACUCHO	2016-2-H	IS-446	Sistemas de Información Gerencial	3	REGULAR	
3	27021206	APERRIGUE CABRERA RENAN CARLOS	v	1984	SAN AGUSTIN	AYACUCHO HUAMANGA AYACUCHO	2016-2-H	AD-444	Gestión de Recursos Humanos	3	REGULAR	
4	27021206	APERRIGUE CABRERA RENAN CARLOS	v	1984	SAN AGUSTIN	AYACUCHO HUAMANGA AYACUCHO	2016-2-H	AD-444	Gestión de Recursos Humanos	3	REGULAR	

Figura 19: Visualización de los 5 primeros registros del Dataframe.

Para poder realizar un mejor análisis estadístico, se creó una columna adicional “Rendimiento” donde se resume si el alumno aprobó o no; el siguiente bloque de código lo realiza:

```
data['Rendimiento'] = ''
data.loc[data['Promedio_curso'] > 10, 'Rendimiento'] = 'Aprobado'
data.loc[data['Promedio_curso'] < 11, 'Rendimiento'] = 'Desaprobado'
data.head()
```

Para la visualización adecuada de los gráficos se usa la función pyplot de la librería matplotlib, el cual ya fue importado al inicio junto con la carga del archivo csv; se procederá a mostrar los gráficos más relevantes relacionados al rendimiento académico.

RESUMEN GENERAL

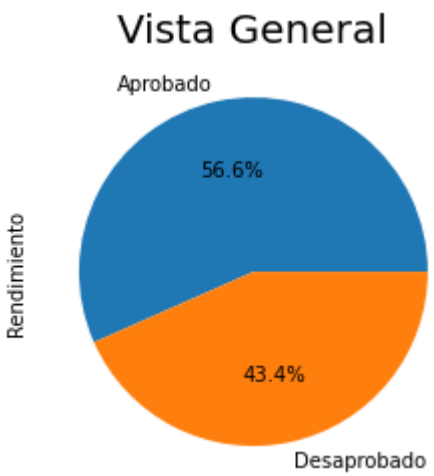


Figura 20: Distribución general del rendimiento académico.

La imagen anterior fue generada por el siguiente código:

```
data.Rendimiento.value_counts().plot.pie(autopct = '%1.1f%%')
plt.title('Vista General', fontsize = 20)
```

En la figura 20 se puede notar que la cantidad de desaprobados es muy grande pues la diferencia entre aprobados y desaprobados no es mucha;, a simple vista podría decirse que los alumnos que aprueban son un poco más de la mitad y que el resto se desaprueba, pero el dataframe tiene registros de alumnos que repiten los cursos en diferentes semestres además de las otras modalidades como son los aplazados, vacacional, curso único, etc.; los cuales añaden la cantidad de veces que el alumno se desaprobó, por tanto hay que tener eso en cuenta.

GÉNERO

```
graf_sexo = data.groupby(['Rendimiento',"Genero"]).count()['Codigo']
graf_sexo.unstack(level=0).plot.bar()
plt.title("Rendimiento por Género", fontsize = 20)
plt.xlabel("Género", fontsize = 12, color = 'blue')
plt.ylabel("Cantidad", fontsize = 12, color = 'blue')
```

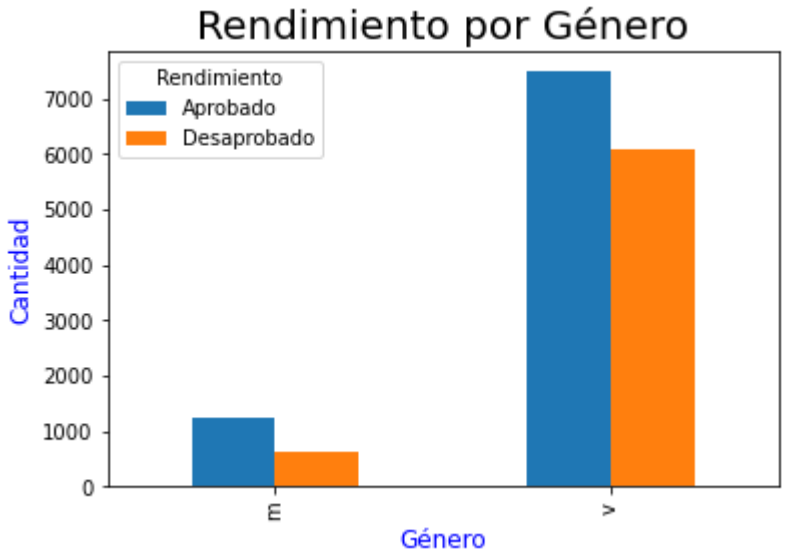


Figura 21: Distribución del rendimiento académico según el género.

En la figura 21 se logra observar que la cantidad de aprobados en ambos géneros es relativamente mayor comparado con los desaprobados, se podría decir que la relación entre aprobados y desaprobados es similar a la mostrada en la imagen 4.5, pues en ambos gráficos se observa que no hay una brecha grande; también es notorio la mayor cantidad de varones, esto era de esperarse pues la cantidad de mujeres en la carrera es poca.

SEMESTRE ACADÉMICO

```
graf_semestre = data.groupby(['Rendimiento',"Semestre_academico"]).count()['Codigo']
graf_semestre.unstack(level=0).plot.bar(figsize=(10, 5))
plt.title("Rendimiento por Semestre Académico", fontsize = 20)
plt.xlabel("Semestre Académico", fontsize = 12, color = 'blue')
plt.ylabel("Cantidad", fontsize = 12, color = 'blue')
```

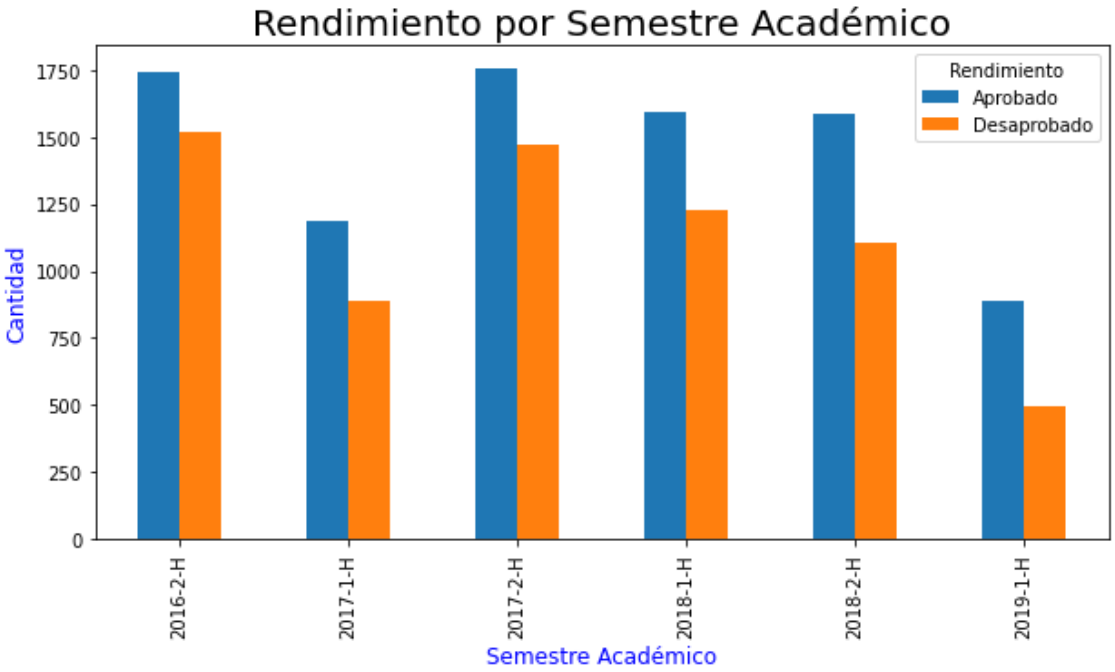


Figura 22: Distribución del rendimiento académico según el semestre académico.

En la figura 22 se corrobora la relación anteriormente dicha, pues aquí también los aprobados no se alejan mucho de los desaprobados; se observa que las alturas de las barras es fluctuante debido al movimiento que generan los alumnos, por ejemplo, en el semestre 2016-II aprobaron y desaprobaron una gran cantidad de estudiantes lo que influyó en el siguiente semestre impar, pues los desaprobados no pueden matricularse en las secuencias de los cursos y aumentando el registro de matriculados en el siguiente semestre par.

MODALIDAD DEL CURSO

```
graf_modalidad = data.groupby(['Rendimiento',"Modalidad_curso"]).count()['Codigo']
graf_modalidad.unstack(level=0).plot.bar()
plt.title("Rendimiento por Modalidad del Curso", fontsize = 20)
plt.xlabel("Modalidad de curso", fontsize = 12, color = 'blue')
plt.ylabel("Cantidad", fontsize = 12, color = 'blue')
```

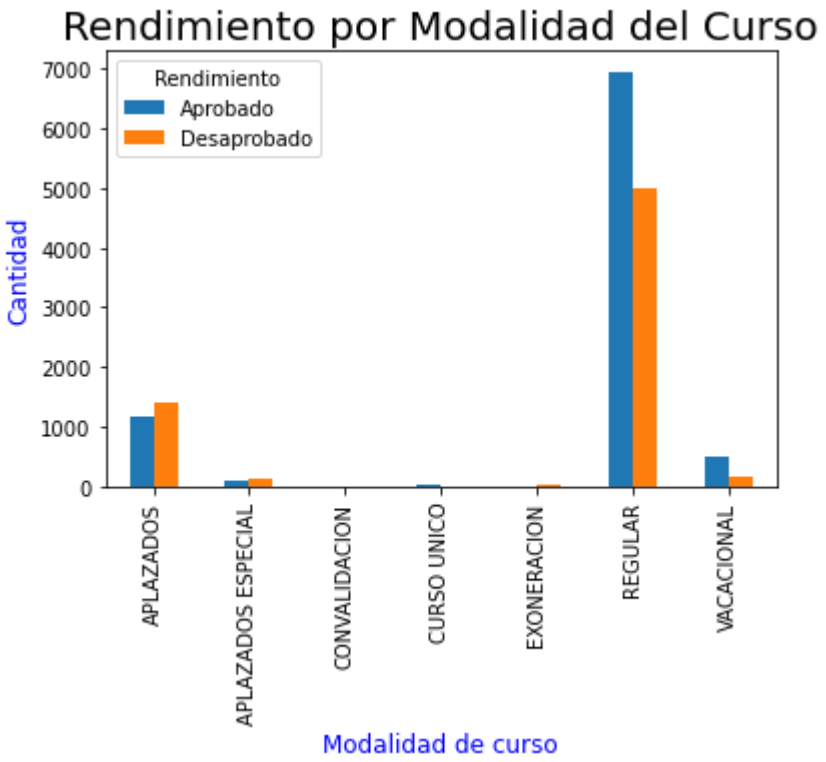


Figura 23: Distribución del rendimiento académico según la modalidad del curso.

En la figura 23 se puede notar que es más probable aprobar una asignatura si se cursa regularmente en vez de rendir el examen de aplazados, pues la cantidad de desaprobados en el aplazado es mayor que el de aprobados; pero también se debe nombrar el hecho que en la universidad existe un limitante respecto a la cantidad de cursos que se pueden dar en los aplazados lo que influye en la gráfica. Continuando, se debe indicar que: en las modalidades vacacional y curso único se desaprueban pocos alumnos, en la modalidad de convalidación todos aprobaron y en la modalidad de exoneración casi todos desaprueban. A modo de síntesis se puede indicar que la mejor modalidad para aprobar es el vacacional, pues el porcentaje de éxito es mayor teniendo en cuenta la cantidad de alumnos que se matriculan; y, que la modalidad menos recomendable es la exoneración.

CRÉDITOS DEL CURSO

```
graf_creditos = data.groupby(['Rendimiento',"Creditos"]).count()['Codigo']
graf_creditos.unstack(level=0).plot.bar()
plt.title("Rendimiento por Créditos del Curso", fontsize = 20)
plt.xlabel("Créditos del Curso", fontsize = 12, color = 'blue')
plt.ylabel("Cantidad", fontsize = 12, color = 'blue')
```

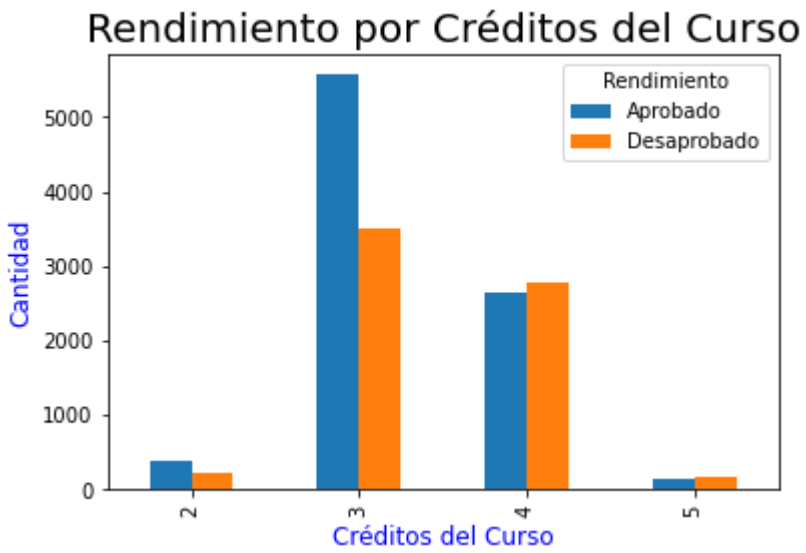
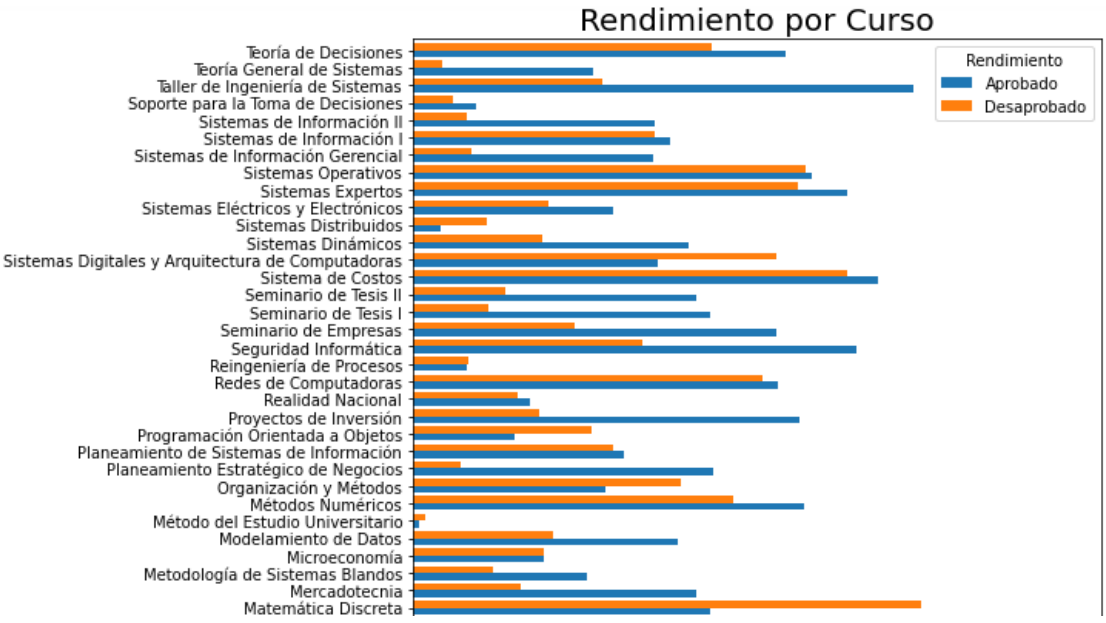


Figura 24: Distribución del rendimiento académico según los créditos.

En la figura 24 se puede notar que hay más desaprobados en los cursos de 4 y 5 créditos, si se observa la base de datos se puede notar que la mayoría de los alumnos que llevan esos cursos lo dejan a la mitad del semestre o desde el inicio, pues existen muchos registros con notas de cero,

CURSO

```
graf_curso = data.groupby(['Rendimiento',"Curso"]).count()['Codigo']
graf_curso.unstack(level=0).plot.barh(width = 0.8,figsize = (8,15))
plt.title("Rendimiento por Curso", fontsize = 20)
plt.xlabel("Cantidad", fontsize = 12, color = 'blue')
plt.ylabel("Curso", fontsize = 12, color = 'blue')
```



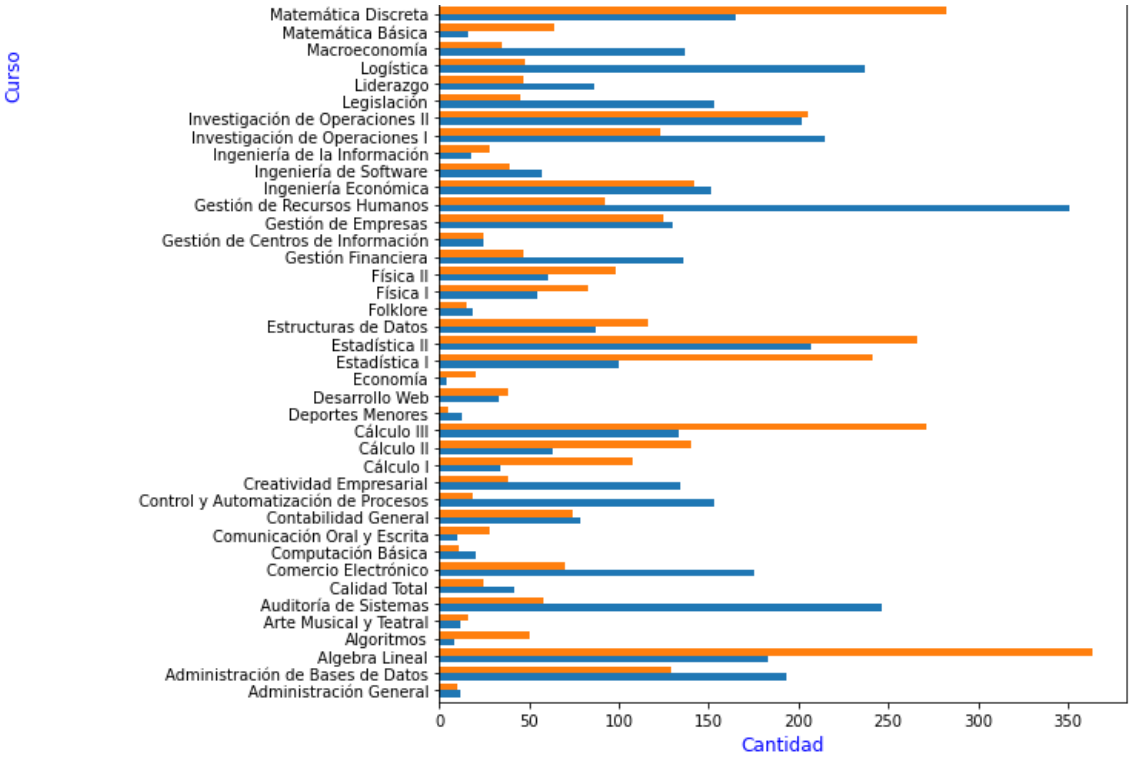
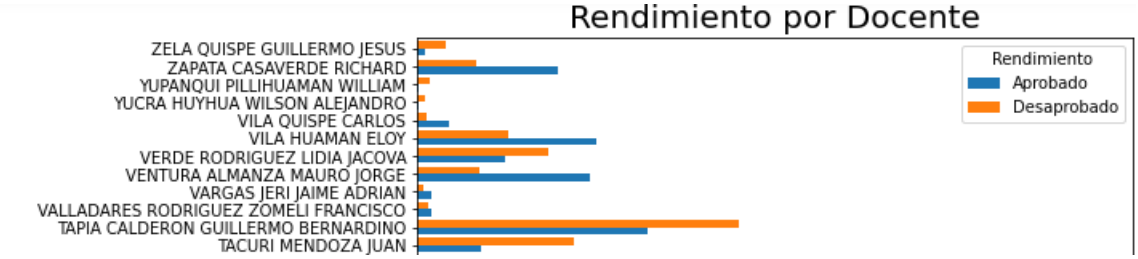


Figura 25: Distribución del rendimiento académico según los cursos.

En la figura 25 se puede notar que los cursos de Gestión de Recursos Humanos y Taller de Ingeniería de Sistemas son los que presentan un mayor número de aprobados, ambos son dictados por diferentes departamentos académicos, por tanto no se puede sacar una conclusión; sin embargo los cursos Algebra Lineal y Matemática Discreta presentan el mayor número de desaprobados, pero en la gráfica se observa que todos los cursos de matemática son los que presentan una predominancia de desaprobados, esto nos indica que el cuello de botella más estrecho está en esos cursos.

DOCENTE

```
graf_docente = data.groupby(['Rendimiento',"Docente"]).count()['Codigo']
graf_docente.unstack(level=0).plot.barh(width = 0.8,figsize = (10,20))
plt.title("Rendimiento por Docente", fontsize = 20)
plt.xlabel("Cantidad", fontsize = 12, color = 'blue')
plt.ylabel("Docente", fontsize = 12, color = 'blue')
```



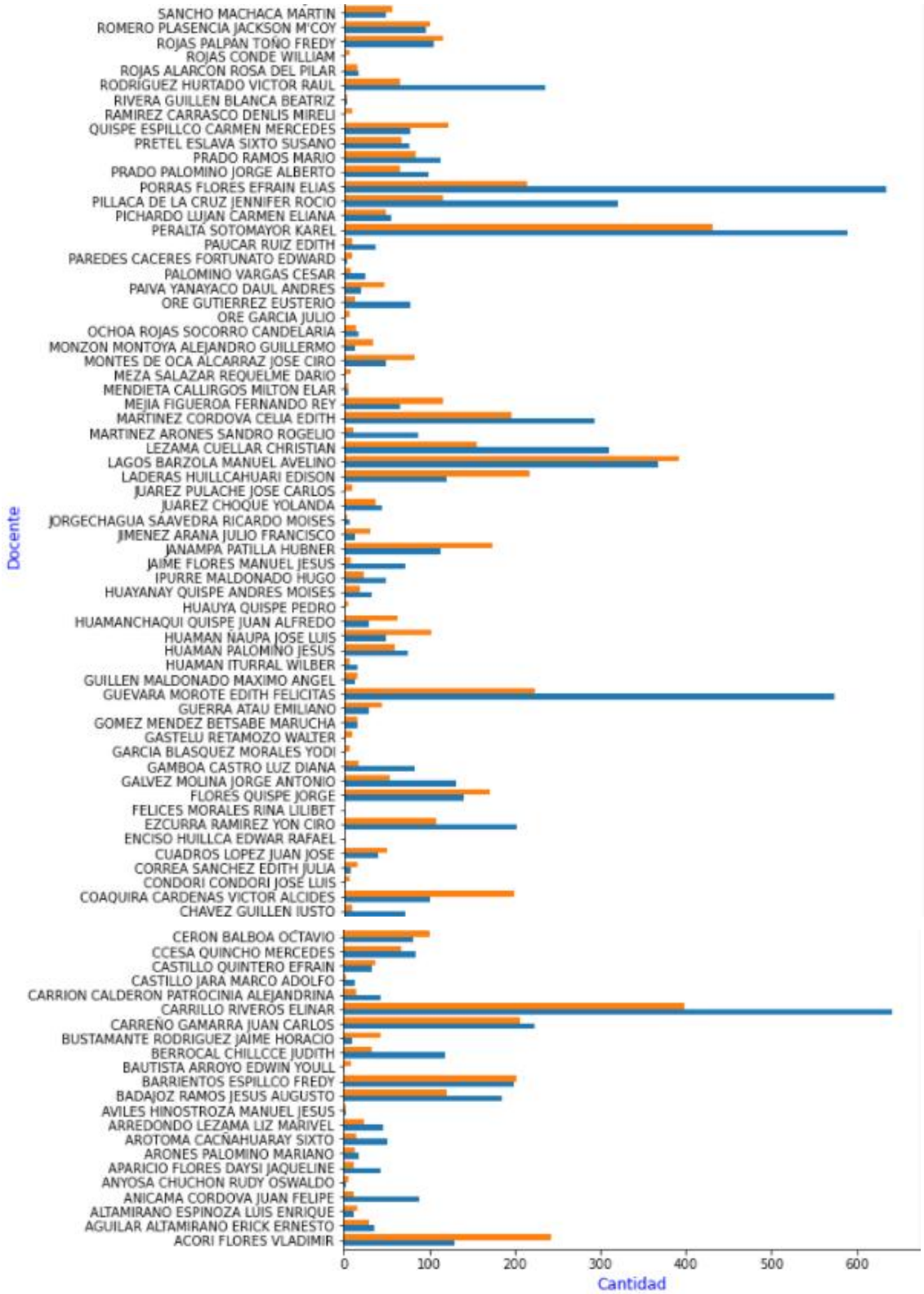


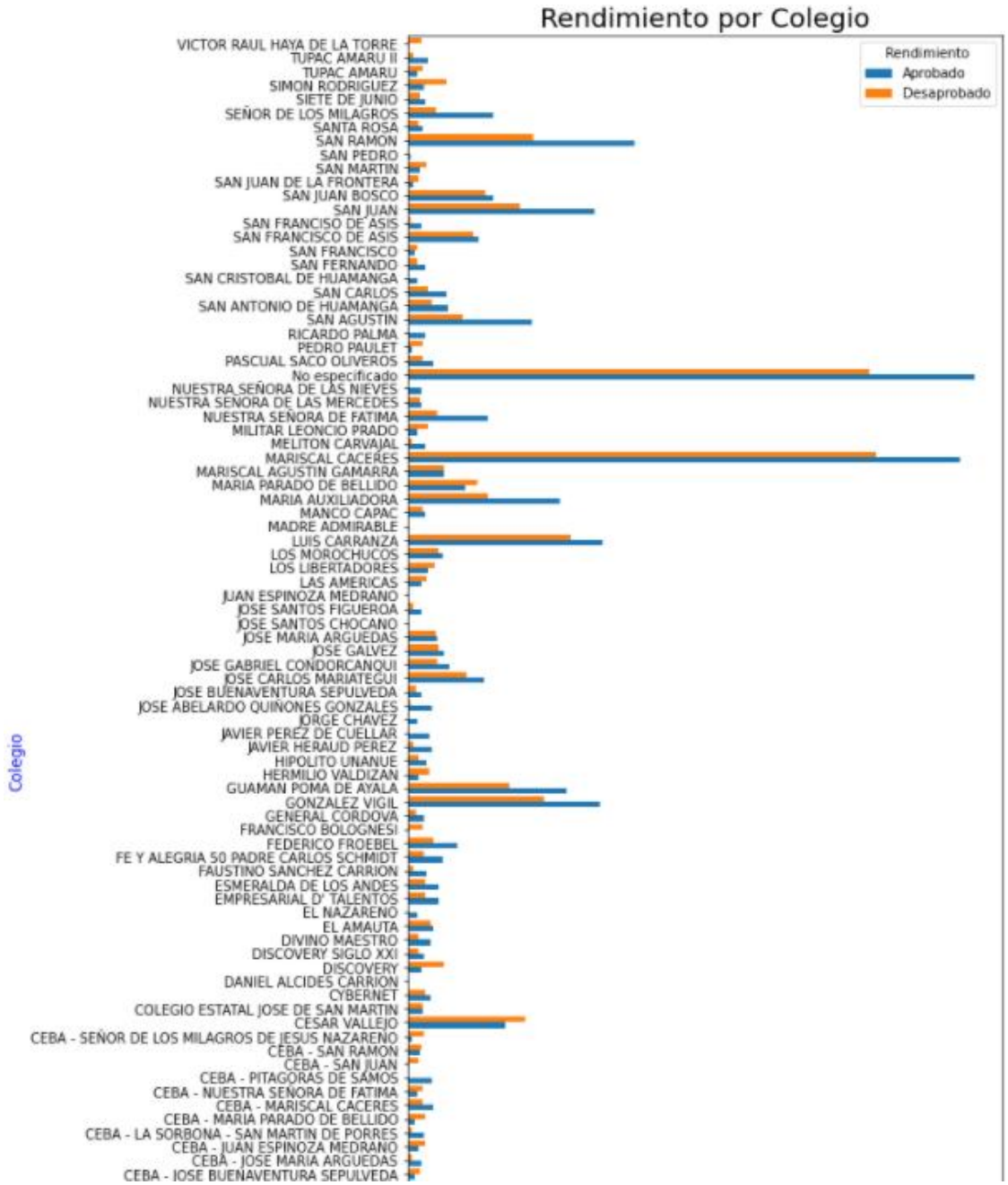
Figura 26: Distribución del rendimiento académico según docentes.

En la figura 26 se observa que los docentes con el mayor número de aprobados son los ingenieros Efraín Porras y Elinar Carrillo, los docentes con más estudiantes desaprobados son los ingenieros Karel Peralta y Elinar Carrillo. Se debe tener en cuenta que todo depende de cuantos cursos dicta el docente, cuántos alumnos se

matricularon, la modalidad del curso, la dificultad del curso y la exigencia del docente; pues, como ya se dijo antes, existen muchos registros que muestran que los alumnos simplemente dejaron el curso ya sea en el ciclo regular o en el aplazado.

COLEGIO DE EGRESO

```
graf_colegio = data.groupby(['Rendimiento',"Colegio"]).count()['Codigo']
graf_colegio.unstack(level=0).plot.barh(width = 0.8,figsize = (8,20))
plt.title("Rendimiento por Colegio", fontsize = 20)
plt.xlabel("Cantidad", fontsize = 12, color = 'blue')
plt.ylabel("Colegio", fontsize = 12, color = 'blue')
```



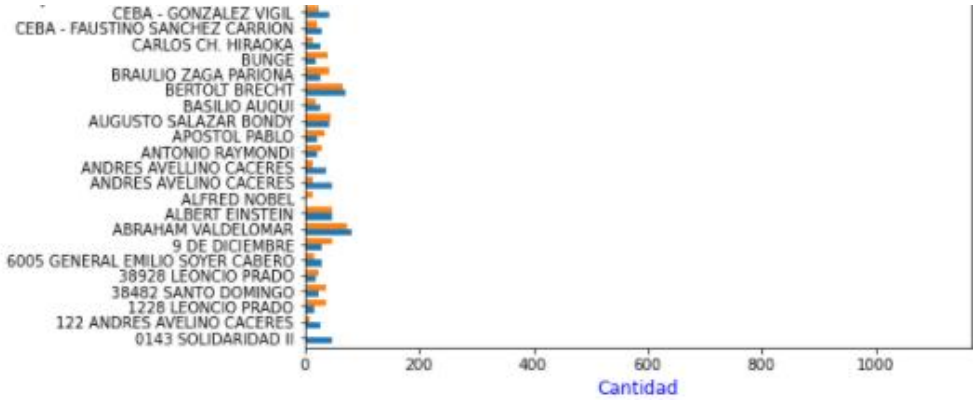


Figura 27: Distribución del rendimiento académico según el colegio de egreso.

En la figura 27 se puede notar claramente que los alumnos proceden de diferentes colegios resaltando el interés de superación y la competitividad; se debe indicar que hubo un gran número de alumnos que no tenían este dato y se le agrupó bajo la denominación de “No especificado” pues no había manera de ubicarlos en un colegio o de eliminarlos; dicho esto, se observa que la mayor cantidad de alumnos proceden del colegio Mariscal Cáceres y muy por debajo se encuentra San Ramón como segunda fuente de estudiantes.

4.1.7 PREPARACIÓN DE DATOS

Como siguiente paso, en esta etapa se procede a preparar los datos para ejecutar los modelos predictivo; debemos recordar que en la etapa de recopilación de datos se detalló el análisis previo que se realizó, pero a modo de resumen se listará lo hecho:

- Se filtraron los registros para que sólo quedasen los registros de Ingeniería de Sistemas.
- Se eliminó la columna Lugar_nacimiento pues estaba vacía.
- Se eliminaron las columnas Modalidad_ingreso y Plan_estudios pues ambos tenían sólo un tipo de dato.
- Se completaron las celdas vacías.
- Se eliminó la columna Tipo_matrícula por ser incoherente.
- Se eliminaron registros duplicados.

SELECCIÓN DE DATOS

Teniendo en cuenta los objetivos del proyecto, se observó y analizó la data que se tiene y se llegó a la siguiente conclusión:

- Se debe añadir una nueva columna llamada “Rendimiento” que tenga los valores 1 (Aprobado) y 0 (desaprobado); pues esta columna se definirá como la

variable a predecir y su valor será establecido por la columna “Promedio_curso”.

- Se estableció como variables predictoras a las columnas: Colegio, Curso, Creditos, Modalidad_curso y Docente.

LIMPIEZA DE DATOS

Para el uso de los modelos de machine learning se necesitan sólo las variables predictoras, para ello se procedió a borras las columnas que no se van a usar o que su aporte al momento de predecir el rendimiento va a ser insignificante, las columnas eliminadas son: código, estudiante, género, fecha de nacimiento, lugar del colegio, semestre académico, sigla y promedio del curso; el siguiente código realiza dicha acción:

```
dataRendimiento = dataRendimiento.drop(['Codigo','Estudiante','Genero','Fecha_nacimiento',
                                        'Lugar_colegio','Semestre_academico','Sigla',
                                        'Promedio_curso'],axis=1)
```

	Colegio	Curso	Creditos	Modalidad_curso	Docente	Rendimiento
0	SAN AGUSTIN	Sistemas Expertos	4	REGULAR	CARRILLO RIVEROS ELINAR	0
1	SAN AGUSTIN	Gestión Financiera	3	REGULAR	PRADO PALOMINO JORGE ALBERTO	1
2	SAN AGUSTIN	Sistemas de Información Gerencial	3	REGULAR	PERALTA SOTOMAYOR KAREL	0
3	SAN AGUSTIN	Gestión de Recursos Humanos	3	REGULAR	MARTINEZ ARONES SANDRO ROGELIO	1
4	SAN AGUSTIN	Gestión de Recursos Humanos	3	REGULAR	ANICAMA CORDOVA JUAN FELIPE	1

Figura 28: Columnas a usar para el modelo.

Es necesario aclarar que ya no se buscará en este dataframe espacio vacíos o repetidos pues este proceso ya se realizó; además que la información fue solicitada a la Oficina General de Informática y Sistemas especificando los detalles que deseábamos de los estudiantes, es decir que ya se tenía una idea de qué información usar.

TRANSFORMACIÓN DE DATOS

Los modelos de machine learning solamente usan tipos de datos numéricos por tanto es conveniente transformar todas las variables cualitativas a cuantitativas, por tanto, se deben usar lógicas propias o usando las librerías de Python para poder realizar dichas transformaciones. Por ejemplo, la columna rendimiento almacena un dato dicotómico pues solo puede tomar dos valores (aprobado y desaprobado), por tanto, se puede asignar al 1 como aprobado y al 0 como desaprobado; el siguiente código realiza esta acción:

```
dataRendimiento["Rendimiento"] = (dataRendimiento["Rendimiento"]==1).astype(int)
dataRendimiento["Rendimiento"] = (dataRendimiento["Rendimiento"]==0).astype(int)
```

Para saber qué tipo de dato fueron almacenados en las columnas, el siguiente código muestra la información del dataframe:

```
dataRendimiento.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15441 entries, 0 to 15440
Data columns (total 6 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   Colegio          15441 non-null  object
 1   Curso            15441 non-null  object
 2   Creditos         15441 non-null  int64 
 3   Modalidad_curso  15441 non-null  object
 4   Docente          15441 non-null  object
 5   Rendimiento      15441 non-null  int64 
dtypes: int64(2), object(4)
memory usage: 723.9+ KB
```

Figura 29: Tipo inicial de datos que tiene el dataframe.

Se puede notar que las columnas que son de tipo numérico son créditos y rendimiento; las demás son de tipo object, esto nos indica que las otras cuatro columnas deben ser modificadas.

Pero observando el contenido de las columnas se observa que no hay un patrón pequeño de datos o una relación que se repita, por ejemplo, hay más profesores que modalidades del curso; por tanto, para transformar los datos categóricos a numéricos se usa la librería panda pues tiene una codificación llama “one-hot” que básicamente lo que hace es crear variables ficticias (dummy) por cada tipo de dato diferente.

Las variables dummy se crean en un nuevo dataframe por tanto se debe añadir al original para tener un solo data y claro que después de concaterar todas las nuevas columnas creadas se debe eliminar la columna original para que no genere confusión. El siguiente código realiza esta transformación para la columna colegio:

```
ColegioDummy = pd.get_dummies(dataRendimiento["Colegio"], prefix="Colegio")
dataRendimiento = pd.concat([dataRendimiento,ColegioDummy],axis=1)
dataRendimiento = dataRendimiento.drop(["Colegio"], axis=1)
```

Se procede a realizar la misma acción para las demás columnas: curso, modalidad del curso y docente:

```
CursoDummy = pd.get_dummies(dataRendimiento["Curso"], prefix="Curso")
dataRendimiento = pd.concat([dataRendimiento,CursoDummy],axis=1)
dataRendimiento = dataRendimiento.drop(["Curso"], axis=1)

ModalidadDummy = pd.get_dummies(dataRendimiento["Modalidad_curso"], prefix="Modalidad_curso")
dataRendimiento = pd.concat([dataRendimiento,ModalidadDummy],axis=1)
dataRendimiento = dataRendimiento.drop(["Modalidad_curso"], axis=1)

DocenteDummy = pd.get_dummies(dataRendimiento["Docente"], prefix="Docente")
dataRendimiento = pd.concat([dataRendimiento,DocenteDummy],axis=1)
dataRendimiento = dataRendimiento.drop(["Docente"], axis=1)
```

Para comprobar se llama nuevamente a la función info() del detaframe para observar los detalles.

```
dataRendimiento.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15441 entries, 0 to 15440
Columns: 283 entries, Creditos to Docente_ZELA QUISPE GUILLERMO JESUS
dtypes: int64(2), uint8(281)
memory usage: 4.4 MB
```

Figura 30: Tipo final de datos que tiene el dataframe.

En la figura 30 se puede observar que hay 283 columnas que empiezan en créditos y termina en Docente_ZELA QUISPE GUILLERMO JESUS; y que los tipos de datos son int64 (sólo 2 columnas) y uint8 (son 281 columnas creadas usando dummy); el código siguiente muestra una parte del dataframe creado:

```
dataRendimiento[700:705]
```

	Creditos	Rendimiento	Colegio_0143 SOLIDARIDAD II	Colegio_122 ANDRES AVELINO CACERES	Colegio_1228 LEONCIO PRADO	Colegio_38482 SANTO DOMINGO	Colegio_38928 LEONCIO PRADO	Colegio_6005 GENERAL EMILIO SOYER CABERO	Colegio_9 DE DICIEMBRE	Colegio_ABRAHAM VALDELOMAR	...
700	4	0	0	0	0	0	0	0	0	0	...
701	4	0	0	0	0	0	0	0	0	0	...
702	3	0	0	0	0	0	0	0	0	0	...
703	3	0	0	0	0	0	0	0	0	0	...
704	4	0	0	0	0	0	0	0	0	0	...

5 rows × 283 columns

Figura 31: Dataframe final.

4.1.8 MODELADO

Para la presente investigación se indicó que se usarían como algoritmos de aprendizaje los de regresión logística y bosques aleatorios (random forest) para el desarrollo del modelo predictivo; como ya la data se encuentra limpia y con los tipos de datos adecuados se debe definir las variables objetivo o a predecir y las variables predictoras.

VARIABLES PREDICTORA Y A PREDECIR

Conociendo los datos que se posee, primero se separará los datos en dos conjuntos: el primero será X que almacenará las variables independientes y el segundo será Y que tendrá a la variable dependiente; los siguientes bloques de código realizan dichas creaciones:

```
dataColumnas = dataRendimiento.columns.values.tolist()

Y = ["Rendimiento"]
X = [v for v in dataColumnas if v not in Y]

var_predictoras = dataRendimiento[X]
var_predictoras[700:705]
```

	Creditos	Colegio_0143 SOLIDARIDAD II	Colegio_122 ANDRES AVELINO CACERES	Colegio_1228 LEONCIO PRADO	Colegio_38482 SANTO DOMINGO	Colegio_38928 LEONCIO PRADO	Colegio_6005 GENERAL EMILIO SOYER CABERO	Colegio_9 DE DICIEMBRE
700	4	0	0	0	0	0	0	0
701	4	0	0	0	0	0	0	0
702	3	0	0	0	0	0	0	0
703	3	0	0	0	0	0	0	0
704	4	0	0	0	0	0	0	0

5 rows × 282 columns

Figura 32: Conjunto X: variables independientes.

```
Y = ["Rendimiento"]
var_predecir = dataRendimiento[Y]
var_predecir[700:705]
```

	Rendimiento
700	0
701	0
702	0
703	0
704	0

Figura 33: Conjunto Y: variable dependiente.

DATOS DE ENTRENAMIENTO Y MUESTRA

También los datos generales se dividirán en dos grupos: uno de prueba y otro de entrenamiento; en machine learning se debe tener en cuenta que la división de datos deja al modelo de predicción con menor cantidad para su entrenamiento, pero si se asigna una cantidad pequeña para el grupo de prueba, irónicamente el modelo se haría impreciso por no tener datos suficientes para testear.

Sabiendo esto, se tomó la decisión de usar el 20% del conjunto de datos como prueba y el resto como datos de entrenamiento; el siguiente código realiza dichas operaciones:

```
from sklearn import linear_model
from sklearn.model_selection import train_test_split

X_train, X_test, Y_train, Y_test = train_test_split(var_predictoras, var_predecir,
                                                    test_size = 0.2, random_state = 0)

X_train.shape, X_test.shape, Y_train.shape, Y_test.shape
```

```
((12352, 282), (3089, 282), (12352, 1), (3089, 1))
```

APLICACIÓN DE LOS ALGORITMOS

A. REGRESIÓN LOGÍSTICA

Una vez separada la data en los grupos de datos adecuados y necesarios, se procede a aplicar los algoritmos y uno de ellos es la regresión logística.

Entrenamiento

Para el entrenamiento se emplea la librería sk-learn, el cual en términos generales somete al dataframe a analizar los datos con la finalidad de obtener un patrón, con la finalidad que después pueda predecir la variable objetivo a partir de datos nuevos o con una muestra de testeo; el siguiente código realiza el entrenamiento del modelo de regresión logística:

```
from sklearn.linear_model import LogisticRegression

model_logist = LogisticRegression(solver='lbfgs', max_iter=1000)
model_logist.fit(X_train, Y_train.values.ravel())
```

```
LogisticRegression(max_iter=1000)
```

```
model_logist.score(X_train, Y_train)
```

```
0.6823186528497409
```

Se puede observar que se obtuvo una calificación de proximidad del 68%, empleando los datos de entrenamiento; podría decirse que casi un 70% lo cual puede considerarse aceptable.

Validación Cruzada

Esta es una manera de estimar la precisión del modelo, como sabemos que la validación cruzada requiere un número de particiones, se decidió que k = 10, para que la data sea dividida en 10 subconjuntos; a grandes rasgos, la validación cruzada consiste en hallar la media aritmética operando con los datos de las particiones y en diferente orden; es decir que la primera iteración tomará al primer grupo como prueba y a los nueve restantes como entrenamiento, la segunda iteración tomara el segundo grupo y así hasta llegar a la décima iteración. EL siguiente código realiza dicha operación:

```
from sklearn.model_selection import cross_val_score
valid_cruz = cross_val_score(LogisticRegression(solver='lbfgs', max_iter=1000), var_predictoras,
                             var_predecir.values.ravel(), scoring="accuracy", cv=10)
valid_cruz
```

```
array([0.6381877 , 0.58290155, 0.61139896, 0.54663212, 0.57772021,
       0.51165803, 0.56735751, 0.52784974, 0.62888601, 0.64119171])
```

```
valid_cruz.mean()
```

```
0.5833783557187652
```

Se puede observar que se primero se obtuvo un array con el resultado de las 10 iteraciones y que el promedio final es de 0.583.

Matriz de Confusión y Curva ROC

Esta matriz es la herramienta más útil y es muy usada en el campo de la ciencia de datos, pues nos muestra de manera sencilla el desenvolvimiento del algoritmo usado, las columnas de la matriz nos mostrarán los valores del rendimiento (1 = aprobado y 0 desaprobado), mientras que las filas serán llenadas por datos reales de acierto o no; el siguiente código nos permite graficar dicha matriz:

```
probs = model_logist.predict_proba(X_test)
prob = probs[:,1]
prob_df = pd.DataFrame(prob)
threshold = 0.15
prob_df["prediccion"] = np.where(prob_df[0]>=threshold, 1, 0)
prob_df["actual"] = np.array(Y_test)
confusion_matrix = pd.crosstab(prob_df.prediccion, prob_df.actual)
confusion_matrix
```

actual	0	1
prediccion		
0	34	7
1	1302	1746

Figura 34: Matriz de confusión del algoritmo de regresión logística.

En la figura 34 se puede observar que el algoritmo predijo correctamente que 34 alumnos se desaprobarán, esto representa casi 5 veces la cantidad que erró (7), pero al momento de predecir los alumnos aprobados no tuvo tanto éxito pues acertó 1756 y se equivocó en 1302.

Gracias a la matriz ya podemos calcular la sensibilidad y especificidad, donde los datos son sacados de dicha matriz, las fórmulas son:

$$Sensibilidad = \frac{VP}{VP + FN}$$

$$Especificidad = \frac{VN}{VN + FP}$$

Donde: VP = Verdaderos positivos
VN = Verdaderos negativos
FP = Falsos positivos
FN = Falsos negativos

EL siguiente código realiza el calcula para hallar dichas variables:

```
TN = confusion_matrix[0][0]
TP = confusion_matrix[1][1]
FP = confusion_matrix[0][1]
FN = confusion_matrix[1][0]

sens = TP/(TP+FN)
espc = 1-TN/(TN+FP)
sens, espc
```

```
(0.9577866514546491, 0.8360778443113772)
```

Una vez hallados la sensibilidad y especificidad, ahora se puede graficar la curva ROC (Receiver Operating Characteristic), para tal fin se emplea la librería sk-learn y matplotlib; los detalles del código se muestran a continuación:

```
import matplotlib.pyplot as plt
from sklearn import metrics

espc, sens, _ = metrics.roc_curve(Y_test, prob)
%matplotlib inline
auc = metrics.auc(espc, sens)
plt.plot(espc, sens, marker="o", linestyle="--", color="r", label="Data, auc="+str(auc))
x = [i*0.01 for i in range(100)]
y = [i*0.01 for i in range(100)]
plt.plot(x,y)
plt.xlabel("Especifidad")
plt.ylabel("Sensibilidad")
plt.title("CURVA ROC")
plt.legend(loc=4)
```

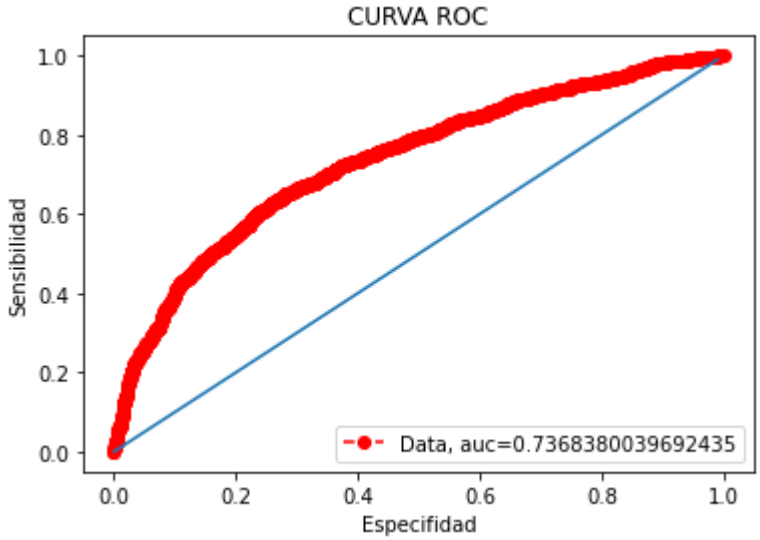


Figura 35: Curva ROC del modelo de regresión logística.

En la figura 35 se observa la curva ROC, y el área debajo de dicha curva representa que tan eficaz es nuestro modelo, y resulta que tiene una probabilidad del 73.6% lo cual es aceptable; lo ideal sería que el área nos devuelva un resultado de 1 pero eso, además de ser imposible, nos indicaría que los datos han sido manipulados pues todos los modelos predictivos siempre tendrán errores por más mínimo que sea.

Como dato adicional se debe indicar que, si la curva está por encima de la línea, significa que sí cumple su función de predicción; en cambio, cuando la cuerva está por debajo, simplemente el modelo no sirve.

B. K VECINOS CERCANOS

Este tipo de algoritmo pertenece a los algoritmos de clasificación, siendo uno de los más básicos, pero no por eso dejan de ser eficientes; en esencia busca en los datos cercanos a la variable a predecir un comportamiento del cual realizar sus cálculos.

Entrenamiento

La sintaxis del código es la misma que la mostrada en el algoritmo de regresión logística, por lo que no se redundará explicando cómo se entrena, en cambio, se mostrará directamente el código de entrenamiento:

```
from sklearn.neighbors import KNeighborsClassifier

model_neigh = KNeighborsClassifier()
model_neigh.fit(X_train, Y_train.values.ravel())
```

```
KNeighborsClassifier()
```

Validación Cruzada

Se aplicará la validación cruzada de la misma manera, eso quiere decir que con 10 iteraciones (k = 10); el siguiente código realiza dicha operación:

```
from sklearn.model_selection import cross_val_score
valid_cruz_kn = cross_val_score(KNeighborsClassifier(), var_predictoras,
                                var_predecir.values.ravel(), scoring="accuracy", cv=10)
valid_cruz_kn
```

```
array([0.64983819, 0.59909326, 0.62305699, 0.61852332, 0.61722798,
       0.58743523, 0.61528497, 0.6003886 , 0.62888601, 0.62694301])
```

```
valid_cruz_kn.mean()
```

```
0.6166677565940606
```

Se puede observar que el promedio de las 10 iteraciones es de 0.616.

Matriz de Confusión y Curva ROC

Los pasos son los mismos que fueron explicados anteriormente, entonces primero se hace la matriz de confusión, el código es el siguiente:

```
probs_kn = model_neigh.predict_proba(X_test)
prob_kn = probs_kn[:,1]
prob_df_kn = pd.DataFrame(prob_kn)
threshold = 0.15
prob_df_kn["prediccion"] = np.where(prob_df_kn[0]>=threshold, 1, 0)
prob_df_kn["actual"] = np.array(Y_test)
confusion_matrix_kn = pd.crosstab(prob_df_kn.prediccion, prob_df_kn.actual)
confusion_matrix_kn
```

actual	0	1
prediccion		
0	104	42
1	1232	1711

Figura 36: Matriz de confusión del algoritmo de k vecinos cercanos.

En la figura 36 se puede observar que el algoritmo es mejor prediciendo a los desaprobados que a los aprobados, pues acertó que 104 desaprobarían y falló 42; ahora, si vemos la cantidad de aprobados que acertó es de 1711 pero la cantidad que falló supera más de mil casos, eso es algo a tener en cuenta.

Continuando, es momento de calcular la sensibilidad y especificidad; no es necesario describir las fórmulas sino escribir el código que lo calcula:

```
TN = confusion_matrix_kn[0][0]
TP = confusion_matrix_kn[1][1]
FP = confusion_matrix_kn[0][1]
FN = confusion_matrix_kn[1][0]

sens_1 = TP/(TP+FN)
espc_1 = 1-TN/(TN+FP)
sens_1, espc_1
```

```
(0.9760410724472333, 0.9221556886227544)
```

Ya se tiene lo necesario para graficar la curva ROC, para ello se escribe el siguiente código:

```
import matplotlib.pyplot as plt
from sklearn import metrics

espc_1, sens_1, _ = metrics.roc_curve(Y_test, prob_kn)
%matplotlib inline
auc = metrics.auc(espc_1, sens_1)
plt.plot(espc_1, sens_1, marker="o", linestyle="--", color="r", label="Data, auc="+str(auc))
x = [i*0.01 for i in range(100)]
y = [i*0.01 for i in range(100)]
plt.plot(x,y)
plt.xlabel("Especifidad")
plt.ylabel("Sensibilidad")
plt.title("CURVA ROC")
plt.legend(loc=4)
```

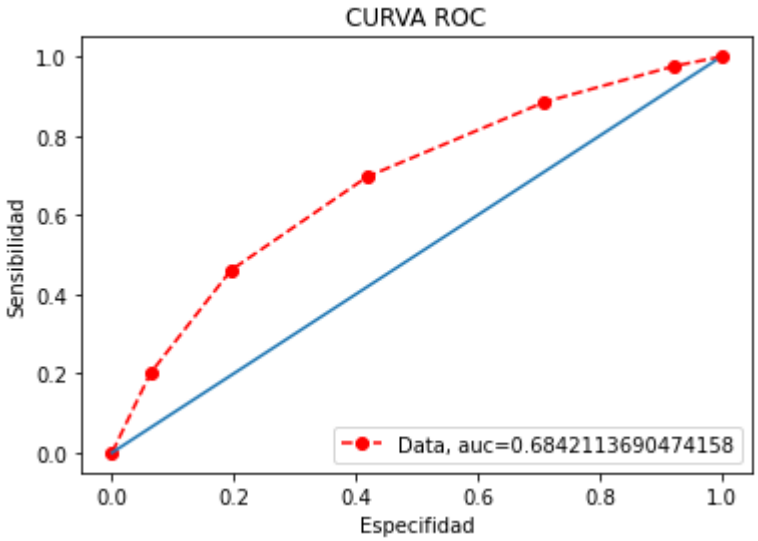


Figura 37: Curva ROC del modelo de k vecinos cercanos.

En la figura 37 se puede indicar que el modelo es apropiado pues sirve para predecir (la curva está por encima de la línea) teniendo una probabilidad del 68.4%, lo cual es aceptable.

C. MÁQUINA DE VECTOR DE SOPORTE

Este tipo de algoritmo que básicamente se centra en dividir los datos en 2 grupos creando un margen de separación, cuya separación determinará si la elección fue razonable.

Entrenamiento

La sintaxis del código es la misma que las mostradas anteriormente, por lo que no se redundará explicando cómo se entrena, en cambio, se mostrará directamente el código de entrenamiento:

```
from sklearn.svm import SVC

model_svc = SVC(probability = True)
model_svc.fit(X_train, Y_train.values.ravel())
```

```
SVC(probability=True)
```

Validación Cruzada

Como todos los algoritmos son de agrupación, los métodos de validación son las mismas, por tanto, se aplicará la validación cruzada de la misma manera, eso quiere decir que con 10 iteraciones (k = 10); el siguiente código realiza dicha operación:

```
from sklearn.model_selection import cross_val_score
valid_cruz_svc = cross_val_score(SVC(), var_predictoras,
                                 var_predecir.values.ravel(), scoring="accuracy", cv=10)
valid_cruz_svc
```

```
array([0.65436893, 0.59715026, 0.64119171, 0.60103627, 0.62823834,
       0.57124352, 0.60816062, 0.57901554, 0.65867876, 0.66515544])
```

```
valid_cruz_svc.mean()
```

```
0.6204239398360079
```

Se puede observar que el promedio de las 10 iteraciones es de 0.62.

Matriz de Confusión y Curva ROC

Los pasos son los mismos que fueron explicados anteriormente, entonces primero se hace la matriz de confusión, el código es el siguiente:

```
probs_svc = model_svc.predict_proba(X_test)
prob_svc = probs_svc[:,1]
prob_df_svc = pd.DataFrame(prob_svc)
threshold = 0.15
prob_df_svc["prediccion"] = np.where(prob_df_svc[0]>=threshold, 1, 0)
prob_df_svc["actual"] = np.array(Y_test)
confusion_matrix_svc = pd.crosstab(prob_df_svc.prediccion, prob_df_svc.actual)
confusion_matrix_svc
```

actual	0	1
prediccion		
0	2	2
1	1334	1751

Figura 38: Matriz de confusión del algoritmo máquina de vector de soporte.

En la figura 38 se puede observar que el algoritmo tuvo una inclinación para predecir si el alumno aprobará, pues se observa claramente que la cantidad de alumnos que fueron predichos su desaprobación fue solamente de 4.

Continuando, es momento de calcular la sensibilidad y especificidad; no es necesario describir las fórmulas sino escribir el código que lo calcula:

```
TN = confusion_matrix_svc[0][0]
TP = confusion_matrix_svc[1][1]
FP = confusion_matrix_svc[0][1]
FN = confusion_matrix_svc[1][0]

sens_2 = TP/(TP+FN)
espc_2 = 1-TN/(TN+FP)
sens_2, espc_2
```

```
(0.9988590986879635, 0.9985029940119761)
```

Ya se tiene lo necesario para graficar la curva ROC, para ello se escribe el siguiente código:

```
import matplotlib.pyplot as plt
from sklearn import metrics

espc_2, sens_2, _ = metrics.roc_curve(Y_test, prob_svc)
%matplotlib inline
auc = metrics.auc(espc_2, sens_2)
plt.plot(espc_2, sens_2, marker="o", linestyle="--", color="r", label="Data, auc="+str(auc))
x = [i*0.01 for i in range(100)]
y = [i*0.01 for i in range(100)]
plt.plot(x,y)
plt.xlabel("Especifidad")
plt.ylabel("Sensibilidad")
plt.title("CURVA ROC")
plt.legend(loc=4)
```

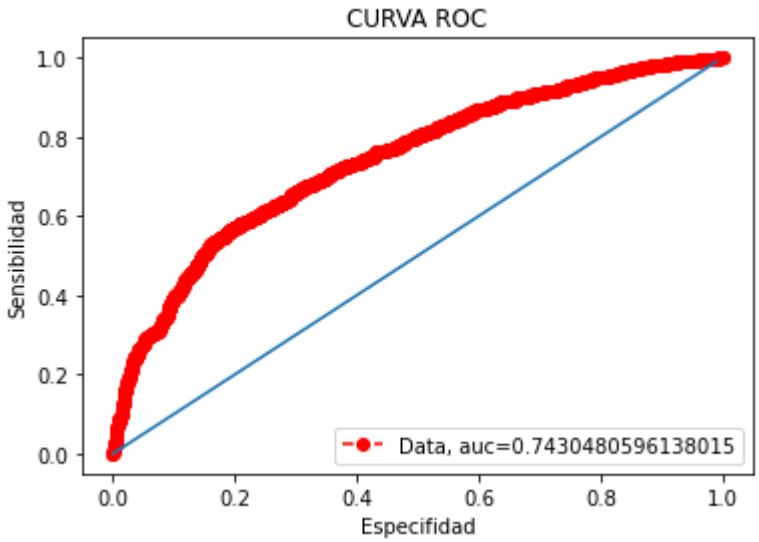


Figura 39: Curva ROC del modelo de máquina de vector de soporte.

En la figura 39 se puede observar que la probabilidad de predicción es del 74.3%, la más alta registrada, lo que a priori, nos indicaría que este es el mejor algoritmo, pero, para determinar dicha afirmación se deben evaluar los demás factores, pues debemos recordar que este algoritmo tiende a inclinarse hacia el grupo de aprobados.

D. ÁRBOL DE DECISIÓN

Este algoritmo es flexible ya que los datos pueden ser tanto cualitativas como cuantitativas, pero en esta investigación se tuvo que cuantificar todas las variables para poder trabajar junto con los demás algoritmos.

Entrenamiento

La sintaxis del código es la misma que las mostradas anteriormente, por lo que no se redundará explicando cómo se entrena, en cambio, se mostrará directamente el código de entrenamiento:

```
from sklearn.tree import DecisionTreeClassifier

model_tree = DecisionTreeClassifier()
model_tree.fit(X_train, Y_train.values.ravel())
```

```
DecisionTreeClassifier()
```

Validación Cruzada

Como todos los algoritmos son de agrupación, los métodos de validación son las mismas, por tanto, se aplicará la validación cruzada de la misma manera, eso quiere decir que con 10 iteraciones (k = 10); el siguiente código realiza dicha operación:

```
from sklearn.model_selection import cross_val_score
valid_cruz_tree = cross_val_score(DecisionTreeClassifier(), var_predictoras,
                                  var_predecir.values.ravel(), scoring="accuracy", cv=10)
valid_cruz_tree
```

```
array([0.66796117, 0.59909326, 0.61658031, 0.58937824, 0.625     ,
       0.56023316, 0.58031088, 0.5861399 , 0.60686528, 0.62435233])
```

```
valid_cruz_tree.mean()
```

```
0.6055914532924191
```

Se puede observar que el promedio de las 10 iteraciones es de 0.605.

Matriz de Confusión y Curva ROC

Los pasos son los mismos que fueron explicados anteriormente, entonces primero se hace la matriz de confusión, el código es el siguiente:

```
probs_tree = model_tree.predict_proba(X_test)
prob_tree = probs_tree[:,1]
prob_df_tree = pd.DataFrame(prob_tree)
threshold = 0.15
prob_df_tree["prediccion"] = np.where(prob_df_tree[0]>=threshold, 1, 0)
prob_df_tree["actual"] = np.array(Y_test)
confusion_matrix_tree = pd.crosstab(prob_df_tree.prediccion, prob_df_tree.actual)
confusion_matrix_tree
```

actual	0	1
prediccion		
0	492	278
1	844	1475

Figura 40: Matriz de confusión del algoritmo árbol de decisión.

En la figura 40 se puede observar que el algoritmo tiene muchos valores errados, pues la cantidad de verdaderos negativos y la de falsos negativos no están muy separados de sus respectivos valores positivos.

Continuando, es momento de calcular la sensibilidad y especificidad; no es necesario describir las fórmulas sino escribir el código que lo calcula:

```
TN = confusion_matrix_tree[0][0]
TP = confusion_matrix_tree[1][1]
FP = confusion_matrix_tree[0][1]
FN = confusion_matrix_tree[1][0]

sens_3 = TP/(TP+FN)
espc_3 = 1-TN/(TN+FP)
sens_3, espc_3
```

```
(0.8414147176269253, 0.6317365269461077)
```

Ya se tiene lo necesario para graficar la curva ROC, para ello se escribe el siguiente código:

```
import matplotlib.pyplot as plt
from sklearn import metrics

espc_3, sens_3, _ = metrics.roc_curve(Y_test, prob_tree)
%matplotlib inline
auc = metrics.auc(espc_3, sens_3)
plt.plot(espc_3, sens_3, marker="o", linestyle="--", color="r", label="Data, auc="+str(auc))
x = [i*0.01 for i in range(100)]
y = [i*0.01 for i in range(100)]
plt.plot(x,y)
plt.xlabel("Especifidad")
plt.ylabel("Sensibilidad")
plt.title("CURVA ROC")
plt.legend(loc=4)
```

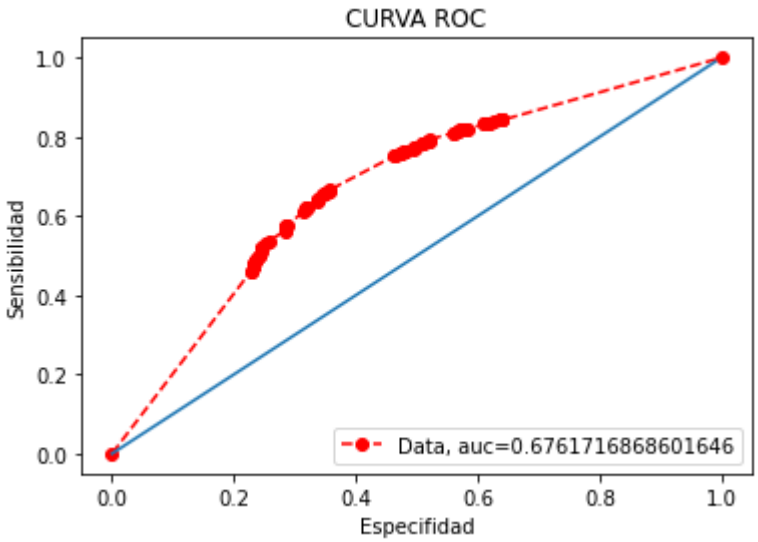


Figura 41: Curva ROC del modelo de árbol de decisión.

En la figura 41 se puede indicar que el modelo es apropiado pues sirve para predecir (la curva está por encima de la línea) teniendo una probabilidad del 67.6%, lo cual es aceptable

E. BOSQUES ALEATORIOS (RANDOM FOREST)

Los bosques aleatorios emplean a muchos árboles de decisión para obtener un modelo más eficiente en comparación a que si usáramos un solo árbol, pues como su nombre lo indica se crean muchos árboles formando un bosque aleatorio.

Entrenamiento

La sintaxis del código es la misma que las mostradas anteriormente, por lo que no se redundará explicando cómo se entrena, en cambio, se mostrará directamente el código de entrenamiento:

```
from sklearn.ensemble import RandomForestClassifier

model_forest = RandomForestClassifier()
model_forest.fit(X_train, Y_train.values.ravel())
```

```
RandomForestClassifier()
```

Validación Cruzada

Como todos los algoritmos son de agrupación, los métodos de validación son las mismas, por tanto, se aplicará la validación cruzada de la misma manera, eso quiere decir que con 10 iteraciones (k = 10); el siguiente código realiza dicha operación:

```
from sklearn.model_selection import cross_val_score
valid_cruz_rf = cross_val_score(RandomForestClassifier(), var_predictoras,
                                var_predecir.values.ravel(), scoring="accuracy", cv=10)
valid_cruz_rf
```

```
array([0.6828479 , 0.61463731, 0.62953368, 0.59196891, 0.62240933,
       0.55181347, 0.59909326, 0.57772021, 0.61852332, 0.63536269])
```

```
valid_cruz_rf.mean()
```

```
0.6123910072605934
```

Se puede observar que el promedio de las 10 iteraciones es de 0.612.

Matriz de Confusión y Curva ROC

Los pasos son los mismos que fueron explicados anteriormente, entonces primero se hace la matriz de confusión, el código es el siguiente:

```
probs_rf = model_forest.predict_proba(X_test)
prob_rf = probs_rf[:,1]
prob_df_rf = pd.DataFrame(prob_rf)
threshold = 0.15
prob_df_rf["prediccion"] = np.where(prob_df_rf[0]>=threshold, 1, 0)
prob_df_rf["actual"] = np.array(Y_test)
confusion_matrix_rf = pd.crosstab(prob_df_rf.prediccion, prob_df_rf.actual)
confusion_matrix_rf
```

actual	0	1
prediccion		
0	219	74
1	1117	1679

Figura 42: Matriz de confusión del algoritmo random forest.

En la figura 42 se puede observar que el algoritmo es mejor prediciendo a los desaprobados que a los aprobados, pues acertó que 219 desaprobarían y falló 74; ahora, si vemos la cantidad de aprobados que acertó es de 1679 pero la cantidad que falló supera más de mil casos, eso es algo a tener en cuenta.

Continuando, es momento de calcular la sensibilidad y especificidad; no es necesario describir las fórmulas sino escribir el código que lo calcula:

```
TN = confusion_matrix_rf[0][0]
TP = confusion_matrix_rf[1][1]
FP = confusion_matrix_rf[0][1]
FN = confusion_matrix_rf[1][0]

sens_1 = TP/(TP+FN)
espc_1 = 1-TN/(TN+FP)
sens_1, espc_1
```

```
(0.9577866514546491, 0.8360778443113772)
```

Ya se tiene lo necesario para graficar la curva ROC, para ello se escribe el siguiente código:

```
import matplotlib.pyplot as plt
from sklearn import metrics

espc_1, sens_1, _ = metrics.roc_curve(Y_test, prob_rf)
%matplotlib inline
auc = metrics.auc(espc_1, sens_1)
plt.plot(espc_1, sens, marker="o", linestyle="--", color="r", label="Data, auc="+str(auc))
x = [i*0.01 for i in range(100)]
y = [i*0.01 for i in range(100)]
plt.plot(x,y)
plt.xlabel("Especifidad")
plt.ylabel("Sensibilidad")
plt.title("CURVA ROC")
plt.legend(loc=4)
```

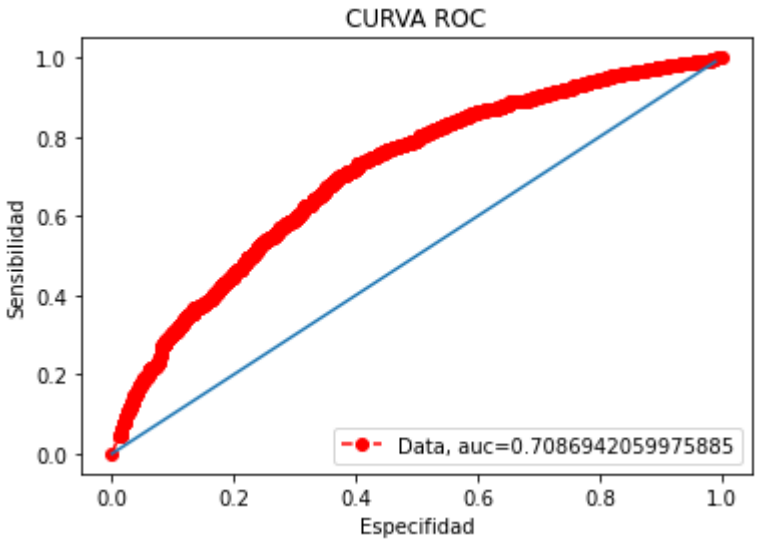


Figura 43: Curva ROC del modelo de random forest.

En la figura 43 se puede confirmar la afirmación inicial, que con el random forest se obtiene una mejor probabilidad (70.8%) el cual supera a la obtenida en el árbol de decisión.

4.1.9 EVALUACIÓN

Para esta etapa se comparará los resultados obtenidos de los 2 diferentes modelos predictivos, se tomará los valores arrojados por las herramientas de validación y se mostrarán en tablas para su posterior análisis:

Tabla 5
Resultados del modelo de regresión logística.

Métrica de Validación	Probabilidad
Validación cruzada	58.3%
Sensibilidad	99.6%
Especificidad	97.4%
Curva ROC	73.6%

Nota. Fuente: Elaboración propia.

Tabla 6
Resultados del modelo de k vecinos cercanos.

Métrica de Validación	Probabilidad
Validación cruzada	61.6%
Sensibilidad	97.6%
Especificidad	92.2%
Curva ROC	68.4%

Nota. Fuente: Elaboración propia.

Tabla 7
Resultados del modelo de máquina de vector de soporte.

Métrica de Validación	Probabilidad
Validación cruzada	62.0%
Sensibilidad	99.8%
Especificidad	99.8%
Curva ROC	74.3%

Nota. Fuente: Elaboración propia.

Tabla 8
Resultados del modelo de árbol de decisión.

Métrica de Validación	Probabilidad
Validación cruzada	60.5%
Sensibilidad	84.1%
Especificidad	63.1%
Curva ROC	67.6%

Nota. Fuente: Elaboración propia.

Tabla 9
Resultados del modelo de random forest.

Métrica de Validación	Probabilidad
Validación cruzada	61.3%
Sensibilidad	95.9%
Especificidad	85.3%
Curva ROC	70.5%

Nota. Fuente: Elaboración propia.

Analizando las tablas se puede indicar:

- Todos rondan entre el 60% y 75% de probabilidad de predicción.
- Se observa que el algoritmo de árbol de decisión es el que obtuvo los resultados más bajos, indicándonos que para esta investigación no es el mejor algoritmo a usar.
- En todas las mediciones mostradas, es el algoritmo de máquina de soporte de vectores el que obtiene las mediciones más altas, lo que nos indicaría que es el mejor algoritmo a emplear, pero, la gran desventaja que tiene este algoritmo es que el proceso de entrenamiento se hace muy largo cuando los registros de la base de datos pasan los 1000, cosa que no cumple la base de datos que maneja la Escuela Profesional de Ingeniería de Sistemas, pues sus registros son extensos, concluyendo que por temas de rendimiento, este algoritmo no será usado en este modelo.
- Como opción de reemplazo se tomará al algoritmo de regresión logística, pues es el segundo en tener las mediciones más altas y el tema de rendimiento no es problema pues es más veloz en el análisis de datos.

4.1.10 IMPLEMENTACIÓN

Primero se hablará acerca del concepto de persistencia de modelo, significa que un modelo se debe guardar con el algoritmo entrenado para después usarlo en otros archivos y/o otros scripts de Python; esto es muy importante para no tener que entrenar al algoritmo desde cero sino simplemente guardarlo, llamarlo y usarlo en el futuro.

Este mismo concepto puede usarse en la implementación, pues se lleva el algoritmo entrenado para que puedan usarlo a su criterio, se debe recalcar y reincidir que lo que se guarda es el algoritmo no el modelo, por tanto, es responsabilidad del destinatario contar con las tecnologías para continuar usando dicho modelo, en el caso de la

universidad no existe problema pues la Oficina General de Informática y Sistemas cuenta con un área especializada de desarrollo.

Primero guardamos el algoritmo seleccionado que viene siendo el de regresión logística, se realiza con el siguiente código:

```
from sklearn.externals import joblib

joblib.dump(model_logist, 'modelo_entrenado.pkl')
```

Esto nos guarda el algoritmo en el archivo modelo_entrenado.pkl el cual se ubica en la misma carpeta donde está el script. Ahora, para usar este archivo, se debe abrir un nuevo script del Jupyter notebbok y cargar el algoritmo entrenado, se realiza escribiendo el siguiente código:

```
from sklearn.externals import joblib

model_logist_trained = joblib.load('modelo_entrenado.pkl')
```

Con esto ya se tiene el algoritmo entrenado con los datos del dataframe limpiado, es decir que ahora el algoritmo puede seguir siendo entrenado con nuevos datos y mejorar su precisión.

4.1.11 RETROALIMENTACIÓN

Estas son algunas de las maneras de retroalimentación:

- Como el algoritmo fue entrenado con datos hasta el semestre 2019-I, eso significa que puede continuar mejorando con los datos de los siguientes semestres.
- Analizar y comparar con otros modelos que tienen distinto tipo de algoritmo, para corroborar si nuestras variables predictoras fueron correctas.
- Mejorar el dataframe, ya sea aumentando otras variables o quitándolos.
- Analizar si otras variables tendrían un mejor efecto al momento de predecir el rendimiento académico.

4.2 DISCUSIÓN

Antes de iniciar la discusión, se tiene que aclarar que, para esta investigación, el concepto de rendimiento académico se está limitando estrictamente a la nota final que

obtuvo el estudiante en un determinado curso, e inclusive se generalizó a sólo dos opciones que puede tenerse (aprobado o desaprobado), es decir que las demás dimensiones que influyen y definen el concepto de rendimiento académico no serán tomados en cuenta pues no se cuenta con una información histórica, además que analizar el impacto que cada uno de esta dimensiones ejercen sobre el rendimiento académico requeriría una investigación más profunda.

Al indicar que la nota final será la que determine el rendimiento académico, se debe entender que las variables a las cuales está sujeto esta calificación también entran al análisis, los cuales fueron tomados como variables predictoras.

La presente investigación es legítima, pues los datos empleados fueron extraídos de la base de datos del sistema de matrícula (SIMA), el cual es el encargado de la recopilación de registros de todos los movimientos de matrícula en cada semestre académico, que incluye los tipos de matrícula que se pueden realizar, los cursos, el docente y las notas; por tanto, es la base de datos fiable para poder diseñar un modelo predictivo del rendimiento académico.

Las limitaciones que esta investigación presenta y que deben ser consideradas para futuros estudios son:

- Tener en cuenta que las variables cualitativas que se obtuvo de la data, tienen una limitada cantidad de valores, por ejemplo, en la variable colegio de egreso se tiene 105 colegios, lo que significa que, si se quiere predecir con un nuevo colegio que no está dentro de esa lista, el modelo empezará a crear un patrón con ese nuevo colegio, lo que implica que sus primeras predicciones no serán confiables, lo recomendable es entrenar al modelo con información que incluya al nuevo colegio.
- Como consecuencia de lo anterior se debe indicar que, para entrenar al modelo con nuevas variables predictoras, lo más recomendable es iniciar desde cero pues no sería lo más conveniente, además que se podrán analizar los resultados de ambos modelos.
- Los nuevos datos con los cuales entrenar al modelo, deben ser lo más confiable posible, pues la data inicial con la cual se trabajó tenía errores de duplicación de datos y de registros vacíos; y al no tratarse de variables numéricas no se puede aplicar las técnicas de llenado que tiene machine learning, ya que no puedes hacer un promedio de docentes ni de colegio como para llenar los registros vacíos. Para esta investigación ese fue un problema

que definitivamente influyó en los resultados, ya que además de los problemas anteriores se notó que existen registros que se duplican y solo se difieren en el nombre del docente lo que implica que el curso fue dictado por 2 docentes y el alumno tiene un registro duplicado por esta causa; finalmente se tiene que indicar que existen muchos registros con nota de cero, lo que indica el abandono del alumno o que el proceso de desmatrícula no eliminó ese registro.

Este modelo predictivo, puede seguir siendo entrenado a través del tiempo, ya que en cada semestre académico se obtendrán datos para las variables predictoras, es decir que el aprendizaje inicial de un determinado curso va a ser reforzado cuando se vuelva a entrenar con los datos nuevos, los cuales van a variar ya sea en el colegio de egreso del estudiante y del docente que lo dicta, mejorando así la eficiencia del modelo. Esto, también implica, que el modelo puede ser empleado para predecir el rendimiento académico de otras escuelas, ya que esa fue una de las razones de tomar esas variables predictoras, que tenga la capacidad de ser registrados por cualquier escuela profesional.

Para finalizar, se harán las comparaciones con las otras investigaciones presentadas:

- En la investigación de Murat Pojon se centró en la comparación de dos modelos ya creados con dos grupos de prueba, en la nuestra se centró en diseñar un nuevo modelo a partir de una base de datos histórica.
- En la investigación de Lizares Castillo se realizó las comparaciones de dos tipos de algoritmos: regresión logística y árbol de decisión, además que dentro de las variables predictoras se encuentran variables propias de las condiciones socioeconómicas, lo que difiere con esta investigación, ya que no se tiene ese tipo de datos y se centró netamente en datos académicos, por tanto, el resultado también es diferente, en su investigación el algoritmo con mejor precisión es el árbol de decisión.
- Con la investigación de Zevallos Salazar, la diferencia radica en el uso del algoritmo (empleó redes neuronales) pero más importante es el porcentaje de muestra que él uso para su validez o testeo sólo fue del 10%, en cambio esta investigación empleo el 20%, lo que implica que nuestro modelo predictivo tenía más muestra para ser testeado.
- Con la investigación de Menacho Chiok la diferencia radica en que su modelo se centra en predecir del rendimiento académico en los estudiantes de un solo curso, en cambio esta investigación no.

- En la investigación de Camborda Zamudio, se centró en predecir el rendimiento académico de los estudiantes de los primeros ciclos de la carrera de Ingeniería Civil, lo que se parece mucho a esta investigación, pero difiere solamente en el uso del algoritmo de aprendizaje pues sólo uso el árbol de decisión, pero los datos que obtuvo eran más variados, pues contaba con datos demográficos, académicos, institucionales y actitudinales, lo que nos indica que su investigación estaba mejor nutrida.

CAPÍTULO V
CONCLUSIONES Y RECOMENDACIONES

5.1 CONCLUSIONES

a. Teniendo en cuenta los conceptos presentados en el capítulo II, donde se explica qué son las técnicas predictivas, y, junto con el trabajo desarrollado en el capítulo IV; se determinó que la técnica de análisis predictivo más adecuada es la llamada técnica de aprendizaje supervisado, porque nuestro modelo es de clasificación, debido a que la predicción a realizar nos brinda un resultado binario, en nuestro caso el rendimiento académico, pudiendo tomar sólo los valores de 1 como aprobado y 0 como desaprobado.

b. Teniendo en cuenta los conceptos presentados en el capítulo II, donde se explica cuáles son los algoritmos de predicción que se van a usar, y, junto con el trabajo desarrollado en el capítulo IV; se determinó que el algoritmo predictivo más eficiente para el rendimiento académico es la regresión logística pues obtuvo un probabilidad más alta (73.6%), se debe aclarar que el éxito de este modelo sobre los demás está relacionado a los tipos de datos seleccionados como variables predictoras, por tanto para esta investigación el mejor algoritmo es el de regresión logística.

c. Teniendo en cuenta los conceptos presentados en el capítulo II, donde se explica cuáles son las métricas de validación que se van a usar en el modelo, y, junto con el trabajo desarrollado en el capítulo IV; se determinó que la métrica de la curva ROC es la más recomendable para usar en la validación de este modelo predictivo, porque nos muestra el porcentaje de probabilidad, una gráfica simple y fácil de comprender, emplea los datos críticos como sensibilidad y especificidad; además de usar los datos en su totalidad.

d. Finalmente, como se obtuvo un resultado exitoso en el logro de los objetivos específicos, se alcanzó a realizar el objetivo principal que consistía en diseñar un modelo predictivo basado en machine learning, a través del uso de la

técnica predictiva llamada de aprendizaje supervisado, además aplicamos el algoritmo de regresión logística que nos brindó una probabilidad predictiva del 73.6% según la métrica de validación de la curva ROC, dando como resultado un modelo que predice el rendimiento académico de los alumnos de la Escuela Profesional de Ingeniería de Sistemas, el cual contribuirá aportando información valiosa sobre la cantidad de aprobados que el modelo pueda predecir para que se puedan tomar las decisiones adecuadas y a tiempo evitando el incremento de la población estudiantil o establecer mejores estrategias para su manejo; además que el modelo usará los datos que permanecen almacenados sin generar alguna información útil,

5.2 RECOMENDACIONES

1. Se recomienda analizar y contrastar los resultados que ofrecen los demás tipos de algoritmos predictivos, pues en esta investigación sólo se usaron cinco.
2. Se recomienda analizar nuevas variables predictoras y entrenar con ellas un algoritmo predictivo con el fin de hallar un modelo más eficiente que tenga una mejor probabilidad de predicción.
3. Se recomienda aprovechar toda la información guardada por el SIMA-UNSCH, ya que son demasiados y deben usarse para generar beneficios, pueden explorar los datos, hacer un tratamiento, analizar patrones de comunes y determinar su influencia realizando diferentes estudios.
4. Se recomienda mejorar la calidad de los datos que se guardan en el SIMA-UNSCH pues se observó que el problema más notorio es la duplicidad de registros; las futuras investigaciones pueden centrarse en el tratamiento de los datos o creación de aplicativos que mejoren la data general para futuros trabajos de data science.
5. Se recomienda recopilar información adicional de los estudiantes para realizar nuevas investigaciones, como por ejemplo tener detalles de la condición socioeconómica del estudiante al inicio de cada semestre para estudiar su influencia en un modelo predictivo.

REFERENCIAS BIBLIOGRÁFICAS

Álvaro, T. (1990). *Hacia un modelo causal del rendimiento académico*. Madrid: Centro De Publicaciones.

Arias, F. (2012). *El proyecto de investigación: Introducción a la metodología científica*. Caracas, Venezuela: Editorial Episteme.

Barrero, G. (2019). *Evaluación de la eficiencia de los modelos machine learning para la predicción de la calidad del software desarrollado en ibm rpg usando la matriz de confusión y las curvas roc*. Lima, Perú: Universidad César Vallejo.

Barrios, J. (2019). *La matriz de confusión y sus métricas*. Recuperado de https://www.cienciadatos.com/la-matriz-de-confusion-y-sus-metricas.

Bavaresco, A. (2006). *Proceso Metodológico en la Investigación. (Cómo hacer un diseño de investigación).* Maracaibo: La Universidad del Zulia.

Bernal, C. (2010). *Metodología de la investigación administración, economía, humanidades y ciencias sociales. (3° Ed.),* Bogotá, Colombia: Pearson Educación.

Boschetti, A. y Massaron, L. (2016). *Python Data Science Essentials (2ª ed.).* Reino Unido: PACKT.

Brownlee, J. (2004). *Machine Learning Resource Guide*. Estados Unidos: C&D.

Camilo, C., Silva, J., El-jaick, D., Hendrickx, T., Cule, B., Meysman, P., Oliveira, F. (2015). *From Data Mining to Knowledge Discovery in Databases. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9078(3), 637–648. https://doi.org/10.1007/978-3-319-18032-8_50

Castillo, N. (2015). *Técnicas de Machine Learning para el Post-Proceso de la predicción de la Irradiancia (Tesis de maestría).* Universidad de Granada, España.

Chávez, N. (2007). *Introducción a la Investigación Educativa*. Maracaibo.

Craing, I. (2002). *The Interpretation of object-oriented programming languages (2ª Ed.).* Gran Bretaña: Springer-Verlag London.

Churpek, M. (2018). *Big Data and Data Science In Critical Care*. Researchgate.

Cios, K., Pedrycz, W.; Swiniarski, R. (2007). *Data Mining a Knowledge Discovery Approach. Springer Science+Business* Media 2007, LLC, pp: 606, ISBN: 978-0-387-33333-5

Córdova, M. (2006). *Estadística Inferencial (2ª ed.).* Lima. Perú: Librería MOSHERA S.R.L.

DAMA. (2017). *DAMA-DMBOK: Data Management Body of Knowledge* (2da Ed.). Bradly Beach, Nueva Jersey: Technics Publications.

Deisenroth, M., Faisal, A., Ong, C. (2020). *Mathematics for Machine Learning*. Inglaterra: Cambridge University Press.

Fayyad, U., Piatestsky-Shapiro, G., Smith, P. (1996). *From Data Mining to Knowledge Discovery: An Overview. Advances in Knowledge Discovery and Data Mining*, pp:1-34, AAAI/MIT.

García, J. y Molina, J. (2012). *Técnicas de análisis de datos.* Recuperado de http://ocw.uc3m.es/ingenieria-informatica/analisis-de-datos

Gareth, J., Witten, D. y Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Los Angeles. USA: Editorial SPRINGER.

Gómez, A. Gualo, F. y Caballero, I. (2019). *Calidad de datos*. Ediciones de la U.

Hernández, A. (2015). *Diagnóstico del rendimiento académico de estudiantes de una escuela de educación superior en Mexico*. Revista Complutence de Educación, Obtenido de https://revistas.ucm.es/index.php/RCED/article/view/48551

Hernandez, R., Fernández, C., y Baptista, P. (2014). *Metodología de la investigación. (6° Ed.),* Distrito Federal, México: Editorial McGraw-Hill.

Huddleston, S., y Brown, G. (2018). *Machine learning. In Informs Analytics Body of Knowledge (1st ed., Vol. 38*). https://doi.org/10.1002/9781119505914.ch7.
Hurwitz, J. y Kirsch, D. (2018). *Machine Learning for Dummies*. Estados Unidos: IBM Limited Edition.
IBM (2015). *Metodología Fundamental para la Ciencia de Datos*. New York, Estados Unidos: IBM Analitics.
Igual, L., y Seguí, S. (2017). *Introduction to Data Science*. Barcelona, Spain: Springer.
Juan B. (2020). *Aprende Machine Learning en español*. Leanpub.
Kumar, R. (2017). *Machine Learning and Cognition in Enterprises.* India: TODD GREEN.
Lizares, M. (2017). *Comparación de modelos de clasificación: regresión logística y árboles de clasificación para evaluar el rendimiento académico*. Universidad Mayor de San Marcos.
Liu, A. (2015). *Data Science and Data Scientist*. IBM.
Lun, W. (2011). *Machine Learning Tutorial*. Taiwan, China: National Taiwan University.
Marsland, S. (2015). *Machine Learning An Algorithmic Perspective (2ª ed.).* Estados Unidos: Taylor & Francis Group.
Massaron, L., & Boschetti, A. (2018). *Python Data Science Essentials (Third Edition ed.).* Packt Publishing. Recuperado el 23 de octubre de 2018, de https://www.safaribooksonline.com/library/view/python-data-science/9781789537864/
Mathews, A. (2019). *What can machine learning do for information security? Network Security*, 2019, 15–17. https://doi.org/10.1016/S1353-4858(19)30050-9
McGilvray, D. (2008). *Executing Data Quality Project: Ten Steps to Quality Data and Trusted Information*. Burlington, MA, USA, Morgan Kaufmann.
Miroslav, K. (2017). *An Introduction to Machine Learning.* Miami Estados Unidos: Springer.
Moreno, A., Armengol, E., Béjar, J., Belanche, L., Córtes, U., Sanchez, M. (1994). *Aprendizaje automático*. Barcelona, España: Universidad Politécnica de Catalunya.
MS. (2018). *Machine Learning una pieza clave en la transformacion de los Modelos de Negocios*. España: MANAGEMENT SOLUTIONS.
Müller, A. y Guido, S. (2017). *Introduction to Machine Learning with Python*. Boston: OREILLY.
Orallo J, Ramírez C. (s. f.) *Minería de Datos y Extracción de Conocimiento de Bases de Datos*., De: http://users.dsic.upv.es/~jorallo/docent/doctorat/t2a.pdf
Osorio, F. (s.f.). *Lógica de programación orientada a objetos: un inicio al desarrollo de software*. Madrid, España: ITM.
Ozdemir, S. (2016). *Principles of Data Science*. Packt Publishing. Recuperado el 23 de octubre de 2018, de https://www.safaribooksonline.com/library/view/principles-of-data/9781785887918
Pojon, M. (2017). *Using Machine Learning to predict student performance.* University of Tampere.
Qin, S. J., y Chiang, L. H. (2019). *Advances and opportunities in machine learning for process data analytics*. Computers and Chemical Engineering, 126, 465–473. https://doi.org/10.1016/j.compchemeng.2019.04.003
Raschka, S., Mirjalili, V. (2017). *Python Machine Learning*. Inglaterra: Packt Publishing.
Roman, V. (2019). *Machine Learning: Cómo Desarrollar un Modelo desde Cero*. Recuperado de https://medium.com/datos-y-ciencia/machine-learning-c%C3%B3mo-desarrollar-un-modelo-desde-cero-cc17654f0d48.
Romero, F. (2019). *¿Por qué necesitamos una metodología para la ciencia de datos?* Recuperado de https://fyaromo.com.co/2019/03/03/por-que-necesitamos-una-metodologia-para-la-ciencia-de-datos/.
Salinas, P. (2010). *Metodología de la investigación científica.* Mérida, Venezuela: Universidad de los Andes.

Samuel, A. L. (1959). *Some Studies in Machine Learning Using the Game of Checkers.* IBM Journal of Research and Development. http://doi.org/10.1147/rd.33.0210

Soysal, M., y Guran, E. (2010). *Machine learning algorithms for accurate flow-based network traffic classification Evaluation and comparison*. Performance Evaluation, 67(6), 451–467. https://doi.org/10.1016/j.peva.2010.01.001

Supo, J. (2014). *Seminario de Investigación Científica: Metodología de la Investigación para las Ciencias de la Salud*. Arequipa, Perú: Bioestadístico EIRL.

Tamayo, M. y Tamayo, A. (1997). *El proceso de la Investigación Científica*. *(Tercera Edición).* México: LIMUSA

Turing, A. M. (1950). *Computing Machinery and Intelligence*. Mind, 49, 433–460.

Weitzenfeld, A. (s.f.). *Ingeniería de software orientada a objetos: con UML, Java e internet.* Madrid, España: Thomson.

Zevallos, R. (2017). *Predicción del rendimiento académico mediante redes neuronales*. Universidad Nacional del Callao.

LISTA DE ABREVIATURAS

UNSCH: Universidad Nacional San Cristóbal de Huamanga.
EPIS: Escuela Profesional de Ingeniería de Sistemas.
SIMA: Sistema de matrícula
Jupyter: Julia – Python – R.
Numpy: Numerical Python.
POSMAD: Planificar – obtener – almacenar y compartir – mantener – aplicar – eliminar.
ML: Machine Learning.
CSV: Comma-separated values (valores separados por comas).
XLSX: Extensión de los archivos EXCEL.
PKL: Extensión Python de un algoritmo entrenado.
POO: Programación orientada a objetos.
GUI: Graphical user interface
IBM: International Business Machines
ROC: Receiver - Operating – Characteristic.
VP: Verdaderos positivos
VN: Verdaderos negativos.
FP: Falsos positivos.
FN: Falsos negativos.

GLOSARIO

Modelo:
En machine learning un conjunto de instrucciones y pasos que se siguen para obtener un resultado establecido, dentro del modelo se ubica el algoritmo de aprendizaje, y es el cual le da el nombre al modelo.

DUMMY:
En machine learning, es una característica que se da a un dato ficticio, un dato que se crea intencionalmente para ayudar al análisis de los datos.

Dataset:
Es un conjunto de datos que se encuentran contenidos en una sola tabla o en una matriz, donde la ubicación es importante, pues cada columna es un dato independiente.

Dataframe:
Es una estructura especial, producto del uso de la librería pandas de Python, es similar a una matriz con la diferencia que es indizada; es el modelo que los datos deben tener para poder ser tratados y/o analizados mediante técnicas e instrumentos propios de la ciencia de datos.

Indizar:
Registrar ordenadamente la información, siguiendo un determinado criterio con la finalidad de personalizar al dato para que sea único y fácil de ubicar.

Predicción:
Declaración exacta de lo que va a suceder en un determinado tiempo, o anticipar el resultado de alguna acción o evento.

ANEXO A

MATRIZ DE OPERACIONALIZACIÓN DE VARIABLES

VARIABLE	VARIABLES DESCRIPTIVAS	INDICADORES	ITEMS	INSTRUMENTO
Modelo de machine learning	Técnica predictiva	Estandarización de los datos	¿Cuál es el procedimiento de estandarizar y normalizar los datos para entrenar un modelo predictivo?	Ficha bibliográfica
			¿Qué herramientas usar para la limpieza y preparación de los datos de manera que entrene un modelo predictivo?	Ficha bibliográfica
		Técnicas de análisis de datos	¿Cuáles son las técnicas de análisis de datos para crear patrones predictivos?	Ficha bibliográfica
			Qué técnicas y herramientas se usa para analizar grandes volúmenes de datos	Ficha bibliográfica
	Algoritmo predictivo	Algoritmos supervisados	¿Qué tipo de algoritmos supervisados se usan para crear un modelo de machine learning?	Ficha bibliográfica
			¿Cómo elegir un algoritmo de regresión o de clasificación para predecir el rendimiento académico?	Ficha bibliográfica
		Métricas de los algoritmos	¿Cuáles son las métricas para medir la probabilidad de acierto de un modelo predictivo?	Ficha bibliográfica
			¿Qué herramientas usar para medir el rendimiento de un algoritmo predictivo?	Ficha bibliográfica
	Métrica de validez	Fidelidad	¿Cómo evaluar la fidelidad de un modelo para garantizar el cumplimiento de los objetivos?	Ficha bibliográfica
			¿De qué manera evaluar el rendimiento para poner en marcha un modelo predictivo?	Ficha bibliográfica
		Precisión	¿Cuáles son las métricas para evaluar la precisión del modelo predictivo?	Ficha bibliográfica
			¿Qué criterio estadístico o prueba estadística usar para evaluar la precisión de un modelo predictivo?	Ficha bibliográfica

ANEXO B

FICHA DE REGISTRO

Código	Nombres	Género	Fecha de nacimiento	Lugar de nacimiento	Modalidad de ingreso	del estudiante	Colegio de egreso	Ubicación del colegio	Semestre académico	Sigla del curso	Curso	Plan de estudios	Créditos	Modalidad del curso	Docente	Notas	Promedio Final

ANEXO C
FICHA BIBLIOGRÁFICA

FICHA BIBLIOGRÁFICA	
Título:	
Autor:	
N.º Edición:	Año:
Editorial:	Ciudad, País:
Resumen:	

ANEXO D
SOLICITUD DE OBTENCIÓN DE INFORMACIÓN PARA LA OFICINA GENERAL DE INFORMÁTICA Y SISTEMAS

SOLICITO: INFORMACIÓN DE ESTUDIANTES DE LA EPIS PARA ELABORACIÓN DE TRABAJO DE TESIS

SEÑOR JEFE DE LA OFICINA GENERAL DE INFORMÁTICA Y SISTEMAS

Yo, **ARONÉS AYALA, Ever**; identificado con DNI 44106805, con domicilio en Jr. Parinacochas 471. Sta. Elena, distrito de Andrés Avelino Cáceres Dorregaray; alumno egresado de la Universidad Nacional de San Cristóbal de Huamanga de la Escuela Profesional de Ingeniería de Sistemas con código universitario **N° 27061107** y correo institucional **ever.arones.27@unsch.edu.pe**; ante usted con el debido respeto me presento y expongo:

Que, para la elaboración de mi tesis "PREDICCIÓN DEL RENDIMIENTO ACADÉMICO BASADO EN MACHINE LEARNING, ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS, AYACUCHO 2021" y teniendo como asesor designado al Dr. Ing. Manuel Avelino LAGOS BARZOLA; es necesario contar con la **INFORMACIÓN DE LOS ESTUDIANTES DE LA ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS CORRESPONDIENTES DESDE EL PERÍODO ACADÉMICO 2015-I HASTA EL 2019-I**, con los siguientes datos:

- Código del estudiante
- Nombres y apellidos
- Género
- Fecha de nacimiento
- Lugar de nacimiento
- Modalidad de ingreso
- Condición del estudiante (regular, irregular, etc.)

- Nombre del colegio de egreso
- Ubicación del colegio de egreso
- Semestre académico
- Código del curso
- Nombre del curso
- Plan de estudios
- Créditos del curso
- Modalidad del curso (regular, vacacional, curso único, etc.)
- Docente del curso
- Notas
- Promedio final del curso

Solicito a usted que tenga la bondad de asignar a quien corresponda se me otorgue la información solicitada, adjunto el archivo con la resolución de aprobación del plan de tesis.

POR TANTO:

Ruego a usted acceder a mi petición por ser de necesidad.

Ayacucho, 12 de noviembre de 2021

Aronés Ayala, Ever

Celular: 900173346