UNIVERSIDAD NACIONAL HERMILIO VALDIZÁN

ESCUELA DE POSGRADO

INGENIERÍA DE SISTEMAS, MENCIÓN EN TECNOLOGÍA DE INFORMACIÓN Y COMUNICACIÓN

"RENDIMIENTO ACADÉMICO MEDIANTE TÉCNICAS DE MINERÍA DE DATOS EN ESTUDIANTES DE LA UNIVERSIDAD NACIONAL AGRARIA DE LA SELVA"

LÍNEA DE INVESTIGACIÓN: INGENIERÍA DE SISTEMAS
SUBLÍNEA DE INVESTIGACION: SISTEMAS DE AUTOMATIZACIÓN Y CONTROL DE PROCESOS

TESIS PARA OPTAR EL GRADO DE MAESTRO EN INGENIERIA DE SISTEMAS, MENCIÓN EN TECNOLOGIAS DE LA INFORMACIÓN Y COMUNICACIÓN

TESISTA: PONCE GUIZABALO SANTOS VICTOR

ASESOR(A): Dr. PASQUEL CAJAS ALEXANDER FRANK

HUÁNUCO – PERU

2023

DEDICATORIA

Dedicado a Dios por ser mi guía y protector, a mi hijo que en paz descanse en la gloria de Dios, a mis padres que tengan muchos años más de vida.

AGRADECIMIENTO

A todos los que brindaron aporte para el desarrollo de la presente investigación, a Dios por la vida y la salud, a las personas que están mi lado.

RESUMEN

El presente trabajo de investigación tiene como objetivo determinar la incidencia de la minería de datos en el rendimiento académico de los estudiantes del primer ciclo de la Universidad Nacional agraria de la selva. Para ello se planteó la hipótesis la minería de datos incide en el rendimiento académico de los estudiantes del primer ciclo de la Universidad Nacional Agraria de la Selva, Tingo María 2021. La técnica que se usó para recolectar los datos es una ficha de análisis documental, inicialmente se contaba con 2900 registros, luego con la depuración nos quedamos con 2404 registros, con el software WEKA 3.9 se analizó el resultado de los algoritmos de aprendizaje automático resultando con mejor predicción regresión logística de 72.79 % de exactitud (acuracy). En conclusión, El rendimiento académico de los estudiantes es un tema muy complejo y con los indicadores de ingreso como económicos, sociales o académicos, a través de la aplicación de técnicas de minería de datos y la metodología CRISP-DM, usando la aplicación de diferentes técnicas de minería de datos se logró determinar que el algoritmo que mejor incidencia tiene en el rendimiento académico de los alumnos ingresantes es la regresión logística que llega a una exactitud (acuracy) de 72.79 %.

Palabras claves: *minería de datos, rendimiento académico, regresión logística, algoritmos de aprendizaje automático, metodología CRISP-DM.*

ABSTRACT

The objective of this research work is to determine the incidence of data mining in the academic performance of the students of the first cycle of the National Agrarian University of the jungle. For this, the hypothesis that data mining affects the academic performance of the students of the first cycle of the National Agrarian University of La Selva, Tingo María 2021, was raised. The technique used to collect the data is a documentary analysis file, Initially there were 2900 records, then with the debugging we were left with 2404 records, with the WEKA 3.9 software the result of the automatic learning algorithms was analyzed, resulting in the best logistic regression prediction of 72.79% accuracy (acuracy). In conclusion, the academic performance of students is a very complex issue and with income indicators such as economic, social or academic, through the application of data mining techniques and the CRISP-DM methodology, using the application of different Using data mining techniques, it was possible to determine that the algorithm that has the best impact on the academic performance of incoming students is logistic regression, which reaches an accuracy of 72.79%.

Keywords: data mining, academic performance, logistic regression, machine learning algorithms, CRISP-DM methodology.

ÍNDICE

INTRODUCCIÓN

Actualmente las grandes cantidades de datos que se almacenan en el mundo quedan sin ser analizadas o procesadas, con la cual se pierde información valiosa. Las universidades también recolectan la información de los alumnos que postulan y muchos de estos datos tiene información oculta que al ser tratada nos puede brindar patrones que nos permita por ejemplo determinar el rendimiento de los alumnos durante el primer ciclo, esta información se puede procesar y usar herramientas de minería de datos para predicción y así encontrar indicadores sociales, económicos y académicos asociados al rendimiento académico de los estudiantes del primer ciclo, esto permite a los directivos de la universidad tomar decisiones acertadas para mejorar la calidad de la educación dentro de la universidad.

En el presente trabajo de investigación el objetivo está orientado a determinar la incidencia de la minería de datos en el rendimiento académico de los estudiantes del primer ciclo de la Universidad Nacional agraria de la selva. La presente investigación se ha divido en 5 capítulos.

CAPÍTULO I. Se presenta los aspectos básicos de la investigación dentro del cual está la fundamentación del problema, la justificación e importancia, así como la viabilidad, tenemos la formulación del problema, los objetivos

CAPÍTULO II. Tenemos las hipótesis, la operacionalización de las variables y la definición operacional de las variables.

CAPÍTULO III. Aquí se presenta el marco teórico de la tesis con sus antecedentes similares tanto internacionales como nacionales, con sus respectivas bases teóricas y conceptuales teniendo en cuenta las variables de estudio.

CAPÍTULO IV. Tenemos el marco metodológico, la población y la muestra para el estudio, el tipo y nivel de investigación, las técnicas y los instrumentos utilizados para la recolección y procesamiento de los datos.

CAPÍTULO V. Mostramos los resultados, la contrastación de las hipótesis, discusión de los resultados. Para luego sacar nuestras conclusiones y sugerencias, también tenemos las referencias bibliográficas y los anexos.

CAPÍTULO I. ASPECTOS BÁSICOS DEL PROBLEMA DE INVESTIGACIÓN

1.1. Fundamentación del problema

De acuerdo con Díaz et al. (2002) sostienen que, “para abordar la cuestión de la calidad de la enseñanza superior, el rendimiento académico de los estudiantes universitarios es crucial, ya que sirve de indicador que permite aproximarse al panorama educativo” (pp. 357-383).

Dentro de las universidades se tiene en cuenta que, el rendimiento académico inadecuado es un problema frecuente que enfrentan tanto los estudiantes como los educadores en todos los niveles escolares. La trascendencia de este tema tanto para los individuos como para la sociedad se evidencia a través de dos factores clave : en primer lugar, cuando el bajo rendimiento académico dificulta la capacidad de los estudiantes para alcanzar sus objetivos profesionales ; y en segundo lugar, cuando el alcance de sus conocimientos y habilidades se restringe a los requisitos de su profesión específica ; y en segundo lugar, cuando el alcance de sus conocimientos y habilidades se restringe a los requisitos de profesión específica. (García et al., 2014, p. 272).

Las universidades no son ajenas a esta realidad, el rendimiento académico dé los ingresantes afecta de alguna forma en las universidades, por ende, se tiene deficiencia académica, y es muy importante conocer que tipos de alumnos ingresaron y poder conocer cuál será su rendimiento dentro de la universidad para tomar decisiones que permitan mejorar el rendimiento de estos.

Muchos estudios intentan hacer predicciones sobre los resultados del aprendizaje de los estudiantes mediante el uso de técnicas de minería de datos. Esta es una tarea difícil ya que los estudios han demostrado que muchos factores, tanto personales como socioeconómicos, psicológicos y de otro tipo, afectan el rendimiento escolar. La predicción precisa puede detectar de antemano qué estudiantes tendrán dificultades para aprobar la lección, tomar las decisiones correctas y brindar apoyo adicional en forma de cambios o ajustes editados por los maestros, entre otras cosas (Yamao, 2018).

En la Universidad Nacional Agraria de la Selva los alumnos ingresantes se tocan con una realidad distinta al colegio, y muchos de ellos tienen un rendimiento académico bajo durante el primer ciclo. Si no se conoce quienes son los alumnos que tendrán un rendimiento académico bajo, hace posible que los alumnos terminan por abandonar los estudios universitarios.

1.2. Justificación e importancia de la investigación

En los últimos tiempos la tecnología juega un papel muy importante en las actividades humanas, las técnicas de data mining nos permiten agrupar datos y encontrar patrones en grandes cantidades de datos, categorizar y tener tendencias para poder predecir, ya sea en el negocio, así como en el ámbito académico.

La Universidad Nacional Agraria de la Selva no es ajena a la problemática debido a la gran cantidad de datos que se generan diariamente en las instituciones públicas y privadas de educación superior de nuestro país. Los datos de los estudiantes que postulan y son admitidos en la universidad se generan y almacenan continuamente en el departamento de asuntos académicos de esta escuela secundaria. Aunque estos datos no se utilizan para explorar el conocimiento oculto en ellos, la importancia de este estudio radica en el hecho de que gracias a él podemos predecir resultados de aprendizaje e identificar perfiles de estudiantes con el nivel académico más bajo posible mediante técnicas de minería de datos.

Importancia de la investigación

Este trabajo de investigación quedará a disposición de la universidad, será de mucha utilidad para tomar decisiones en el tiempo y espacio correcto, ya que podemos tener la clasificación de los estudiantes que podrían tener un bajo rendimiento académico.

A nivel teórico, el presente estudio nos permite conocer cómo influyen los diferentes factores económicos, sociales y académicos, con respecto al rendimiento academico de los alumnos. El contribuyente hacia la teoría es de gran importancia, ya que se llevará a cabo una demonstración de direccionalidad entre las variables mediante el aporte de los datos almacenados.

A nivel práctico, los resultados del presente estudio permitirán identificar factores que influyen en el rendimiento académico de los alumnos ingresantes. Determinar las relaciones existentes entre las variables mencionadas, contribuye de manera fundamental en el diseño de estrategias que permiten el mejoramiento de los índices de bajo rendimiento académico, además que existe abandono y muchos prolongan de formas excesiva la duración de su carrera.

1.3. Viabilidad de la investigación

La viabilidad del estudio es muy favorable por que se cuentan con recursos humanos, se tiene acceso a la base de datos de la información de los estudiantes de la universidad, es de bajo costo económico y su implementación es de corto tiempo y se trabajara con datos históricos.

1.4. Formulación del problema

1.4.1. Problema general

¿Cómo la minería de datos puede predecir el rendimiento académico de los estudiantes del primer ciclo de la Universidad Nacional agraria de la selva?

1.4.2. Problema especifico

1. ¿Cuáles son los indicadores sociales, económicos y académicos que tienen mayor incidencia para predecir el rendimiento académico en estudiantes universitarios de primer ciclo de la Universidad Nacional Agraria de la Selva?

2. ¿Qué algoritmos de aprendizaje automático de la minería de datos son capaces de predecir el rendimiento académico de los estudiantes del primer semestre de la Universidad Nacional Agraria de la Selva?

1.5. Formulación de objetivos

1.5.1. Objetivo general

Predecir con las técnicas de la minería de datos el rendimiento académico de los estudiantes del primer ciclo de la Universidad Nacional Agraria de la Selva.

1.5.2. Objetivos específicos

1. Determinar los indicadores académicas, sociales y económicas que más influyen en la predicción del rendimiento académico. de los estudiantes universitarios de primer ciclo de la Universidad Nacional Agraria de la Selva.
2. Determinar los algoritmos de aprendizaje automático de la minería de datos que pueden predecir el rendimiento académico en estudiantes universitarios de primer ciclo de la Universidad Nacional Agraria de la Selva.

CAPÍTULO II. SISTEMAS DE HIPÓTESIS

2.1 Formulación de la hipótesis

2.1.1 Hipótesis de investigación

HI: Con las técnicas de la minería de datos se puede predecir el rendimiento académico de los estudiantes del primer ciclo de la Universidad Nacional Agraria de la Selva, Tingo María 2021

H0: Con las técnicas de la minería de datos no se puede predecir el rendimiento académico de los estudiantes del primer ciclo de la Universidad Nacional Agraria de la Selva, Tingo María 2021.

2.1.2 Hipótesis especificas

1. HI: Los indicadores académicas y económicas son los que más influyen en la predicción del rendimiento académico. de los estudiantes universitarios de primer ciclo de la Universidad Nacional Agraria de la Selva.

   H0: Los indicadores académicas y económicas no son los que más influyen en la predicción del rendimiento académico. de los estudiantes universitarios de primer ciclo de la Universidad Nacional Agraria de la Selva.

2. HI: Los algoritmos de aprendizaje automático de la minería de datos pueden predecir el rendimiento académico de los estudiantes universitarios de primer ciclo de la Universidad Nacional Agraria de la Selva.

   H0: Los algoritmos de aprendizaje automático de la minería de datos no pueden predecir el rendimiento académico de los estudiantes universitarios de primer ciclo de la Universidad Nacional Agraria de la Selva.

2.2 Operacionalización de las variables

Variable Independiente:

Minería de datos

Variable Dependiente

Rendimiento Académico:

VARIABLES	DEFINICION CONCEPTUAL	DEFINICION OPERACIONAL	DIMENSIONES	INDICADORES
VARIABLE INDEPENDIENTE **MINERIA DE DATOS**	De acuerdo a (Microsoft, 2021), Encontrar información útil a partir de vastas colecciones de datos se denomina minería de datos. Permite inferir patrones y tendencias en los datos mediante análisis matemáticos.	Se avalúan los algoritmos de aprendizaje automático, y con la ayuda de la herramienta WEKA nos proporciona los porcentajes de exactitud de cada modelo, se crea el modelo con todos los datos recolectados en las oficinas de admisión mediante ficha de análisis documental.	Indicadores sociales Indicadores económicos Indicadores académicos	• Sexo • Edad • Provincia • Financiamiento de estudios • Tipo de colegio • Puntaje examen de ingreso • Facultad • Modalidad de ingreso • Colegio de procedencia
VARIABLE DEPENDIENTE **RENDIMINENTO ACADEMICO**	Según (González, 1989), El éxito académico de los estudiantes depende exclusivamente de la universidad a la que vayan, de sus profesores y, lo que es más importante, de su propia capacidad. Es una noción que se utiliza en todos los ámbitos educativos para referirse a la evaluación del conocimiento de los estudiantes expresado en los resultados de sus evaluaciones.	En la herramienta software WEKA se carga el modelo y los datos de prueba para predecir el rendimiento académico el cual nos arroja la exactitud de predicción con la condición de aprobado o desaprobado.	Predicción	• Aprobado • Desaprobado

CAPÍTULO III. MARCO TEÓRICO

3.1 Antecedentes de la investigación

ANTECENDENTES INTERNACIONALES

Según (Oñate, 2016) realizó un estudio titulado: "**Análisis de la Deserción y Permanencia Académica en la Educación Superior Aplicando Minería De Datos**." en la ciudad de Bogotá del país Colombia. El **objetivo** general del estudio fue crear un conjunto de datos de estudiantes que tenga en cuenta el estatus socioeconómico, la formación académica y el rendimiento en el SAT 11.. La **muestra** estuvo constituida por: el conjunto de datos tiene 1.665 registros con 37 características, incluidos 26 atributos numéricos y 11 de categoría. El **diseño** que se utilizó fue el Diseño Experimental. Los **instrumentos** que se usaron fueron: base de datos de los estudiantes matriculados durante los periodos académicos del 2010 al 2014. Y los **resultados** obtenidos han sido. La Tabla 5-14 compara varias técnicas de clasificación en términos de diversos parámetros de rendimiento, presentando los resultados del modelo de predicción de la pérdida de estatus académico utilizando la información del proceso de admisión y el historial académico del semestre anterior con los datos de entrenamiento y validación. Al examinar los resultados del conjunto de datos de entrenamiento y validación, se puede observar cómo el grado de precisión del árbol de decisión aumentó en la segunda y cuarta matrícula. Los registros con bloqueo académico que se categorizaron con éxito tuvieron un mayor grado de precisión gracias al clasificador bayesiano. Del mismo modo que ambos métodos tienen un rendimiento muy superior a 0,7 cuando se observa el área bajo la curva (AUC), también lo tienen en la segunda matriculación. Según las métricas de rendimiento, las distintas estrategias de clasificación se comparan en la Tabla 5-15, que muestra los resultados del modelo de predicción de pérdida de estado académico utilizando la información de entrada del proceso de admisión, el historial académico del semestre anterior y los datos de la prueba. Cuando se tienen en cuenta los resultados del análisis del conjunto de datos de prueba, se observa que el árbol de decisión tiene la mayor proporción de predicciones que tenían bloqueos académicos y se identificaron con precisión en la segunda matriculación. Analizando el área bajo la curva (AUC), el método Naive Bayes supera al enfoque del árbol de decisión con un área superior a 0,9.

Además, (Márquez Vera, 2015) realizó un estudio titulado: "PREDICCIÓN DEL FRACASO Y EL ABANDONO ESCOLAR MEDIANTE TÉCNICAS DE MINERÍA

DE DATOS" en la ciudad CÓRDOBA del país ARGENTINA. El **objetivo** general del estudio fue: encontrar un modelo que permita identificar a los alumnos con más probabilidades de tener dificultades académicas o de abandonar los estudios. La **muestra** estuvo constituida por los 670 alumnos matriculados en el Programa II de la UAPUAZ para el curso 2009-10. El **diseño** que se utilizó fue experimental. Los **instrumentos** que se usaron fueron encuesta a los alumnos Y los **resultados** obtenidos han sido: Como puede verse, hay ocho reglas del tipo SI-ENTONCES, cuatro de las cuales se refieren a las clases que suspendieron y cuatro a las clases que aprobaron. Los únicos factores que aparecen en las reglas para la clase SUSPENDIDA y que, en consecuencia, son culpables del suspenso de los alumnos son haber obtenido bajas calificaciones en las asignaturas del curso (Matemáticas 1, Informática 1, Inglés 1, Ciencias Sociales 1, Física 1, Taller de Lectura y Escritura 1 y Humanidades 1). Es intrigante observar que en las normas que identifican a los alumnos que aprobaron y continuarán en el semestre siguiente aparecen marcadores adicionales, como abstinencia de alcohol o consumo muy moderado de alcohol, buena asistencia a clase y altas expectativas de aprobar el semestre. De los valores de la medida de clasificación del modelo se desprende claramente que son extremadamente altos y muy cercanos al 100%, el valor máximo que se puede alcanzar. Sin embargo, como emplea datos recogidos al final del semestre, cuando hay menos tiempo para aplicar una intervención de apoyo a los alumnos en riesgo de suspender, este modelo de categorización no es útil para la predicción temprana..

Igualmente, Rico (2019) realizó un estudio titulado: "CONSTRUIR UN MODELO PARA PREDECIR EL RENDIMIENTO ACADÉMICO DE ESTUDIANTES UNIVERSITARIOS MEDIANTE EL ALGORITMO *NAÏVE BAYES*, OBTENER SU EXACTITUD EN LAS PREDICCIONES Y APLICARLO MEDIANTE LENGUAJES DE PROGRAMACIÓN PARA SU USO EN LA WEB EN LA CIUDAD DE ZAPOPAN DEL PAÍS DE MEXICO". El **objetivo** general del estudio es crear un modelo para predecir el éxito académico de los estudiantes universitarios utilizando el método Naive Bayes, determinar su precisión de predicción y emplearlo utilizando lenguajes de programación basados en web. La **muestra** estuvo constituida por 122 estudiantes en este estudio. El diseño del modelo predictivo se creó utilizando el método de descubrimiento de extracción de información de bases de datos. Los instructores del curso facilitaron la información sobre aprobados y suspensos de los alumnos que participaron en este estudio. El resto de la información se recogió mediante una encuesta. Los resultados fueron los

siguientes: De este modo, los 122 registros de alumnos se separaron en dos conjuntos de 61 registros cada uno, utilizándose un conjunto para el entrenamiento y el otro para las pruebas a fin de determinar la precisión y viceversa. La precisión predicha del modelo de predicción se mostró mediante la precisión media de estos porcentajes, que fue del 61,4754%. Para comprobar la exactitud de la predicción del modelo de predicción se utilizó el rendimiento académico previsto de 71 alumnos matriculados en el mismo curso pero que lo cursarían el semestre siguiente. En primer lugar, se obtuvieron las predicciones de los estudiantes utilizando el modelo predictivo desarrollado y la plataforma implementada. A continuación, se contrastaron estas expectativas con los resultados reales obtenidos por los alumnos al finalizar el curso. En la Figura 6 se muestra la proporción de predicciones acertadas e inexactas. Además, se comprobó que la precisión de las predicciones era del 70,4225%, superior a lo que había predicho el enfoque de validación cruzada.

ANTECEDENTES NACIONALES

Yamao (2018) realizó un estudio titulado: “PREDICCIÓN DEL RENDIMIENTO ACADÉMICO MEDIANTE MINERÍA DE DATOS EN ESTUDIANTES DEL PRIMER CICLO DE LA ESCUELA PROFESIONAL DE INGENIERÍA DE COMPUTACIÓN Y SISTEMAS, UNIVERSIDAD DE SAN MARTÍN DE PORRES, LIMA-PERÚ” en la ciudad Lima del país Perú. El **objetivo** general de su estudio es el uso de minería de datos para predecir el rendimiento académico de estudiantes de primer año de la Escuela Profesional de Informática y Sistemas de la Universidad de San Martín de Porres. La **muestra** estuvo constituida por de aproximadamente 1300 estudiantes. Su diseño fue el diseño de investigación transaccional del tipo causal correlacional. Los **instrumentos** que se usaron fueron: Los datos históricos sobre los estudiantes y su rendimiento académico, incluyendo el GPA, la modalidad de admisión, el distrito de origen, la situación familiar y socioeconómica, fueron compilados a partir de fuentes de datos de las oficinas de admisiones y de la facultad de ingeniería y arquitectura de la Universidad San Martín de Porres.. Los resultados demuestran que, si los nuevos solicitantes pueden completar con éxito todos los cursos en su primer ciclo, hay datos suficientes para permitir predicciones futuras sobre su clasificación. El mejor modelo identificado tuvo unos valores máximos del 90,4% y una precisión de predicción del 82,87%. Dado que se obtuvo un valor P de la prueba F inferior a 0,05, puede deducirse de los resultados del análisis estadístico ANOVA que existe una diferencia

estadísticamente significativa entre los valores medios con un nivel de confianza del 95,0%. La misma prueba también arrojó un valor P inferior a 0,05 en otras variables, como puede demostrarse mediante enfoques de minería de datos.

Holgado (2018) realizó un estudio titulado: “DETECCIÓN DE PATRONES DE BAJO RENDIMIENTO ACADÉMICO MEDIANTE TÉCNICAS DE MINERÍA DE DATOS DE LOS ESTUDIANTES DE LA UNIVERSIDAD NACIONAL AMAZÓNICA DE MADRE DE DIOS 2018” en la ciudad PUNO del país Perú. El su estudio el **objetivo** general es utilizar la minería de datos para identificar tendencias en el bajo rendimiento académico de los estudiantes de la Universidad Nacional Amazónica de Madre de Dios. La **muestra** estuvo constituida por la población para el presente estudio incluyó 9545 registros de estudiantes matriculados en la Universidad Nacional Amazónica de Madre de Dios entre los semestres de 2001 y 2018-I.. El **diseño** que se utilizó fue documental no experimental (Arias, 2006), se seleccionó CRISP-DM como enfoque para alcanzar los objetivos. A petición del estudiante, este trabajo se llevó a cabo con la ayuda de profesionales con acceso a los datos del DUAA. Y los **resultados** obtenidos han sido. La representación gráfica del modelo de árbol de clasificación se muestra en la Figura 54. La hoja 7 muestra que el 33% de los estudiantes fueron categorizados en el grupo B y el 67% en la categoría C, que juntos constituyen el 25% de los datos globales. El 18% de los alumnos fueron categorizados en la categoría B y el 82% en la categoría C en la hoja 8, que en conjunto constituyen otro 25% de los datos globales. Según este árbol de categorización, las hojas 7 y 8 incluyen el 50% del número total de alumnos, y el 67% y el 82% de ellos pertenecen al grupo C, respectivamente. La calificación de estos alumnos oscila entre cero y diez. Resumiendo, la hoja 8, podemos decir que responden al siguiente perfil: Alumnos que han cursado más de seis pero menos de sesenta y dos asignaturas, que no han prestado servicio en el comedor universitario y que deben algún dinero a la institución. Además, podemos concluir de la ficha 7 que este grupo de estudiantes no debe dinero a la institución y no trabaja en los campos de la administración y los negocios internacionales, la contabilidad y las finanzas, el derecho y las ciencias políticas, o la educación inicial y especial.

Alania (2018) realizó un estudio titulado: “APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA PREDECIR LA DESERCIÓN ESTUDIANTIL DE LA

FACULTAD DE INGENIERÍA DE LA UNIVERSIDAD NACIONAL DANIEL ALCIDES CARRIÓN". en la ciudad PASCO del país PERÚ

El **objetivo** general del estudio fue que utilizando herramientas de minería de datos para pronosticar la deserción estudiantil en la Facultad de Ingeniería de la Universidad Tecnológica Nacional Daniel Carrión. La **muestra** estuvo integrado por 218 estudiantes de la Escuela Profesional de Sistemas y Computación Daniel Alcides Carrión de la Facultad de Ingeniería de la Universidad. El **diseño** que se utilizó fue: transeccional del tipo correlacional causal. En los **instrumentos** se utilizaron datos históricos sobre los estudiantes y su rendimiento académico que fueron tomados de las bases de datos de la Oficina de Admisiones y de la Oficina de Informática de la Universidad Nacional Daniel Alcides Carrión. Estos datos incluían promedios ponderados de cada semestre, métodos de admisión, circunstancias familiares y socioeconómicas, entre otros.. Y los **resultados** obtenidos han sido con un nivel de confianza del 95% y un nivel de significación del 5%, el abandono escolar y la nota media difieren significativamente. De ello se deduce que la tasa de abandono de la escuela de formación profesional de informática y sistemas de la UNDAC está influida significativamente por la nota media.

Candia (2019) realizó un estudio titulado: "PREDICCIÓN DEL RENDIMIENTO ACADÉMICO DE LOS ESTUDIANTES DE LA UNSAAC A PARTIR DE SUS DATOS DE INGRESO UTILIZANDO ALGORITMOS DE APRENDIZAJE AUTOMÁTICO" en la ciudad Cusco del país Perú.

El **objetivo** general del estudio fue emplear técnicas de aprendizaje automáticos para predecir los estudiantes de primer semestre de la UNSAAC a partir de sus datos de egreso. Doce mil estudiantes ingresantes de diversas modalidades de la UNSAAC integraron la manifestación. El diseño utilizado fue del tipo correlacional más que experimental. Los **instrumentos** que se usaron a sido recabar información sobre estadísticas de admisión y encuestas en el sitio web de la UNSAAC. Y los **resultados** obtenidos han sido: El algoritmo Función Logística quedó en segundo lugar con una precisión de predicción del 68,33%, mientras que el algoritmo Bosque Aleatorio obtuvo el mayor rendimiento, es decir, el mejor porcentaje de predicción.

Luna (2020) realizó un estudio titulado: "IMPLEMENTACIÓN DE UN SISTEMA DE PREDICCIÓN DEL RENDIMIENTO ACADÉMICO DE LOS ESTUDIANTES DE

INGENIERÍA DE SISTEMAS DE LA UNIVERSIDAD NACIONAL JOSÉ MARÍA ARGUEDAS UTILIZANDO TÉCNICAS DE MINERÍA DE DATOS PARA LA ADECUADA TOMA DE DECISIONES" en la ciudad ANDAHUAYLAS del país PERÚ.

El **objetivo** general del estudio fue: Evaluar la eficacia de un sistema basado en la minería de datos para predecir el éxito académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas de la UNAJMA. La **muestra** estuvo constituida por se obtuvo un total de 1975 datos. Los instrumentos utilizados como base de datos incluyeron datos extraídos de 1975 de expedientes académicos y nivel socioeconómico familiar, los cuales determinaron las predicciones de los estudiantes de ingeniería de sistemas. Y los resultados obtenidos han sido estudiantes de ingeniería de sistemas proporcionaron una muestra de datos de 1.380 puntos para el experimento. Otros algoritmos utilizados fueron SMOreg, IBK, Kstar, LWL, Bagging, Vote, ZeroR, M5P, RandomForest, RandomTree y REPTree. Para el análisis se utilizó la prueba estadística Weka Crossvalidation 10. Diez submuestras se dividen aleatoriamente de la muestra original para este tipo de prueba. Chaque uno y diez iteraciones del procedimiento de aprendizaje se realizaron diez validaciones cruzadas sobre diez hojas. En el experimento se compararon los errores medios absolutos de cada algoritmo. La tabla estadística muestra que, mientras que el error absoluto medio de los otros algoritmos es mayor, como se muestra en la Tabla 8, el método KStar tiene el error absoluto medio más bajo de 1,18. El algoritmo KStar es por lo tanto más efectivo. Se utilizaron 1674 puntos de datos de estudiantes de ingeniería de sistemas. Entre los algoritmos utilizados en el experimento se encuentran SMOreg, IBK, Kstar, LWL, Bagging, Vote, ZeroR, M5P, RandomForest, RandomTree y REPTree. El tipo de validación cruzada 10 de Weka fue probado estadísticamente. Para este tipo de prueba, la demostración original se divide aleatoriamente en diez submuestras. Chaque uno y diez iteraciones del procedimiento de aprendizaje se realizaron diez validaciones cruzadas sobre diez hojas. En el experimento se compararon los errores medios absolutos de cada algoritmo. La tabla estadística muestra que, mientras que el error absoluto medio de los otros algoritmos es mayor, como se muestra en la Tabla 9, el algoritmo KStar tiene el error absoluto medio más bajo (0,85). El algoritmo KStar es esencial, por tanto. Los datos reales de los estudiantes de Ingeniería de Sistemas del semestre 2018-I y los datos esperados derivados de esos datos muestran una inexactitud del 11%. H2 es por lo tanto cierto, como se esperaba.

3.2 Bases teóricas

3.2.1. MÍNERIA DE DATOS

La minería de datos es muy usada en los últimos tiempos para descubrir patrones por lo mismo podemos decir: la minería de datos es un sistema de información basado en la informática que engloba grandes repositorios de datos con el fin de proporcionar conocimiento e información. Aunque el término proviene de la minería tradicional, el objetivo es encontrar conocimientos que ayuden a descubrir aspectos como patrones interesantes, relaciones entre datos, definición de reglas, predicción de valores desconocidos, agrupación de objetos homogéneos y otras cosas que son difíciles de encontrar en la minería tradicional. sistemas de información tradicionales. (Peña, 2014).

Además (Yamao, 2018) citando a (Han et al., 2012) indica que la mineria de dtos es el proceso de extracción de grandes volúmenes de datos para encontrar nuevos patrones e información. Las bases de datos, los almacenes de datos, Internet, otras formas de repositorios o los datos dinámicos que se introducen directamente en el sistema pueden utilizarse como fuentes de estos datos. La necesidad de examinar de forma automática e inteligente la ingente cantidad de datos que se producen cada día parece estar cubierta por la minería de datos. Las diversas redes e interacciones de las empresas, las personas, la investigación, la ingeniería, la medicina y otros elementos de la vida cotidiana crean terabytes o petabytes de datos (p. 27).

Por otro lado, (Microsoft, 2021) En los conjuntos grandes de datos, se conoce como minería de datos el proceso de identificar la información procesable. Utiliza el análisis matemático para deducir los patrones y tendencias observados en los datos. En términos generales, la exploración de datos tradicionales no puede encontrar estos patrones porque las relaciones son demasiado complicadas o hay demasiados puntos de datos.

Este patrón y tendencia pueden recopilarse y definirse como un modelo de minería de datos. Las siguientes son algunas situaciones en las que se pueden utilizar modelos de minería de datos.:

Pronostico. La previsión incluye el cálculo de los ingresos y la predicción de las cargas y caídas de los servidores.

Riesgo y probabilidad. Los mejores destinatarios de correo directo a los que dirigirse, el punto de equilibrio más probable para situaciones de riesgo. y las probabilidades de diagnóstico y otros resultados son todos aspectos del riesgo y la probabilidad.

Recomendaciones. Creación de sugerencias y determinación de los artículos que pueden comercializarse conjuntamente.

Búsqueda de secuencias: evaluación de productos que los compradores han añadido a sus carros al momento de la compra y previsión de resultados más posibles.

Agrupación: Distribución de consumidores o eventos en grupos de objetos relacionados, análisis y predicción de afinidad.

De acuerdo con (Hernandes, 2004) nos indica que los métodos de análisis de datos y extracción de modelos se combinan bajo la expresión relativamente nueva de "minería de datos". La información informatizada que nos rodea hoy en día, que suele ser heterogénea y abundante, también puede utilizarse para extraer patrones, describir tendencias y regularidades, predecir comportamientos y, en general. Esto permite a las personas y a las organizaciones comprender y modelizar el contexto en el que deben actuar y tomar decisiones de forma más eficaz y precisa.

3.2.1.1. *El proceso de descubrimiento de conocimiento en bases de datos (KDD).*

De acuerdo con (Mondragon, 2007) citando a (Fayyad y otros, 1996), afirma que el KDD es un método no común para encontrar patrones verdaderos, originales, posiblemente beneficiosos y perceptibles en los datos. En este contexto, los términos "datos" y "patrones" se refieren a frases en algunos idiomas que explican sucintamente los datos. En tal sentido, los datos se refieren a una colección de eventos (como los de una base de datos).

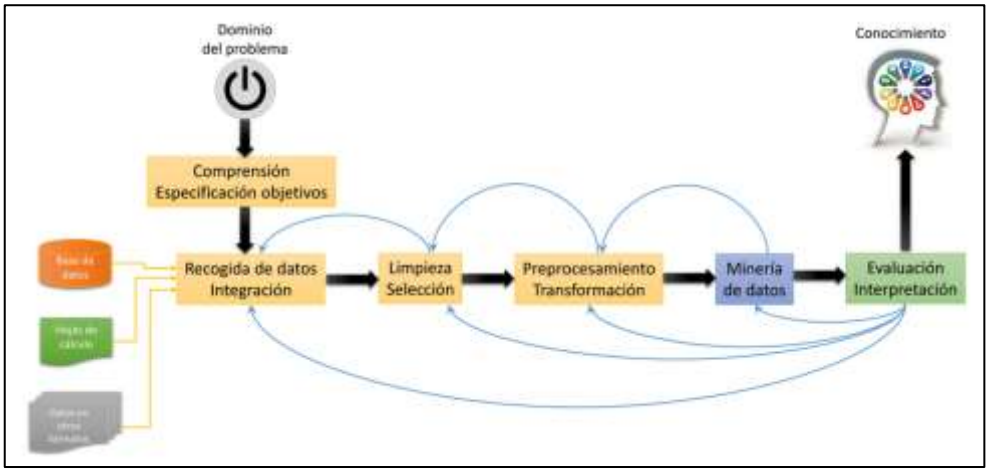
Como implica el nombre "proceso", el KDD conlleva una serie de procesos, como la preparación de datos, la búsqueda de patrones, la evaluación de conocimientos y el refinamiento, todos los cuales pueden realizarse varias veces. Hay que reconocer que por no trivial se está realizando algún tipo de búsqueda o inferencia, es decir, una búsqueda de estructuras, modelos, patrones o parámetros. Para caracterizar y/o pronosticar con precisión el comportamiento futuro de algún objeto, los patrones encontrados deben ser

coherentes con los datos frescos hasta cierto punto. Los patrones también deben ser originales (al menos para el sistema, pero idealmente para el usuario) y posiblemente beneficiosos, es decir, deben ayudar al usuario o al trabajo de alguna manera. Por último, pero no por ello menos importante, los patrones deben ser interpretables; de lo contrario, sería necesario un postprocesamiento (p. 6).

El método KDD es participativo e iterativo; consta de varias fases y se basa en gran medida en las decisiones de los usuarios. Una perspectiva realista del proceso KDD la ofrecen Brachman y Anand (1996), que hacen hincapié en el carácter participativo del proceso, como se ve en la Figura 1.

Figura 1

Esquema del proceso KDD



Nota. La figura muestra el esquema del proceso KDD. Fuente: Charte (2020).

De acuerdo con la Figura 1, para realizar el KDD se sigue 5 etapas, (Charte, Campusmvp, 2020) nos describe cada una de la las etapas:

Recogida de datos: se cometen errores de forma rutinaria durante la recopilación de datos, ya sea humana o automatizada, así como durante la codificación y transmisión posteriores hasta el punto en que finalmente se ensamblan algunos de los datos. Muchas veces, esto se traduce en dos tipos diferentes de problemas. El primero se basa en datos que parecen inexactos y, como tales, serían una barrera para el proceso de extracción de conocimiento en lugar de un beneficio. El segundo resulta en la eliminación de algunos datos, esto que podría causar problemas en el conjunto. Estos obstáculos podrían

abordarse mediante tareas de limpieza de datos, por ejemplo, utilizando algoritmos para la eliminación de ruido o la imputación de valores ausentes.

Selección de datos relevantes: el KDD no requiere que todos los datos obtenidos de las fuentes originales sean relevantes, por lo que el preprocesamiento avance en la selección de aquellos que adecuadamente sean útiles. Las dos tareas más comunes en esta fase son la selección de variables y la selección de usuarios. Fundamentalmente, incluyen eliminar cualquier dato que no mejore la extracción de conocimiento, ya que es redundante o puede inferirse de otros datos. Estas dos técnicas están incluidas en los métodos de reducción de dimensionalidad.

Transformación de los datos: Es posible considerar que los datos pueden ser empleados para el aprendizaje de un modelo ya que están limpios y sin redundancias, aspectos de los que se ocupan las operaciones previas. Sin embargo, existen medidas que pueden optimizar los datos de cara para que el aprendizaje sea más eficaz. Entre estos se encuentran la normalización, el escalado y la discretización. Estas son operaciones que casi siempre producen cambios reversibles en los datos originales, creando una nueva versión que es más adecuada para el análisis KDD.

Minería de datos: después del preprocesamiento, los datos están listos para la siguiente fase. En esto, se emplea un algoritmo de minería de datos para extraer de estos conocimientos el cual, aunque este implícito en ellos, no resulte obvio ni simple.

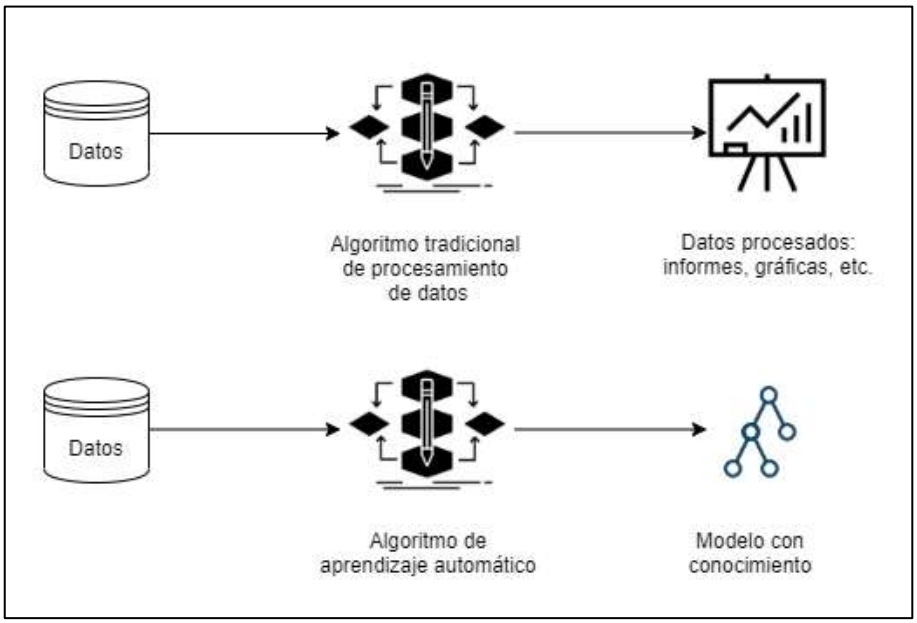
Numerosas tácticas estadísticas, algoritmos matemáticos de optimización y, de otra manera, métodos de aprendizaje automático son entre los que se pueden utilizar en este punto.

A diferencia de los pasos anteriores, cuando es necesaria la exploración de datos para establecer qué operaciones deben realizarse y la intervención de expertos es esencial, la fase de aprendizaje automático es donde se suele utilizar el aprendizaje automático.

El proceso de datos por un algoritmo de aprendizaje automático resulta en un modelo que representa el conocimiento extraído, y no en nuevos datos, como es habitual en la mayoría de algoritmos de ordenador:

Figura 2

Modelo de conocimiento extraído



Nota: La figura muestra el modelo de conocimiento extraído. Fuente: Charte (2020)

Luego de haber realizado todos los procesos anteriores se tiene que interpretar y evaluar los resultados para ello Timarán et al. (2016) nos dice que la:

Método de interpretación y evaluación de datos. Los patrones descubiertos se interpretan en la etapa de interpretación/evaluación, y es posible regresar a etapas anteriores para iteraciones posteriores. Parte de esta etapa podría incluir la visualización de patrones extraños, la eliminación de patrones redundantes o innecesarios y la traducción de patrones útiles a términos que el usuario pueda comprender. Por el contrario, el conocimiento descubierto se combina para integrarse en otro sistema para acciones futuras o, en pocas palabras, para ser registrado y reportado a las partes interesadas. También se utiliza para confirmar y resolver cualquier conflicto con conocimientos descubiertos previamente. (p. 67)

De acuerdo con Perez & Santin (2007) las siguientes etapas comprenden el proceso de extracción de conocimientos de KDD:

Figura 3

Esquema del proceso de extracción del conocimiento KDD

Nota. La figura muestra el proceso de extracción del conocimiento KDD. Fuente: Extraida de (Perez & Santin, 2007)

El proceso de extracción de conocimiento KDD Fuentes et al. (2014) lo clasifican mediante una secuncia de fases:

Selección: La fase de selección produce la integración y compilación de los datos, la identificación de posibles fuentes de información y su ubicación, la identificación y selección de las variables relevantes en los datos y el uso de técnicas de prueba adecuadas.

Todo eso se facilita tener un almacén de datos con la información en formato uniforme y libre de errores.

Exploración: Los datos provienen de muchas fuentes, por lo que es necesario investigarlos utilizando técnicas de análisis de datos exploratorios. Estos enfoques incluyen la búsqueda de correlaciones en los datos, así como su distribución, simetría y normalidad.

Limpieza: Aquí se verifican los datos, ya que pueden incluir valores atípicos, números que faltan o valores incorrectos.

Transformación: Esta etapa consiste en la transformación del atributo (numeración, discretización, etc.).

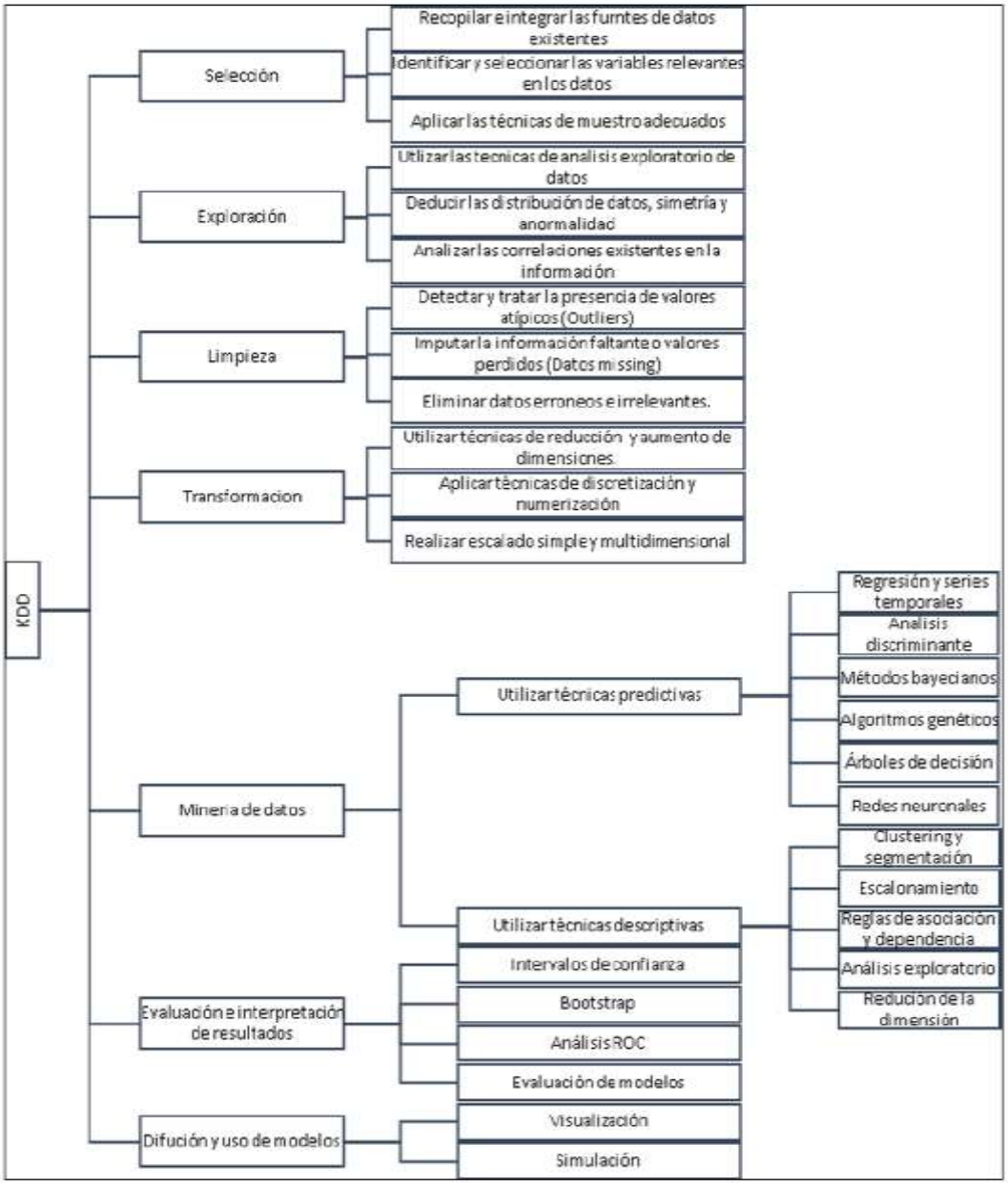
La preparación de los datos suele referirse a las cuatro primeras etapas. Durante la fase de extracción de datos se determina el trabajo que debe realizarse (clasificación, agrupación, etc.) y el enfoque descriptivo o predictivo que debe aplicarse.

Evaluación e interpretación: Los expertos evalúan e interpretan los patrones durante la fase de evaluación e interpretación. Si es necesario, se revisan las fases anteriores para la siguiente iteración. Los nuevos conocimientos se utilizan y difunden a todos los posibles consumidores en este punto, durante la fase de difusión (p. 35).

A este respecto, las etapas del proceso de extracción de conocimientos pueden clasificarse según el siguiente esquema:

Figura 4

Esquema de clasificación del proceso de extracción del conocimiento



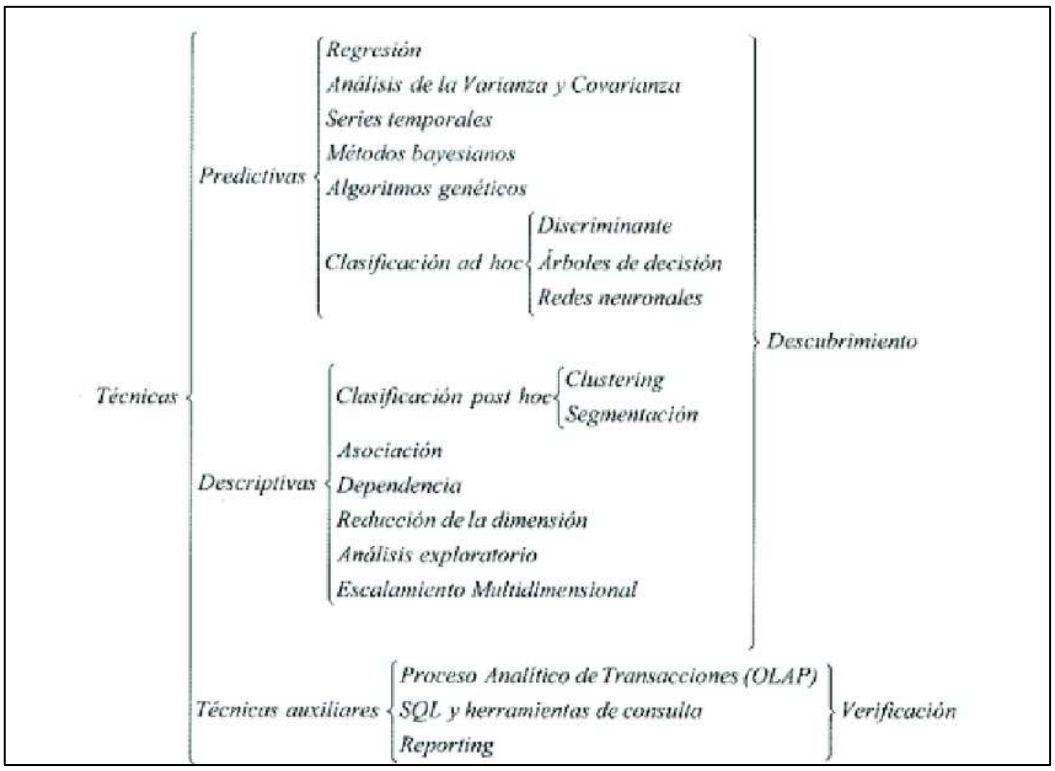
Nota. Esquema de clasificación del proceso de extracción del conocimiento. Fuente: Holgado (2018)

3.2.1.2. Clasificación de las Técnicas de Minería

A continuación, se muestra una clasificación de las técnicas de Data Mining.

Figura 5

Esquema de clasificación de las técnicas de Data Mining



Nota. Diagrama de clasificación de las técnicas de Data Mining. Fuente. (Perez & Santin, 2007).

Técnicas predictivas o supervisadas

De acuerdo con Moreno et al. (2001) afirme que: Estos algoritmos calculan la valoración de un atributo (etiqueta) basándose en una colección de datos, comparativamente referidos como otros atributos (atributos descriptivos). Se puede inferir una relación entre un rasgo identificado y otro conjunto de atributos basándose en datos. Estas relaciones ayudan a predecir datos cuya etiqueta se desconoce. Este método de aprendizaje es el aprendizaje supervisado y avanza a través de dos etapas: entrenamiento implica construir un modelo utilizando un subconjunto de datos con una etiqueta conocida, mientras que prueba consiste en probar el modelo con los datos restantes. (p. 3).

A su vez Rosado & Verjel (2014) indica que las técnicas de predicción manejan tareas de regresión y clasificación, por su parte (Valcárcel, 2004) indica que para obtener de un

modelo que posibilita predecir el valor numérico de una variable se lleva a cabo mediante tareas de regresión logística, de igual manera (Aluja, 2001) afirma que si la respuesta es categórica (por ejemplo, comprar o no comprar un producto), decimos que existe un problema de clasificación.

Técnicas descriptivas o No supervisadas

Sobre las técnicas descriptivas o no supervisadas podemos encontrar diferentes puntos de vista, en los enfoques descriptivos no se asigna a las variables ningún papel preconcebido. No se espera que existan variables dependientes o independientes ni que haya un modelo previo para los datos. La creación automática de modelos se basa en el reconocimiento de patrones. Esta categoría incluye técnicas de asociación y reducción de dimensionalidad (factorial, componentes principales, correspondencia, etc.), escalamiento multidimensional y enfoques de agrupamiento y segmentación (que también son técnicas de clasificación hasta cierto punto).

La identificación de la información oculta en los datos es el principal objetivo tanto de los enfoques descriptivos como de los predictivos (Perez & Santin, 2007, p. 9 citado por Holgado, 2018, p. 30).

De acuerdo con Moreno et al. (2001) sin utilizar datos históricos, Estas técnicas permiten identificar patrones y tendencias en los datos presentes. El descubrimiento de dicha información es ventajoso para emprender acciones y obtener un beneficio (comercial o científico) de las mismas.

3.2.1.3. Técnicas de minería de datos

Algoritmos de clasificación

Un método de análisis de datos conocido como clasificación permite el desarrollo de modelos que describen las clases esenciales de datos. Estos modelos permiten predecir valores categóricos mediante un proceso de dos fases que incluye el aprendizaje (el proceso en el que se forma el modelo) y la clasificación (en el que el modelo se utiliza para predecir los valores de otros datos). (Han et al., 2012 citado por Yamao, 2018, p. 37).

La descripción de los principales algoritmos de clasificación según lo mencionado por Yamao (2018) cita a (Lantz, 2013) y Han et al. (2012) se encuentran a continuación:

Árbol de decisión. representación en forma de árbol de una colección de criterios de categorización. Los árboles de decisión originales fueron creados por especialistas

humanos, pero en la actualidad cada vez son más frecuentes los árboles creados por un algoritmo. Los más conocidos son ID3 y C4.5. El principio fundamental de la técnica de los árboles de decisión es que se puede determinar el punto final del camino a partir de las propiedades de cada clase.

Entre los beneficios de los árboles de decisión se encuentran su simplicidad y facilidad de comprensión, su capacidad para trabajar con variables numéricas y de categorización, clasificación rápida de nuevos datos y relativa simplicidad y flexibilidad en la gestión del ruido y los valores perdidos dentro de una clase. Por el contrario, si hay pocos datos de entrenamiento, puede conducir a un modelo muy específico y difícil de generalizar.

La suposición de que todos los datos pueden clasificarse de manera determinista en una clase es la principal limitación de un árbol de decisión. Por lo tanto, todas las discrepancias se consideran errores, lo que hace inapropiado clasificar datos como el desempeño de los estudiantes en una clase, que a menudo incluye discrepancias.

Clasificadores bayesianos

Las relaciones estadísticas se representan gráficamente en una red bayesiana para mostrar su naturaleza visual. Considera la independencia condicional de toda la información y representa las características con una estructura de dependencia mínima. Las aristas del grafo reflejan el conjunto de propiedades de las que depende cada vértice del grafo.

El grado de dependencia se muestra mediante una probabilidad condicional. Antes de utilizar redes bayesianas para hacer predicciones, primero se debe comprender la red para determinar las dependencias de cada atributo y desarrollar un modelo fundamental en torno al cual se realizarán futuras predicciones de datos.

En realidad, el problema surge del enorme volumen de probabilidades que debemos estimar para determinar la probabilidad que tiene cada característica. Un modelo bayesiano ingenuo puede resolver este problema, ya que se necesitan menos estimaciones porque se supone que todas las características son condicionalmente independientes. El modelo bayesiano ingenuo ha demostrado un gran rendimiento de predicción en la realidad, sin embargo, debido a las suposiciones que se hacen en su uso, su capacidad de representación del modelo es inferior a la de otros modelos, como los árboles de decisión. Sin embargo, el modelo bayesiano ingenuo tiene varias ventajas, como ser sencillo, eficaz, resistente al ruido y fácil de entender. Dado que combina un bajo nivel de complejidad con un modelo probabilístico flexible, se adapta convenientemente a muestras de datos minúsculas. Los datos numéricos deben convertirse porque el modelo fundamental sólo tiene en cuenta datos discretos.

Redes neuronales

En términos de reconocimiento de patrones, son muy apreciadas, pero cuando se utilizan en educación, pueden plantear problemas a menos que se disponga de una cantidad significativa de datos numéricos y de la experiencia de un experto para entrenar el modelo.

El tipo más popular de red neuronal es la red neuronal feed-forward (FFNN). Entrada, salida y una o más capas ocultas intermedias son sólo algunas de las capas de nodos que componen su diseño. Se asignan pesos a cada capa oculta, que está conectada a los nodos de las capas superior e inferior. En una red de tres capas puede representarse teóricamente cualquier función, pero, de hecho, el entrenamiento de la red es todo un reto.

Las redes neuronales tienen la ventaja de poder representar modelos no lineales y, en teoría, cualquier tipo de clasificador. Las FFNN pueden modificar implícitamente los datos originales, aunque carezcan de propiedades discriminativas. También son resistentes al ruido y actualizables con información fresca.

El principal inconveniente es que se necesita una cantidad considerable de datos, muchos más de los que se suelen obtener en un contexto educativo. Los valores categóricos deben cuantificarse de algún modo antes de poder utilizarse, y los datos deben ser numéricos. En consecuencia, el modelo se complica y los resultados dependen del método de cuantificación elegido.

Dado que el modelo de red neuronal es una "caja negra", también es difícil ofrecer una explicación directa de los resultados. Además, suelen ser poco fiables y sólo funcionan bien cuando se utilizan correctamente. Puede llevar un tiempo entrenar el modelo, sobre todo si se quiere construir un modelo genérico (pp. 37-39).

3.2.1.4. Herramientas de Minería de datos

Es posible clasificar dos bibliotecas y herramientas diferentes para el uso de técnicas de minería de datos: Bibliotecas de minería de datos es una colección de técnicas que permite el acceso a datos, los modelos de redes neuronales, los enfoques bayesianos, la exportación de resultados y otras capacidades y utilidades fundamentales. Principalmente, las bibliotecas se utilizan para facilitar las tareas de extracción de datos más difíciles, incluido el diseño de experimentos. La dificultad con las bibliotecas es que hay que ser competente en programación para utilizarlas. Entre las bibliotecas más significativas se encuentran estas:

1. Xelopes (Extended Library For Prudys Embedded Solution): Es una librería para crear aplicaciones de Minería de Datos que está liberada bajo la licencia pública GNU.
Es crucial resaltar que el usuario puede crear aplicaciones específicas de Minería de Datos porque esta librería ha sido diseñada para ser efectiva para la mayoría de los métodos de aprendizaje. Los atributos principales son:

1. Acceso a datos
2. Modelos de redes neuronales
3. Técnicas de clustering
4. Técnicas de reglas de asociación
5. Arboles lineales
6. Árboles no lineales

2. Mlc++ (Machine Learning Library In C++): Consiste en una colección de bibliotecas creadas por la Universidad de Stanford. Salvo la versión 1.3.x, que se ofrece bajo licencia de dominio público, la mayoría de las versiones son de dominio de investigación. Los principales atributos son:

1. Acceso a datos.
2. Transformaciones de datos
3. Métodos de aprendizaje mediante objetos

3. Suites: Poseen las mismas capacidades que el procesamiento de datos, los modelos analíticos, el diseño de experimentos y el soporte gráfico para la visualización de resultados. Suites se caracterizan en este caso por la interfaz de usuario que facilita la interacción de los usuarios con la herramienta.

4. R-Project: Es un entorno de trabajo basado en los entornos de programación S y S-PLUS creados a principios de los años 90 por Bill Venables y David M. Comprende, como afirman Venables et al (2011), un conjunto de recursos informáticos integrados para el procesamiento de datos. , cálculo y desarrollo visual. R-Project aspira a convertirse en un sistema internamente coherente que se distinga por un crecimiento basado en la contribución relativamente desinteresada de la sociedad científica. (Alania Ricaldi, 2018) cita a (López Puga, 2010)

5. SPSS Clementine: Uno de los sistemas de minería de datos más conocidos es éste. Tiene una arquitectura cliente/servidor y una herramienta visual desarrollada por ISL. Estas son las características de este sistema:

1. Acceso a datos.
2. Tratamiento de Datos.

3. Técnicas de aprendizaje.

4. Métodos de evaluación de modelos.

5. La visualización de resultados.

6. Exportaciones.

6. Weka (Waikato Environment For Knowledge Analysis): Investigadores de la Universidad de Waikato (Nueva Zelanda) crearon esta herramienta visual de acceso público. Sus principales atributos son:

1. Acceso a datos desde un archivo con formato ARFF.

2. Datos preprocesados.

3. Modelos de Aprendizaje.

4. Visualización del ambiente.

7. Kepler: Dialogis creó un sistema y lo convirtió en un producto con ánimo de lucro que se distribuye. Presenta varios modelos analíticos. Sus principales recursos didácticos son:

1. Árboles de decisión.

2. Redes neuronales.

3. Regresión no lineal.

4. Aplicaciones estadísticas.

8. Odms (Oracle Data Mining Suite): Se basa en una arquitectura cliente-servidor y ofrece una gran flexibilidad a la hora de acceder a grandes cantidades de datos.

1. Acceso a datos en diferentes formas, como bases de datos relacionales como SQL y Oracle.

2. Datos preprocesados: datos muestreo, datos patrones.

3. Modelos de aprendizaje: redes neuronales y regresión lineal.

4. Instrumentos de visualización.

9. Yale: Herramienta de aprendizaje automático desarrollada en Java por la Universidad de Dormund.

El sistema tiene operaciones para:

1. Importar y preprocesar datos

2. Aprendizaje autónomo

3. Validación de modelos

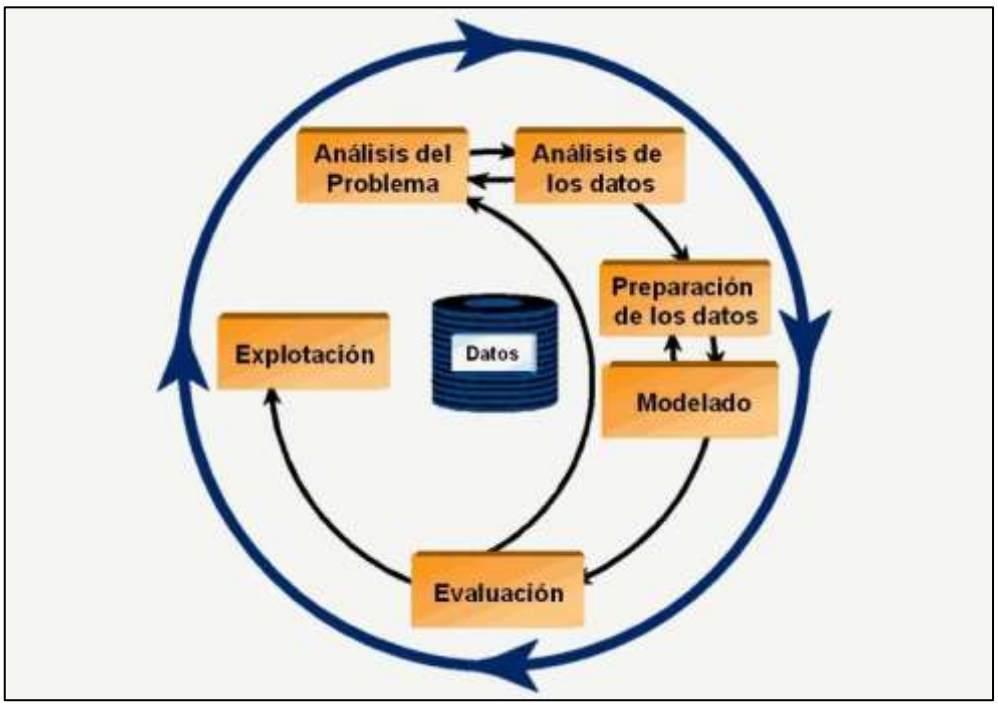
3.2.1.5. Metodología de minería de datos

CRISP-DM (Cross Industry Standard Process for Data Mining)

De acuerdo a Moine et al. (2012) fue desarrollado en 2000 por los grupos de empresas SPSS, NCR y Daimer Chrysler, y es actualmente el manual de referencia más utilizado por quienes trabajan en proyectos de minería de datos. El proceso se divide en seis fases: comprensión del negocio, preparación de los datos, interpretación de los datos, modelización, evaluación e implementación. El orden de las fases no tiene por qué ser estricto. Cada fase se divide en una serie de actividades de segundo nivel a nivel general. Para cada fase del proyecto, CRISP-DM define una lista de tareas y actividades, pero no menciona cómo completarlas (p. 933).

Figura 6

Fases del proceso KDD según la metodología CRISP-DM



Nota: Fases del proceso KDD según la metodología CRISP-DM. Fuente: Ordoñes & Grass (2011)

A continuación, detallamos cada una de las seis etapas, teniendo un conjunto de tareas detalladas en cuatro niveles diferentes de abstracción, desde el más general al más concreto: fase, tarea general, tarea específica y instancia de proceso.

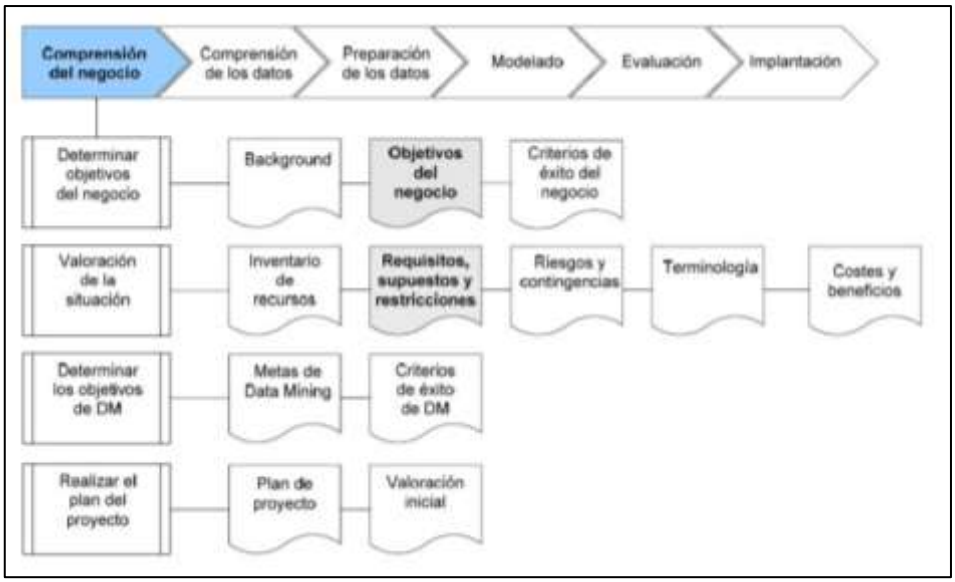
1. Fase de compresión del problema o negocio

De acuerdo con (Gallardo Arancibia, 2009) para convertir los objetivos y requisitos del proyecto en objetivos técnicos y un plan de proyecto, esta fase es probablemente la más

importante. Sin comprender estos objetivos, ningún algoritmo, por sofisticado que sea, puede proporcionar resultados fiables. (p. 17)

Figura 7

Fase de comprensión del negocio



Nota: La figura muestra la fase de comprensión del negocio. Fuente: Gallardo (2009).

Las descripciones de cada una de las tareas forman parte de esta fase son las siguientes:

Determinar los objetivos del negocio

Para determinar los objetivos del negocio HOLGADO (2018**)** Afirma textualmente:

Significa determinar el alcance del proyecto; Para ello desarrollamos las siguientes preguntas: ¿Cuál problema debemos resolver? ¿Qué queremos lograr? ¿Qué ventajas aportaremos al cliente? ¿Por qué es necesario utilizar la minería de datos y establecer los criterios de éxito para el objetivo de la empresa? Se aceptan criterios cuantitativos como cualitativos.

Evaluación de la situación actual

Antes de comenzar el proyecto de minería de datos, debemos considerar el estado de la situación en el momento en que se realiza este trabajo. En tal sentido, serán de utilidad las siguientes preguntas: ¿Qué recursos o necesidades (hardware, software o recursos humanos) utilizamos o necesitamos? ¿Qué información previa tienes sobre el tema? ¿Cuáles son las premisas y restricciones subyacentes? ¿Cuál es la relación coste-beneficio del proyecto de minería de datos? En esta sección, esbozamos las necesidades desde el punto de vista de la empresa y de la minería de datos (pp. 25-26).

Determinación de los objetivos de minería de datos

Para esta tarea Gallardo (2009) nos indica textualmente: Este trabajo representa los objetivos comerciales en términos de los objetivos del proyecto DM. Suponiendo, por ejemplo, que el objetivo empresarial es lanzar una campaña de marketing para aumentar la asignación de crédito hipotético, el objetivo de DM sería determinar el perfil del cliente con respecto a su capacidad de pagar la deuda. (p. 18)

Producción del plan de proyecto

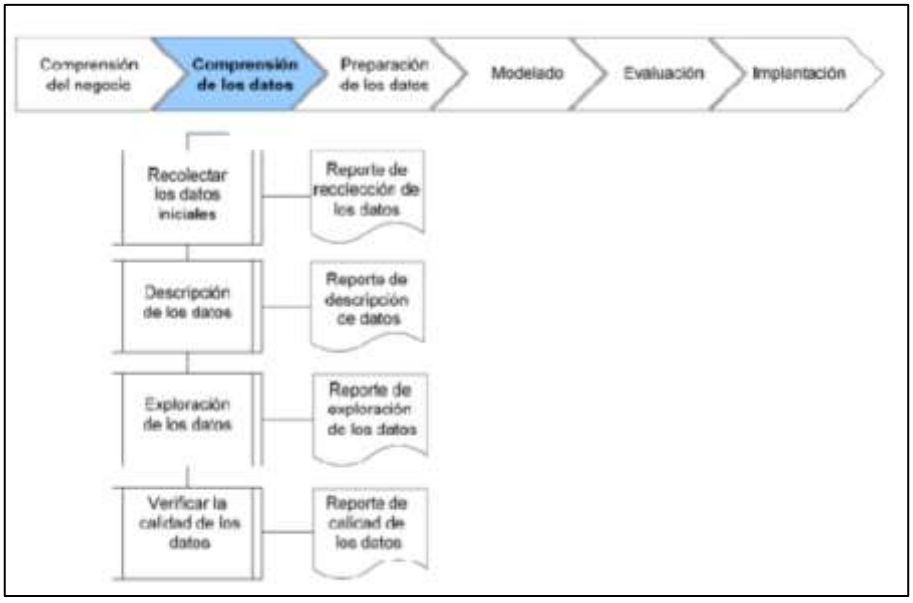
Desde la posición de Holgado (2018) afirma que: "En este caso, la última tarea de la primera fase es crear un plan de proyecto que delinee los procedimientos a seguir y las técnicas a utilizar en cada uno de ellos". (p. 26)

2. Fase de compresión de los datos

La segunda fase (Figura 8) es para la recolección de los datos, tenemos que familiarizarnos con los datos y la información que vamos a utilizar además de evaluar su calidad e identificas las relaciones.

Figura 8

Fase de comprensión de los datos



Nota: La figura muestra la fase de comprensión de los datos. Fuente: [CRISP-DM, 2000].

De acuerdo a Gallardo (2009) nos indica que esta fase comprende las siguientes tareas:

Recolección de datos iniciales. La recopilación de los datos iniciales y la determinación de si son adecuados para su procesamiento posterior es la primera tarea de esta segunda fase del proceso CRISP-DM. El objetivo de esta tarea es proporcionar informes que

describan los datos recopilados, su ubicación, los métodos utilizados para recopilarlos, los problemas que surgieron a lo largo de este proceso y las soluciones para dichos problemas.

Descripción de los datos. Una vez recogido el primer conjunto de datos, hay que caracterizarlo. Este procedimiento consiste en establecer los volúmenes de datos (número de registros y campos por registro), su identificación, la definición de cada campo y la primera descripción del formato.

Exploración de datos. A continuación, se investigan los datos para determinar una estructura general de los mismos. Se elaboran tablas de frecuencias, gráficos de distribución y se aplican pruebas estadísticas básicas para mostrar las cualidades de los datos recién recogidos. El resultado de la tarea es un informe sobre la exploración de los datos.

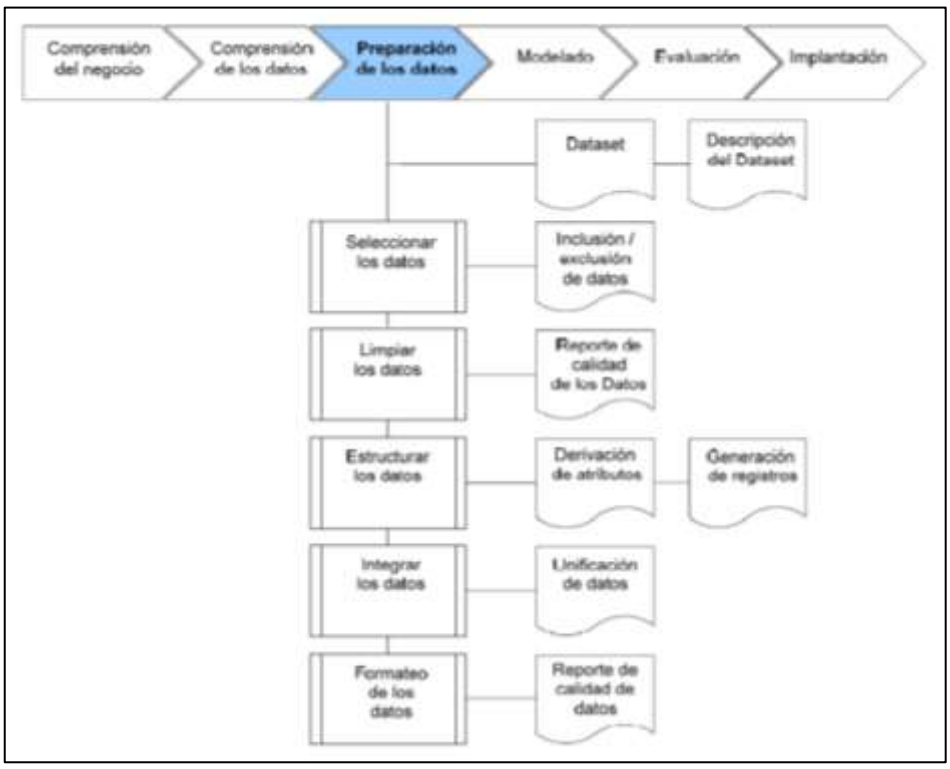
Verificación de la calidad de los datos. En esta tarea se comprueba si los valores de cada campo son coherentes, cuántos valores nulos hay y cómo están distribuidos, y si hay valores fuera de rango que puedan introducir ruido en el proceso. En este caso, el objetivo es garantizar la exactitud e integridad de los datos (p. 19).

3. Fase de preparación de los datos

Para la tercera fase Gallardo (2009) afirma textualmente: Una vez finalizada la recopilación inicial de datos, esta fase produce la preparación de los datos para su uso posterior utilizando técnicas de minería de datos, como herramientas de visualización de datos, búsquedas de conexiones variables y otras técnicas de exploración de datos. La preparación de los datos incluye una amplia gama de procesos de selección de datos a los que se aplicará un enfoque de modelado específico, limpieza de datos, creación de datos, integración de datos de muchas fuentes de datos y cambios de formato (p. 19).

Figura 9

Fase de preparación de los datos



Nota: La figura muestra la fase de preparación de los datos. Fuente: [CRISP-DM, 2000]

De acuerdo a lo que se observa en la (figura 9) comprende las siguientes tareas que Gallardo (2009) lo detalla a continuación de la siguiente manera:

Selección de datos.

Sobre la base de los criterios previamente establecidos en fases anteriores, en esta fase se elige un subconjunto de los datos adquiridos en la fase anterior. Estos criterios incluyen la calidad de los datos en términos de exhaustividad y exactitud, así como las restricciones sobre el volumen o los tipos de datos que son relevantes para las técnicas de gestión de datos elegidas.

Limpieza de los datos.

Debido a la variedad de enfoques que se pueden utilizar para mejorar la calidad de los datos y prepararlos para la fase de modelado, esta actividad complementa la anterior y es la que requiere más tiempo y trabajo. Para ello se pueden aplicar los siguientes enfoques: normalización de datos, discretización de campos de datos, tratamiento de valores perdidos, reducción del volumen de datos, etc.

Estructuración de los datos.

Este trabajo contiene las tareas de preparación de datos, como crear nuevos atributos a partir de los existentes, integrar nuevos registros o cambiar los valores de los atributos existentes.

Integración de los datos.

El proceso de integración de datos es la creación de nuevas estructuras de datos a partir de los datos seleccionados, como la creación de nuevos campos a partir de los existentes, la creación de nuevos registros, la combinación de tablas de campos o la creación de nuevas tablas que resumen las propiedades. de varios registros u otros campos.

Formateo de los datos.

Con el objetivo de posibilitar o simplificar el uso de una técnica específica de DM (eliminar comas, tabuladores, caracteres especiales, valores máximos y mínimos de las cadenas de caracteres, etc.), este trabajo consiste en reorganizar los campos y/o registrar la tabla. o ajustar los valores de los campos para que funcionen dentro de las limitaciones de las herramientas de modelado (pp. 20-21).

4. Fase de modelado

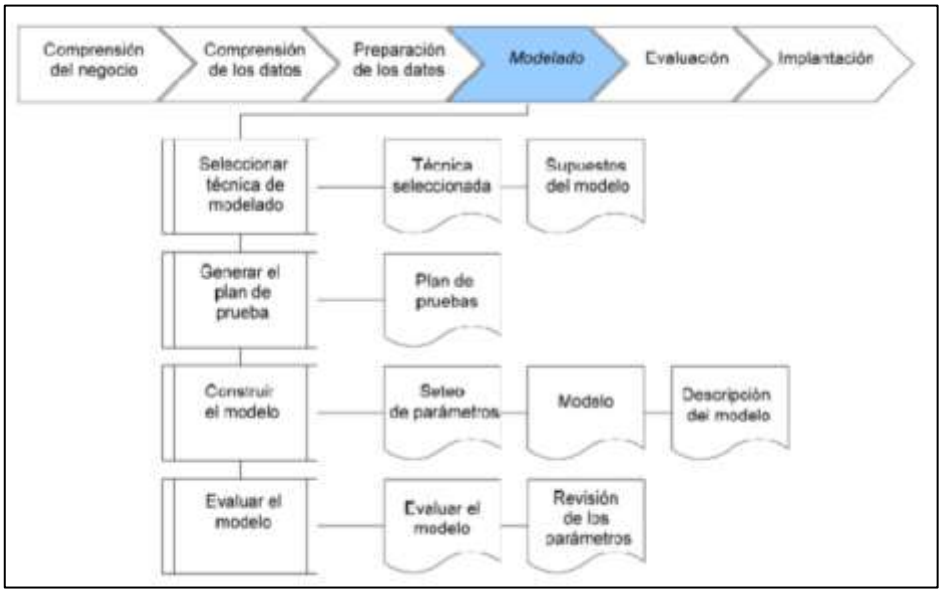
Aquí en esta fase se eligen las técnicas más adecuadas para el proyecto de minería de datos, como nos informa Gallardo (2009):

Asumiendo los siguientes criterios, se eligen estas técnicas:

- Ser apropiado al problema.
- Adecuada disponibilidad de datos.
- Cumpla las condiciones del problema.
- Tiempo sucesivo para conseguir un modelo.
- El conocimiento técnico (p. 21).

Figura 10

Fase de modelado



Nota: La figura muestra la fase de modelado. Fuente: [CRISP-DM, 2000].

Una descripción de cada de las principales tareas de esta fase Gallardo (2009) los detalla de la siquiente manera:

Antes de modelizar los datos debe elegirse una técnica de evaluación del modelo para identificar el nivel de bondad de ajuste del modelo. La creación y evaluación del modelo se realizan una vez finalizadas estas actividades generales. Los parámetros utilizados para crear el modelo vienen determinados por las propiedades de los datos y los objetivos de precisión del modelo. Las actividades y resultados de esta etapa se representan en la Figura 10. A continuación se describen las principales tareas de esta etapa:

Selección de la técnica de modelado

La elección de la mejor estrategia de minería de datos para el trabajo en cuestión se aborda especialmente en esta tarea. Para ello, es necesario tener en cuenta el objetivo principal del proyecto y su relación con las tecnologías de minería de datos.

Generación del plan de prueba

Implica crear un proceso para evaluar la precisión y eficacia del modelo construido. Para desarrollar un modelo basado en el conjunto de entrenamiento y evaluar su rendimiento en el conjunto de prueba, los datos suelen dividirse en dos partes: uno para el entrenamiento y el segundo la prueba.

Construcción del modelo

Se utiliza el programa de modelización para generar uno o varios modelos a partir de los datos previamente preparados. Para cada enfoque de modelización existe un grupo de parámetros que definen las propiedades del modelo que se creará.

Evaluación del modelo

El examen y la revisión de los parámetros del modelo constituyen este trabajo. Los profesionales de minería de datos y los especialistas en el ámbito temático pertinente realizan estas actividades y evalúan los modelos a la luz del ámbito (pp. 21-22).

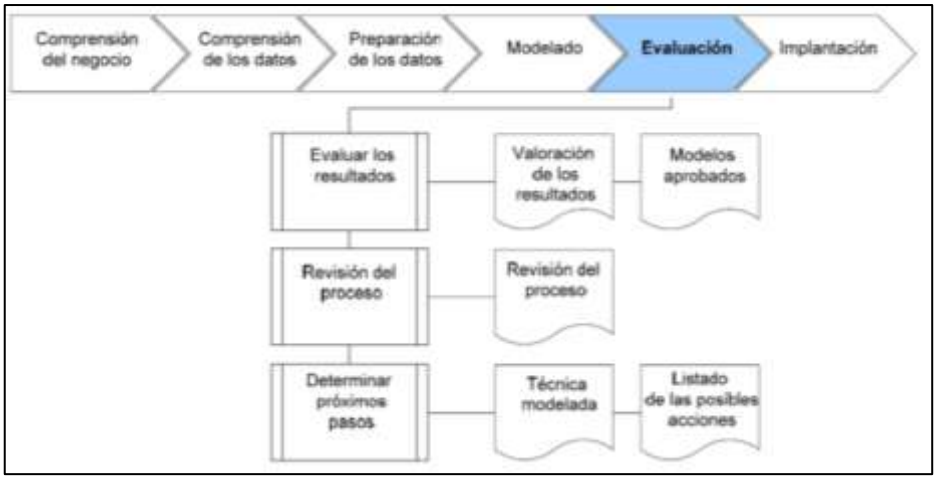
5. Fase de evaluación

De acuerdo con Gallardo (2009) nos explica lo siguiente: Esta fase analiza el modelo considerando si se han alcanzado los criterios de éxito establecidos para el problema. Además, hay que tener en cuenta que la fiabilidad determinada para el modelo sólo se refiere al conjunto de datos utilizados para la investigación.

Evaluar el proceso teniendo en cuenta los resultados obtenidos es necesario para poder replicar cualquier etapa anterior en la que se haya cometido un error. Hay que tener en cuenta que los resultados pueden interpretarse utilizando diversas técnicas (p. 22).

Figura 11

Fase de evaluación



Nota: La figura muestra la fase de evaluación. Fuente: CRISP-DM (2000)

A continuacion Gallardo (2009) nos describe las siguientes tareas de esta fase:

Evaluación de los resultados. La corrección y generalidad del modelo creado se evaluaron en los procesos de evaluación anteriores. Esta actividad consiste en evaluar el modelo en relación con los objetivos empresariales y tiene por objeto establecer si existe alguna razón empresarial por la que el modelo sea inadecuado o si, si las limitaciones de

tiempo y recursos lo permiten, es preferible probar el modelo en una situación real. ¿Es aconsejable analizar el modelo en relación con objetivos distintos de los originales, ya que ello puede revelar nueva información, además de las conclusiones directamente pertinentes para el objetivo del proyecto?

Proceso de revisión. La revisión de todo el proceso de gestión documental para encontrar áreas susceptibles de mejora se denomina proceso de revisión.

Determinación de futuras fases. La siguiente fase podría elegirse si se ha determinado que las fases hasta este punto han producido resultados suficientes; si no, puede decidirse repetir la fase de preparación de datos o de modelización con otros parámetros. Llegados a este punto, podría incluso decidirse lanzar un nuevo proyecto de DM (p. 23).

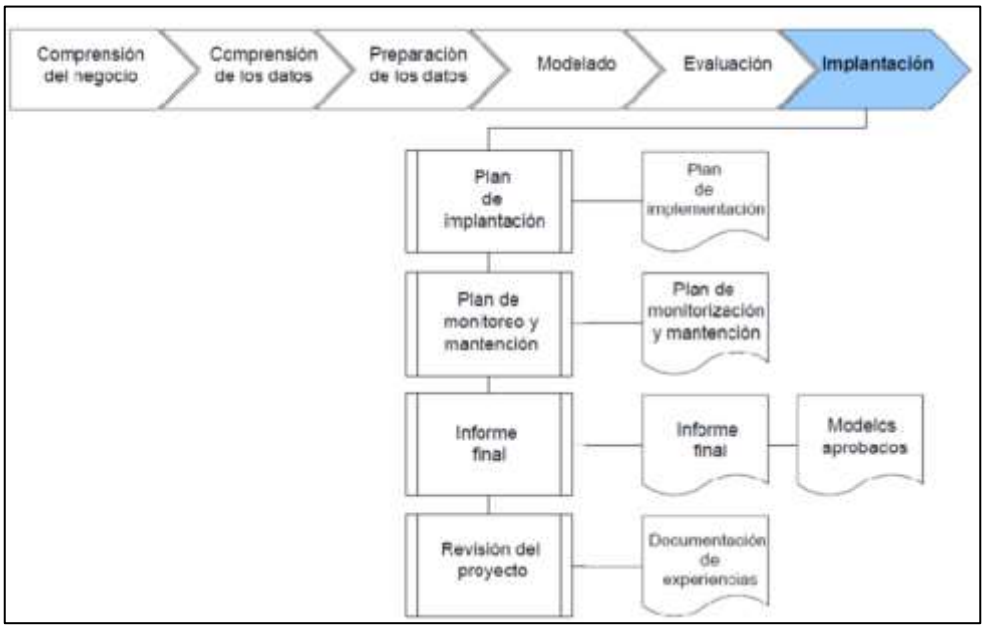
6. Fase de implementación

Cuando el modelo ha sido desarrollado y validado, en esta etapa el conocimiento aprendido se aplica a los procesos de negocio, lo que el analista recomienda o no. Acciones basadas en observaciones y resultados del modelo, aplicando el modelo a muchos conjuntos de datos o utilizando el modelo como parte de un proceso, como la aplicación, detección de fraude, etc. análisis del riesgo crediticio. (Gallardo, 2009).

En términos generales, un proyecto de Minería de Datos no termina con la implementación del modelo ya que los resultados deben ser documentados y presentados para que el usuario pueda entenderlos para lograr el objetivo. Por el contrario, durante la fase operativa se debe garantizar el mantenimiento de la aplicación y difusión de los resultados (Gallardo, 2009).

Figura 12

Fase de implementación



Nota: Fase de implementación

Fuente: CRISP-DM, 2000.De acuerdo a Gallardo (2009) las tareas que se llevan a cabo en esta fase son las siguientes:

Plan de implementación. Esta actividad utiliza los datos de la evaluación para desarrollar una estrategia de aplicación de los resultados de la gestión del cambio dentro de la organización. Si se ha encontrado un enfoque estándar para desarrollar el modelo, debe documentarse para su uso futuro

***Monitorización y mantenimiento*.** Preparar las técnicas de supervisión y mantenimiento que se utilizarán en los modelos si se despliegan en el dominio de la cuestión como parte de las operaciones cotidianas a raíz del proceso de extracción de datos. Las aportaciones de la supervisión y el mantenimiento podrían mostrar si el modelo se está aplicando correctamente.

Informe final. El proyecto de MD ha llegado a su fin. En este informe puede incluirse una presentación final que contenga y analice los logros del proyecto, dependiendo de la estrategia de ejecución, o puede tratarse simplemente de una recapitulación de los principales logros del proyecto.

Revisión del proyecto. Ahora evaluamos lo que ha ido bien y lo que no, lo que se ha hecho correctamente y lo que hay que cambiar (pp. 23-24).

3.2.1.6. Técnicas para evaluar clasificadores

Para los clasificadores binarios y multiclase, se dispone de métricas de rendimiento que incluyen la precisión, el error de clasificación y el coeficiente kappa. Estas mediciones permiten evaluar varios métodos de categorización y elegir el más preciso. En este estudio, el error de clasificación, la precisión y el coeficiente kappa se utilizan para proporcionar un método de evaluación.

Matriz de confusión

Se trata de una tabla de doble entrada que muestra las distintas clases de la variable objetivo según la predicción del clasificador propuesto y según las clases observadas (reales). En la Tabla 1 se muestra una matriz de confusión para dos clases:

Tabla 1

Matriz de confusión

Clasificación Observada	Clasificación Predicha		Total observado
	Positivo (Clase 0)	Negativo (clase1)	
Positiva clase(0)	VP	FN	VP+FN
Negativa (clase1)	FP	VN	FP+VN
Total Predicho	VP+FP	FN+VN	N

Nota: En la tabla se muestra como debe quedar la clasificación de la predicción (Holgado Apaza, 2018)

De acuerdo a Holgado (2018) nos decribe cada uno de los valores mencionados en la tabla de la siguiente manera:

El número de observaciones que el clasificador predijo con precisión como pertenecientes a las clases positiva y negativa está representado por las letras VP y VN, respectivamente. FP significa falsos positivos y FN, falsos negativos. Este es el número de observaciones que el clasificador clasificó erróneamente como pertenecientes a la clase positiva y a la clase negativa, respectivamente. Esta tabla se puede utilizar para calcular el error y la exactitud de la clasificación.

$$Exactitud = \frac{VP + VN}{N}$$

Este valor mide cuántas observaciones clasificó correctamente el modelo de clasificación predictiva.

$$Tasa\ de\ error = \frac{FP + FN}{N}$$

Este valor representa la tasa de error del modelo de clasificación predictiva, es decir, la proporción de observaciones que fueron clasificadas incorrectamente.

Donde: $N = VP + VN + FP + FN$. (p. 36)

Coeficiente de Kappa (k)

En esta parte tenemos a Cerda & Villarroel (2008) que nos indican que la concordancia inter-observador se refleja por el coeficiente kappa, que se puede calcular en tablas de cualquier dimensión, siempre y cuando se contrastan dos observadores (el coeficiente kappa de Fleiss se utiliza para la evaluación de concordancia de tres o más observadores, cuya explicación supersigue el propósito de este artículo). El coeficiente Kappa puede tener valores entre -1 y +1. Por el contrario, cuanto más cerca esté un observador de +1, mayor será el grado de acuerdo entre los observadores; por el contrario, cuanto más cerca esté un observador de -1, mayor será el grado de desacuerdo entre los observadores.

En la Tabla 2, se puede observar la valoración del valor de k que es propuesta por (Landis and Koch, 1977)

Tabla 2

Valoración del coeficiente de kappa (Landis y Koch, 1977)

Coeficiente de kappa	Fuerza de concordancia
0,00	Pobre (Poor)
0,01-0,20	Leve (Slight)
0,21-0,40	Aceptable (Fair)
0,41-0,60	Moderada (Moderate)
0,61-0,80	Considerable (Subtantial)
0,81-1,00	Casi perfecta (Almost perfect)

Fuente: (Cerda & Villarroel, 2008)

3.2.2. RENDIMIENTO ACADEMICO

El rendimiento académico permite medir el resultado del alumno al finalizar el ciclo, en este sentido Reyes (2003) nos dice sobre el rendimiento academico lo siguiente: El sistema educativo considera el rendimiento académico como una medida del nivel de aprendizaje logrado por los estudiantes, por lo tanto, le asigna una gran importancia. De esta manera, el rendimiento académico se convierte en un “referente imaginario” del aprendizaje alcanzado en el aula, que es el objetivo primordial de la educación. Sin embargo, muchos otros factores externos, como la calidad del profesorado, el ambiente del aula, la familia, el programa educativo, etc., así como factores psicológicos o internos,

como la actitud del estudiante ante las tareas, la inteligencia, la personalidad, el autoconcepto, la motivación, etc, también juegan un papel en el rendimiento académico.

Muchos estudios también equiparan el rendimiento académico con el GPA y las calificaciones obtenidas en los cursos. La capacidad de un estudiante para cumplir con los requisitos establecidos para los cursos ofrecidos en una universidad y superar estos desafíos se reflejará en las calificaciones del estudiante y en su capacidad para ser considerado un estudiante exitoso. Por otro lado, quienes no demuestren los conocimientos o habilidades necesarios para cumplir con los requisitos antes mencionados tendrán un bajo rendimiento académico, lo que podría conducir al abandono escolar (Yamao, 2018).

Muchos factores pueden afectar el éxito del desempeño académico de un estudiante. De esta forma, por ejemplo, Rasberry, Lee, Robin, Laris, Russell, Coyle y Nihiser (2011) agrupan los resultados del aprendizaje en tres posibles partes: (1) Rasgos y habilidades cognitivas, que incluyen, entre otras cosas, atención, memoria, comprensión lingüística, procesamiento de información, motivación, autoimagen y satisfacción; (2) comportamiento de aprendizaje, incluyendo, entre otras cosas, organización, planificación y asistencia; y (3) logro académico, incluidas las calificaciones en las materias enseñadas (Yamao, 2018).

Según Beneyto (2015) señala que el: En la actualidad, uno de los temas candentes de la investigación educativa es el examen del rendimiento de los alumnos. El principal problema de la educación en la cultura actual, que se define por el constante bombardeo de información procedente de diversas fuentes, es convertir este vasto volumen de información en conocimientos fiables y prácticos para poder vivir con éxito. Así pues, el éxito o el fracaso académico de una persona tiene un impacto significativo en su capacidad para encontrar empleo en el futuro.

Aunque la regla binómica éxito-fracaso se refiere a una regla general sin tener en cuenta ocasionalmente el proceso evolutivo y las individualidades únicas de cada alumno, la realidad es que el éxito en la escuela está correlacionado con el rendimiento académico, mientras que el fracaso está correlacionado con un bajo rendimiento académico (p.15).

Los resultados del aprendizaje se dividen en dos partes desde un punto de vista práctico, la tendencia más común es equiparar el logro con los resultados, distinguiendo entre los dos: inmediato y diferido. En el caso de la educación superior, el primero vendrá determinado por los diplomas obtenidos por el estudiante durante sus estudios hasta la recepción del diploma correspondiente. La segunda trata sobre el impacto de la formación que reciben los egresados en la vida social; es decir, la utilidad de estos estudios en la integración de los egresados universitarios al mercado laboral. Estos dos criterios, también conocidos como logros internos y externos, son estándares de uso común para evaluar el desempeño académico en la educación superior (Candia, 2019).

El artículo aborda el hecho de que los resultados del aprendizaje se pueden medir a partir de los puntos que reciben los estudiantes durante su carrera académica, lo que se conoce como categoría "instantánea", y el resto del éxito o fracaso en la integración laboral. Para mí lo hace "diferido". (Candia, 2019).

Factores a tener en cuenta para determinar el rendimiento académico

Para elestudio de los factores que vamos a tener en cuenta tenemos a González (1989), quien describe tres factores en su investigación:

1. Factores inherentes al alumno

a) Poca preparación para ingresar a cursos avanzados o niveles de conocimiento que son insuficientes para las demandas de la universidad.

b) Desarrollo inadecuado de competencias especializadas acordes con la trayectoria profesional elegida.

c) El actitudinal de índole

d) Ausencia de métodos de estudio o técnicas de trabajo intelectual

e) Los enfoques de aprendizaje no coinciden con la opción de carrera profesional

2. Factores inherentes con el profesor

a) Deficiencias pedagógicas

b) Tratamiento individualizado inadecuado

c) Dedicación insuficiente

3. Factores inherentes a la organización académica universitaria

a) Falta de objetivos claramente definidos.

b) Falta de coordinación entre diversas materias.

c) Uso de sistemas de selección.

d) Los criterios objetivos de la evaluación. (p. 14)

En resumen, el rendimiento de los estudiantes depende en gran medida de la universidad a la que asisten, los profesores de esa escuela y, especialmente, la capacidad del estudiante, y esto se complica aún más por una serie de factores, factores que determinan el éxito o el fracaso de los estudiantes. , en este estudio, nos centraremos principalmente en la información sobre los estudiantes al observar únicamente sus datos de admisión o admisión a la universidad, como se describirá más adelante (Candia, 2019).

Sobre el rendimiento académico también: (Tejedor, 1998) identifica dos tipos de desempeño académico: el primero es el desempeño estricto, medido mediante la presentación de exámenes o el logro de un cierto nivel de éxito en pruebas (calificaciones); el segundo es el desempeño amplio, medido por alcanzar un cierto nivel de éxito (finalización) o por retroceder o abandonar los estudios. El término "regularidad académica" también se utiliza para describir la operacionalización de la noción de rendimiento académico mediante tasas de presentación o sin convocatorias de exámenes. (Tejedor, 1998, citado por Candia, 2019, p. 15)

Cabe señalar, sin embargo, que el rendimiento académico, estrictamente hablando, analizado en muy pocos estudios de nivel universitario, es más importante que otras medidas. Esto parece razonable considerando el hecho de que las calificaciones más bajas son menos relevantes para la deserción y más relevantes para las calificaciones escolares a la hora de determinar el rendimiento de los estudiantes (Candia, 2019).

Los conceptos y definiciones de aprendizaje son un poco complicados de definir en base a la evidencia encontrada, pero la mayoría de los autores y trabajos se refieren al aprendizaje como multidimensional y multifactorial que el aprendizaje Los estudiantes lo consideran un desafío en el proceso de enseñanza y aprendizaje, a menudo relacionado con las calificaciones que se han obtenido (CANDIA, 2019).

Dada la complejidad y controversia de determinar el rendimiento de los estudiantes, se puede argumentar que el rendimiento académico es una medida de la capacidad de un

estudiante para aprobar o reprobar en la vida académica, en términos de las calificaciones que recibe del proceso de enseñanza y aprendizaje (Candia, 2019).

3.3 Bases conceptuales

Análisis de datos: la aplicación de diversas técnicas de exploración de datos para alcanzar determinados objetivos Visualización, correlación, asociaciones, análisis factorial, segmentación, secuencias y series temporales son algunas técnicas analíticas (Nettleton, 2003).

Árboles de decisión: Un modelo de predicción utilizado en inteligencia artificial se denomina árbol de decisión. Estos diagramas de construcción lógica, que se crean utilizando una base de datos y se asemejan a los sistemas de predicción basados en reglas, se utilizan para expresar y clasificar una secuencia de circunstancias que deben darse simultáneamente para resolver un problema.

Clasificación. Método de análisis de datos que permite la extracción de modelos que describen las clases significativas de los datos.

Cross-Industry Standard Process for Data Mining (CRISP-DM) Esta metodología y modelo de proceso establece un marco para el ciclo de vida del desarrollo de la minería de datos, que esboza un ciclo de seis actividades clave.

Ingresante. Un estudiante que ha sido admitido en la universidad y está matriculado en su primer año de estudios.

Medida de concordancia: El error de medición provoca variabilidad, por lo que uno de los objetivos de los estudios de fiabilidad es determinar cuánta variabilidad existe. Para ello se emplea el índice kappa.

Minería de datos: Un sistema informático de información que busca en grandes conjuntos de datos para proporcionar información y aprender cosas nuevas.

Modelo de datos: Para construir un modelo de los datos, empleamos una serie de técnicas. Una variable de salida que indique el objetivo comercial (comprar sí o no) y una serie de variables de entrada (edad, estado civil, saldo medio, etc.) son los componentes típicos de un modelo. Se puede utilizar la regresión para construir un modelo de datos; puede ser lineal para las tendencias que son lineales, no lineal para las que son no lineales, o logística para los resultados que son binarios por naturaleza.

Asimismo, podemos modelizar con métodos de «aprendizaje automatizado», como la «Inducción de Reglas» o «Redes Neuronales».

Predicción: El objetivo es modelizar datos anteriores con un resultado conocido para poder prever resultados futuros. Es esencial que el entorno de los datos históricos y el

momento futuro que se quiere pronosticar no cambien drásticamente. La salida de un modelo predictivo es el resultado real, y sus entradas incluyen una serie de variables elegidas por su fuerte asociación con resultados anteriores. La inducción de reglas, las redes neuronales y la regresión son algunos ejemplos de métodos utilizados para desarrollar modelos de predicción.

Red neuronal: método de análisis de datos que construye modelos predictivos de datos basados en componentes enlazados (neuronas) que se asemejan a su equivalente biológico. Aunque produce modelos "opacos" con una estructura subyacente ininteligible, es muy adaptable a los datos y robusto frente al "ruido" (errores, escasa importancia de algunas de las variables), a diferencia de la inducción de reglas.

Rendimiento académico: es una prueba que evalúa la capacidad del alumno y expresa lo que ha aprendido durante su trayectoria académica. Además, sugiere que el alumno es capaz de reaccionar a las aportaciones de la instrucción. De este modo, aptitud y éxito académico están relacionados.

CAPÍTULO IV. MARCO METODOLÓGICO

4.1 Ámbito

UBICACIÓN Y DESCRIPCION GEOGRÁFICA

El estudio se realizo en el país de Perú, departamento de Huánuco, provincia de Leoncio Prado, distrito de Rupa-Rupa, ciudad de Tingo María.

La ciudad de Tingo María está situada en el departamento de Huánuco, en el centro-este de Perú, entre la sierra andina y la selva amazónica, a 135 kilómetros de la ciudad de Huánuco, en la margen derecha del río Huallaga. Aquí se encuentra el Parque Nacional de Tingo María, famoso por la Bella Durmiente, una montaña con forma de dama dormida, y por poseer una rica biodiversidad. Además, en las cercanías se encuentra la Cueva de las Lechuzas, que alberga guácharos.

Extensión y tamaño del área: Tiene una superficie de 4,395.46 km2.

Localización geográfica. Exactamente en el centro del Perú, en el punto más septentrional del departamento de Huánuco, se encuentra la provincia de Leoncio Prado. La ciudad de Tingo María es la capital de la provincia y está situada entre los 75° 53' 00" de longitud oeste y los 09° 18' 00" de latitud sur. Está rodeada al norte por los distritos de Nuevo Progreso, Tocache (departamento de San Martín) y Cholón (provincia de Marañón), al sur por los distritos de Chinchao y Churubamba, al este por la provincia de Padre Abad, capital de Aguaytia (departamento de Ucayali), mientras que por el Oeste, con las provincias de Dos de mayo, Huamalies, Marañón y Huacaybamba.

Clima: Es cálido y húmedo (tropical), su temperatura promedio es de 24ºC.

Altitud: 647 msnm

4.2 Tipo y nivel de investigación

Por el tiempo es una investigación longitudinal.

El enfoque de investigación fue cuantitativo, del tipo explicativo y no experimental.

Los estudios correlaciones vinculan variables con un patrón predeterminado para una población o grupo. El objetivo de este tipo de estudio es determinar el grado de asociación o relación entre dos o más conceptos, categorías o variables en un contexto específico (Hernandes, 2004). Estos estudios proporcionan predicciones y explican las relaciones

entre las variables, que es exactamente lo que desarrollaremos para "predeterminar el rendimiento académico".

Nivel de investigación. Por naturaleza del estudio el nivel de investigación es explicativo.

4.3 Población y muestra

4.3.1 Descripción de la población

Para la presente investigación la población en estudio es todos los alumnos ingresantes a la UNAS teniendo en cuenta el semestre 2015-I hasta el 2019-I.

4.3.2 Muestra y método de muestreo

En el presente estudio trabajará y utilizará toda la población para buscar patrones de grandes cantidades de datos.

4.3.3 Criterios de inclusión y exclusión

Criterios de inclusión

- Alumnos ingresantes de cada año desde el 2015 hasta el 2019.
- Alumnos con datos completos llenados en las encuestas previas a la matricula o la postulación.

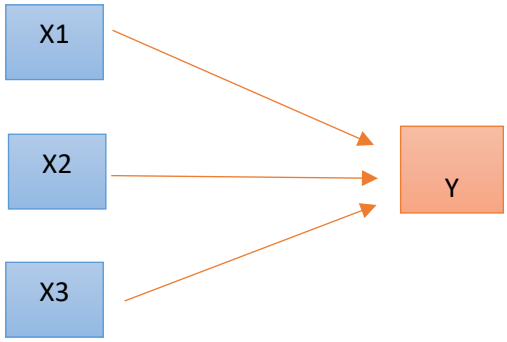
Criterios de exclusión

- Alumnos con datos incompletos en las encuestas.
- Registros repetidos por haber ingresado a dos carreras distintas por diferentes modalidades.

4.4 Diseño de investigación

Para nuestro estudio el diseño de la investigación es transeccional no experimental. Busca describir la relación que existe entre el rendimiento académico y los factores sociales, económicos y académicos de los examinados que se han utilizado para probar la validez de la predicción del resultado del aprendizaje.

Para nuestro estudio se usó el siguiente diseño:



En donde:

x1: Variable de indicadores sociales

x2: Variable de indicadores económicos

x3: Variables de indicadores académicos

Y: Rendimiento académico de los alumnos ingresantes

4.5 Técnicas e instrumentos

4.5.1 Técnicas

Se va a extraer los datos históricos almacenados de la Dirección de Coordinación y Desarrollo Académico (DICDA), y de la oficina de admisión donde se almacenan las encuestas de los postulantes al examen de admisión, en la (DICDA) se tiene las notas del rendimiento académico del primer semestre las cuales se normalizan teniendo en cuenta las teorías consideradas en las bases teóricas para usarlo posteriormente, tales como la preparación y depuración de la data.

Para el desarrollo se aplicarán tecinas de extracción, transformación y posterior carga de los datos y convertirlos en una base de datos transaccional para poder aplicar las técnicas de minerías de datos.

4.5.2 Instrumentos

Para recolectar los datos realizo un análisis documental a la base de datos, con una ficha de análisis documental el cual se encuentra el en anexo 02.

4.5.2.1 Validación de los instrumentos para la recolección de datos. El instrumento ha sido validado por 5 expertos los mismos que se encuentran en el anexo 03.

4.5.2.2 Confiabilidad de los instrumentos para la recolección de datos. La confiabilidad se realizo en el software SPSS obteniéndose un resultado de 0.812 el cual nos indica que el instrumento es muy bueno.

4.6 Técnicas para el procesamiento y análisis de datos

Para el análisis e interpretación de la información extraída de la base de datos de la Dirección de Coordinación y Desarrollo Académico sobre los estudiantes que ingresaron, primero se verifica la información generada y luego se agrega o se eliminan los registros que muestran inconsistencias, valores nulos o incompletos, segundo, los datos se transforman para su procesamiento, tercero, los resultados son analizados, comprendidos y explotados, finalmente interpretados mediante gráficos y matrices, herramientas como Power Pivot, Power Query, Microsoft Excel, así como la herramienta WEKA 3.9 para aprendizaje automático.

4.7 Aspectos éticos

Se tendrán en cuenta las principales cuestiones éticas asociadas a este tipo de investigaciones, teniendo en cuenta la confidencialidad de los datos de los estudiantes de pregrado de la Universidad Nacional Agropecuaria. Previamente autorización y consentimiento son necesarios para usar y divulgar información de identificación personal.

En el proceso de mantener el anonimato de sus datos, se asignarán identificadores únicos a cada uno de sus registros de datos y se eliminarán atributos que permitirán identificar a las personas en diferentes registros, como nombre, apellidos, dirección entre otros que puedan invadir su privacidad. De esta manera, se garantiza la validez de los resultados de las pruebas sin revelar información personal del estudiante.

CAPÍTULO V. RESULTADOS Y DISCUSIONES

5.1.ANÁLISIS DESCRIPTIVO

Vamos a desarrollar usando la metodología CRISP-DM de acuerdo a las fases y las tareas que propone.

5.1.1 Comprensión Del Negocio

Aquí es la fase inicial nos enfocaremos en entender los objetivos y requerimientos del proyecto para esta fase vamos a detallar cada uno de los pasos:

5.1.1.1. ***Determinación del objetivo de negocio***. La universidad Nacional Agraria de la Selva es una comunidad integrada por alumnos, docentes, graduados y trabajadores administrativos. Dentro de la misión de la UNAS es crear y transferir conocimientos científicos, tecnológicos y humanísticos a los estudiantes, formando profesionales holísticos; socialmente responsables y comprometidos con el crecimiento sostenible y competitivo de la nación. Dentro de las políticas de las UNAS y objetivo prioritario es mejorar la calidad de la formación profesional en la calidad de enseñanza y aprendizaje. Sus programas de estudio con acorde a la región. Entre sus objetivos esta mejorar la calidad académica de los alumnos y egresados.

Con el presente trabajo pretendemos predecir el rendimiento académico de estudiantes de primer semestre mediante la estimación de indicadores socioeconómicos y académicos utilizando datos recopilados en las oficinas de admisión. Esto permitirá a las autoridades pertinentes tomar acciones para mejorar el rendimiento académico y así prevenir fracasos y deserciones académicas.

5.1.1.2. ***Evaluación de la situación***. Dentro de la Universidad tiene un problema que muy pocos egresan de la universidad debido a mucha deserción estudiantil, por diferentes motivos los cuales no son todos conocidos, sobre todo esto se da durante los primeros años de estudio en la universidad, lo que hace que la planificación de la UNAS muchas veces se vea afectada.

Los datos de los alumnos son obtenidos desde el momento de la inscripción en la oficina de admisión como postulante o ingresante por el centro Preuniversitarios.

Los cuales contienen datos como apellidos y nombres, edad, sexo, tipo de colegio que egreso, lugar de procedencia, con quienes vive, de quien depende económicamente. También aquí se obtuvieron el puntaje de ingreso y la modalidad por al que ingresaron.

Las notas de los ingresantes por el centro preuniversitario no se registraron en la oficina de admisión, por lo que estos datos se encontraron dentro de las oficinas de los archivos centrales.
Todos los atributos han sido almacenados en Excel y se encuentran centralizados en la oficina de admisión.

Los recursos con los que se dispone para el desarrollo son los siguientes:

- **Los materiales**: Weka versión 3.9.5 se usará como herramienta para la minería de datos también los datos de Excel y el spss para los cruzar los datos de ser necesarios.
- **Recursos humanos:** el autor de la investigación.
- **Datos de trabajo**: los alumnos ingresantes desde el 2015 al 2018 que hacen un total de 2402 registros con sus atributos. Y ingresantes del 2019 para la validación del modelo de aprendizaje automático.

5.1.1.3. ***Determinación de los objetivos de minería de datos***. Dar apoyo a través de tecinas de minería de datos a los objetivos de la investigación es el objetivo de la minería de datos.

Objetivos específicos de la investigación

Realizar estudios estadísticos de los datos.

Encontrar el rendimiento académico de los alumnos mediante los ámbitos académicos, económicos y sociales.
Conocer estos objetivos permitirá tener muy claro el inicio de una planificación estratégica en cada inicio de ciclo.

Se aplicarán lo siguientes pasos para los objetivos de la minería de datos:

- Realizar un análisis de los datos y la limpieza de los datos proporcionada por las oficinas de admisión.
    - Realizar los análisis descriptivos para determinar qué factores son más importantes para predecir el rendimiento académico.
    - Elección de las técnicas de minería de datos que se ajuste al problema que deseamos resolver.

- Analizar y determinar los algoritmos de minería de datos que mejor predicen el rendimiento académico.
- Evaluación de los modelos obtenidos al aplicar los diferentes algoritmos de predicción.
- Comprobación y validación de los modelos obtenidos.

Para realizar la predicción primero hay que realizar el análisis estadístico de los datos para determinar los factores sociales, económicos y académicos, para luego usar la herramienta WEKA y usar los algoritmos de predicción y seleccionar y elegir al mejor algoritmo predictor.

5.1.2 COMPRENSIÓN DE DATOS

5.1.2.1. ***Recolección de datos iniciales***. Los datos que se necesitan están en la oficina de admisión de la UNAS, también se recolecto los datos de las oficinas de la Dirección de asuntos académicos (DICCA) donde se encuentra el promedio ponderado del semestre correspondiente al primer ciclo.

Las notas de los ingresantes por el centro preuniversitario se tuvieron que recolectar en las mismas oficinas del archivo central.

Los datos recolectados en las oficinas de admisión fueron:

Estos datos se recolectaron en hoja de cálculo Excel conteniendo los siguientes campos:

CODIGO

Apellidos y Nombres

OPCION 1

OPCION 2

MODALIDAD

DNI

CODSEDE

INSCRIPCION

UBIGEO PROCEDENCIA

CODCOLEGIO

FECHA EGRESOCOLEGIO

TIPOCOLEGIO

UBIGEO COLEGIO

ESTADO CIVIL

ENCUESTA

INGRESO

INGRESO A

SEXO

NOMBRE

COLEGIO

IDIOMA MAT

TELCELULAR

DIRECCION

UBIGEO

FECNAC

NOTAAC

NOTACO

RESPUESTA

Figura 13

Datos recolectados en la oficina de admisión

	OPCION 2	MODALIDAD	DNI	CODSEDF	INSCRIPCION	UBIGEO PROCEDENCIA	CODCOLEGIO	FECHA EGRESOCN	TIPOCOLEGIO	UBIGEO COLEGIO	ESTADO CIVIL	ENCU
2	AGRONOMIA	Centro Pre Universitario	76070695	1	27/03/2015	005100000010060100000	48	12/01/2014	2	510000001006010000.00	S	1421
3	AGRONOMIA	Exámen Ordinario	47958209	1	23/02/2015	005100000010050700000	9999	12/01/2010	2	510000001001020000.00	S	1233
4	ECONOMIA	Exámen Ordinario	74892660	1	20/03/2015	005100000010060100000	61	12/01/2014	1	510000001006010000.00	S	3113
5	INGENIERIA E	Centro Pre Universitario	71387822	1	03/06/2015	005100000010060100000	63	12/01/2014	1	510000001006010000.00	S	3211
6	INGENIERIA E	Exámen Ordinario	77659378	1	20/03/2015	005100000010060100000	4	12/01/2003	2	510000001006010000.00	S	3243
7	INGENIERIA E	Convenios Especiales	74944883	3	03/04/2015	005100000010060600000	9999	12/01/2012	2	510000001006060000.00	S	2211
8	ADMINISTRAI	Exámen Ordinario	48453953	1	19/03/2015	005100000010060100000	48	12/01/2011	2	510000001006010000.00	CO	3131
9	ADMINISTRAI	Exámen Ordinario	71726259	1	03/11/2015	005100000010060100000	47	12/01/2014	2	510000001006010000.00	S	3111
10	ADMINISTRAI	Deportista Calificado	71726259	1	03/04/2015	005100000010060100000	47	12/01/2014	2	510000001006010000.00	S	3131
11	INGENIERIA E	Exámen Ordinario	48248134	1	20/03/2015	005100000010080300000	54	12/01/2011	1	510000001006010000.00	S	1141
12	ADMINISTRAI	Exámen Ordinario	76452372	1	20/03/2015	005100000010060100000	48	12/01/2013	2	510000001006010000.00	S	3111
13	ECONOMIA	Exámen Ordinario	74041064	1	19/03/2015	005100000021010100000	9999	12/01/2014	2	510000002101010000.00	S	3241
14	INGENIERIA A	Exámen Ordinario	71218195	1	17/03/2015	005100000012060100000	9999	12/01/2014	2	510000001206010000.00	S	3111
15	INGENIERIA F	Exámen Ordinario	74482776	1	17/03/2015	005100000010090500000	9999	12/01/2014	2	510000001009050000.00	S	3211
16	CONTABILID	Centro Pre Universitario	72098573	1	02/09/2015	005100000022100100000	9999	02/09/2013	2	510000002210010000.00	S	3121
17	ECONOMIA	Centro Pre Universitario	75351816	1	14/01/2015	005100000022010500000	9999	12/01/2012	2	510000002201050000.00	S	3221
18	INGENIERIA E	Exámen Ordinario	75162239	1	19/03/2015	005100000012030300000	9999	12/01/2013	2	510000001203030000.00	S	3141
19	INGENIERIA E	Centro Pre Universitario	76068571	1	20/03/2015	005100000010060100000	56	12/01/2014	2	510000001006010000.00	S	3211
20	INGENIERIA F	Exámen Ordinario	73194127	1	20/03/2015	005100000010010100000	47	12/01/2013	2	510000001006010000.00	S	1121
21	INGENIERIA E	Exámen Ordinario	76265642	1	20/03/2015	005100000010060100000	48	12/01/2014	2	510000001006010000.00	S	3123

Nota. La figura muestra los datos recolectados en las oficinas de admisión.

Figura 14

Datos recolectados en la oficina de admisión

	UBIGEO COLEGIO	ESTADO CIVIL	ENCUESTA	INGRESO	INGRESO A	SEXO	NOMBRE COLEGIO	IDIOMA MAT	TELCELULAR	DIRECCION	UBIGEO	FECNAC	NOTAAC	NOTACO	RESP
2	510000001006010000.00	S	142111213	1	AGRONOMIA	M		E	994446735		005100000010	13/11/1997			
3	510000001001020000.00	S	123331241	0		M	MARINO ADE	E	968155906	SVEEN ERICK	005100000010	23/08/1993	24.25	29.5	ABE
4	510000001006010000.00	S	311311212	1	ADMINISTRA	F	CIENCIAS	E	954529179	JR. ELIAS MA	005100000010	21/08/1997	24.25	38.25	EBB
5	510000001006010000.00	S	321111211	1	INGENIERIA	M	CIMAFIQ	E	974044931	JR. MONZON	005100000010	20/07/1998			
6	510000001006010000.00	S	324313311	0		M	GOMEZ ARIA	E	970151062	Jr. GARCILAZ	005100000010	14/09/1996	21.25	26.25	ADE
7	510000001006060000.00	S	221111313	1	INGENIERIA	F		E	991277971		005100000010	20/05/1995			
8	510000001006010000.00	CO	313141451	1	CONTABILID	F	GOMEZ ARIA	E	961653633	MERCEDES A	005100000010	19/08/1999	24.25	39.25	AEE
9	510000001006010000.00	S	311111212	0		F	MARISCAL R	E	978557492	Alberto Fujin	005100000010	29/10/1997	22.25	29.25	ACC
10	510000001006010000.00	S	313111212	0		F	MARISCAL R	E	978557492	Alberto Fujin	005100000010	29/10/1997			
11	510000001006010000.00	S	114111313	0		M		E	62561101		005100000010	03/10/1984	25.25	11.75	ECD
12	510000001006010000.00	S	311111311	0		F	GOMEZ ARIA	E	930960245	JR. CHICLAYO	005100000010	09/12/1997	23.25	31.25	AEB
13	510000002101010000.00	S	324111341	1	ADMINISTRA	M	MANUEL FID	E	948965245	AV. LA FLORI	005100000010	10/01/1997	28.25	37.25	ABA
14	510000001206010000.00	S	311113111	1	INGENIERIA	F	RAFAEL GAS	E	990352046	Calle las mag	005100000010	28/05/1998	28.25	32.5	ACB
15	510000001903040000.00	S	321113122	0		M	JULIO VERA	E	942794198	CASERIO PAI	005100000010	13/04/1999	23.25	27.25	AAD
16	510000002210010000.00	S	312111211	1	CONTABILID	M	I.E. N\272 04	E	966531061	JR. SAN MAR	005100000010	29/09/1997			
17	510000002201050000.00	S	322111411	1	ECONOMIA	F	ALFREDO TEJ	E	927904926	JR. LAS FLOR	005100000010	04/02/1996			
18	510000001203030000.00	S	314111241	1	AGRONOMIA	F		E	929025287		005100000010	27/05/1996	24.25	32.5	EDO
19	510000001006010000.00	S	321111311	1	INGENIERIA	F		E	962988684		005100000010	26/05/1998			
20	510000001006010000.00	S	112111211	1	INGENIERIA	F	RAMON CAS	E	930288054	castillo gran	005100000010	08/02/1995	23.25	32.25	CEA
21	510000001006010000.00	S	312334231	0		F	GOMEZ ARIA	E	962086625	TUPAC AMA	005100000010	12/05/1997	20.25	34	CAB

Nota. Continuación de los datos recolectados en las oficinas de admisión.

Los datos recolectados en DICCA también se recolecto en Excel con los siguientes atributos:

TDOCUMENTO

CODALUMNO

PATERNO

MATERNO

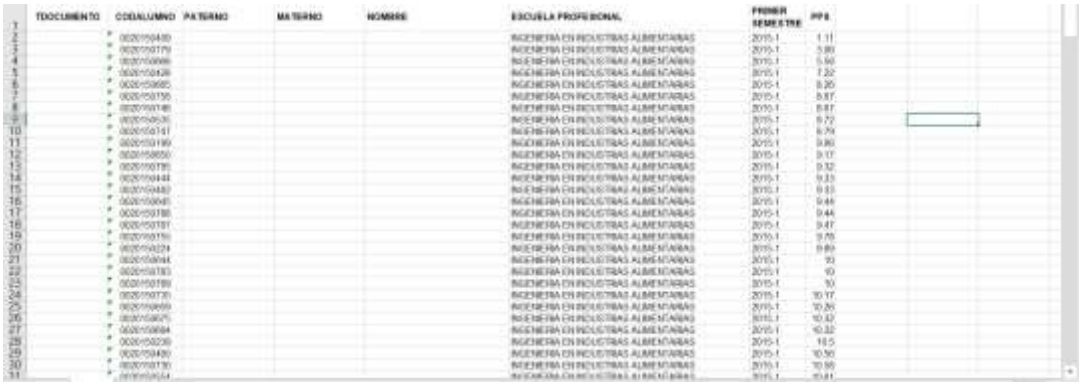
NOMBRE

ESCUELA PROFESIONAL PRIMER SEMESTRE

PPS

Figura 15

Datos recolectados en las oficinas de DICCA



Nota. Aquí se recolecto el promedio final del semestre.

Para no violar la ética de datos no se muestra los datos personales de los alumnos en los registros.

5.1.2.2. ***Describir los datos***. A continuación, describimos cada uno de los campos de la tabla, así como cada uno con sus tipos de datos.

- **Código.** Código único que se asigna a cada postulante para rendir el examen de admisión.
- **Apellidos y Nombres.** De cada alumno esta su apellido y nombre.
- **Opcion1.** Es la primera opción a la que postula el estudiante.
- **Opcion2.** A cada estudiante se le da la oportunidad de elegir dos opciones de carrera.

- **Modalidad.** Nos permitirá obtener la modalidad por la cual el estudiante logro ingresar a la universidad, puede ser por admisión, preuniversitario, primeros puestos, deportistas u otros.
- **DNI.** Documento Nacional de identidad de los estudiantes, será nuestro identificador para relacionar con los datos de la oficina de DICCA.
- **Codsede.** Es el código de la sede donde se tomó el examen de admisión o donde realizaron la preparación pre preuniversitaria.
- **Inscripción.** Es la fecha que se inscribieron para el examen de admisión que restado con la fecha de nacimiento nos permitirá obtener la edad a la que ingresaron a la universidad.
- **Ubigeo de procedencia.** Esto nos permitirá obtener el departamento de donde viene el estudiante.
- **Fecha de egreso del colegio.** Nos brinda la información del año en que el estudiante termino la secundaria.
- **Tipo de colegio.** Se expresa como colegio particular o estatal.
- **Ubigeo de colegio.** La ubicación del distrito, provincia y departamento donde está el colegio en el que culmino los estudios secundarios.
- **Estado civil.** La condición del estudiante si es soltero, casado u otro.
- **Encuesta.** Estas encuestas rellenan los estudiantes al momento de inscribirse y contra de 9 preguntas que serán de mucha importancia para nuestro estudio de predicción, a continuación, se detalla cada uno de las siguientes preguntas:

  C1. Tipo de preparación para su postulación

  1) Autoestudio

  2) Profesor particular.

  3) Academia

  4) Otros

  C2. Como se enteró de las fechas de nuestro concurso de admisión.

  1) Charlas dadas por el personal de la UNAS en el colegio

  2) Familiares y amigos

  3) Radio

  4) Televisión

  5) Internet

  6) Otros

C3. ¿Cuál fue el principal motivo por el que se animó a postular a la UNAS?

1) Prestigio.

2) Nivel académico

3) La carrera que deseo no la ofrecen en otra universidad.

4) Recomendación de familiares o amigos

5) Económico

6) Servicios que ofrece comedor e internado

7) Otros

D1. ¿Trabaja?

1) No

2) Si, tiempo completo

3) Si, por horas

D2. Dependencia económica

1) Sus padres

2) Parientes

3) Si mismo

4) Otros

D3. ¿Viven tus padres?

1) Si los dos

2) Vive solamente el padre

3) Vive solamente la madre

4) Ninguno

D4. ¿Cuantos hermanos son?

1) Ninguno

2) De 1 a 3

3) De 4 a 5

4) De 6 a más hermanos

D5. ¿Con quién vive actualmente?

1) Con mis padres

2) Esposo (a)

3) Parientes

4) Solo

5) Otros

D6. Lugar donde vive

1) En la ciudad
2) Pueblo joven
3) Zona rural

- **Ingreso.** Es la condición del alumno si logro ingresar a la universidad, no ingreso (0), ingreso primera opción (1), ingreso segunda opción (2).
- **Ingreso A.** Indica la escuela profesional a la cual ingreso, por ejemplo, administración, contabilidad, economía entre otros.
- **Sexo.** Esta expresado como Masculino (M) o Femenino (F).
- **Nombre del colegio.** Nombre del colegio de egreso del estudiante.
- **Idioma Mat.** Es el idioma principal del estudiante.
- **TelCelular.** El número de celular del estudiante.
- **Dirección.** Dirección de vivienda del estudiante.
- **Ubigeo.** EL distrito, provincia y departamento de la vivienda del estudiante.
- **FecNac.** Este atributo nos permitirá obtener la edad del estudiante junto a la fecha de inscripción.
- **NotaAC.** Nota de aptitud academia correspondiente a razonamiento matemático y razonamiento verbal.
- **NotaCO.** Nota de conocimientos que incluye matemáticas (algebra, aritmética, geometría, trigonometría), física, química, biología y humanidades.
- **CODALUMNO.** Es el código asignado a cada estudiante al momento de ingresar.
- **Primer semestre.** Es el año en que estudio es primer semestre.
- **PPS.** Es el promedio ponderado del semestre en vigesimal.

La nota de ingreso de los alumnos por el centro preuniversitario no está en las oficinas de admisión, se recopilo en las oficinas de archivo central.

5.1.2.3. ***Explorar los datos.*** Para realizar la exploración de los datos se tuvo que realizar consultas a la base de datos entregado por la oficina de admisión y la oficina de DICCA, para poder unir en una sola base de datos.

Figura 16

Consultas a las bases de datos

Nota. El grafico muestra las consultas realizadas a las bases de datos para poder combinarlas en una sola base de datos.

Luego para realizar la exploración de los datos usaremos el análisis con la estadística descriptiva para las diferentes variables de la base de datos:

Tabla 2

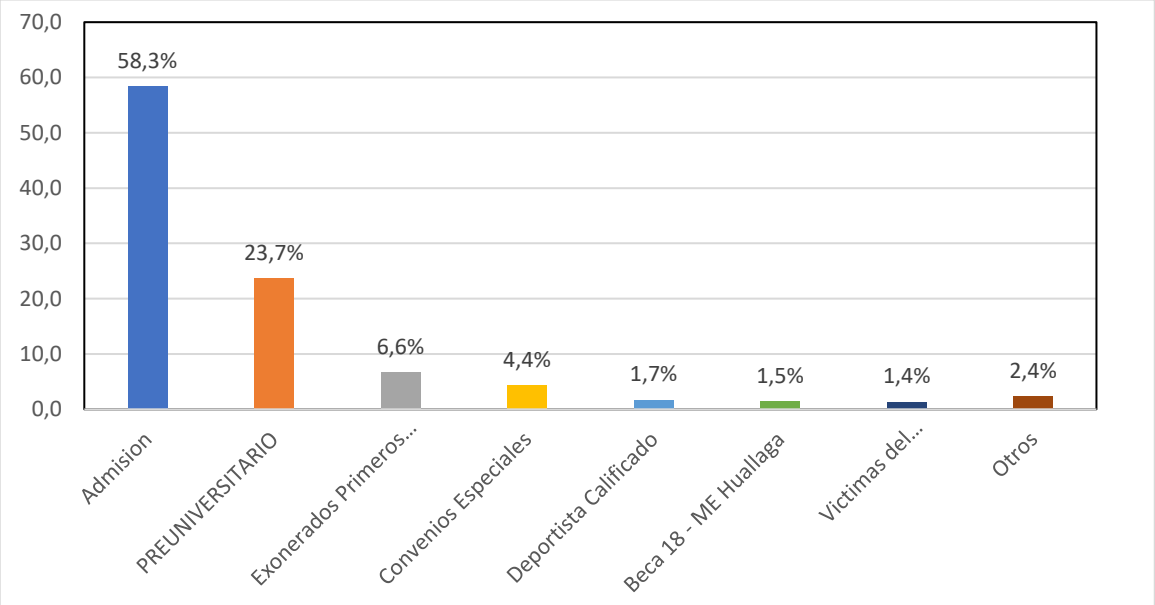
Número de estudiantes ingresantes por distintas modalidades

Modalidad	**Frecuencia**	**Porcentaje**	**Porcentaje acumulado**
Admisión	1699	58.3	58.3
PREUNIVERSITARIO	689	23.7	82.0
Exonerados Primeros Puestos	193	6.6	88.6
Convenios Especiales	128	4.4	93.0
Deportista Calificado	50	1.7	94.7
Beca 18 - ME Huallaga	43	1.5	96.2
Victimas del Terrorismo	40	1.4	97.6
Otros	70	2.4	100.0
Total	**2912**	**100.0**	

Nota. La tabla muestra la cantidad de ingresantes por las distintas modalidades

Figura 17

Distribución de estudiantes por modalidades de ingreso



Nota. El grafico representa la cantidad de estudiantes ingresantes por las distintas modalidades de ingreso.

Como se observa en el grafico el 58,3 % de los alumnos ingresan por la modalidad de admisión y el 23,7 % lo hicieron mediante el centro preuniversitario.

Tabla 3

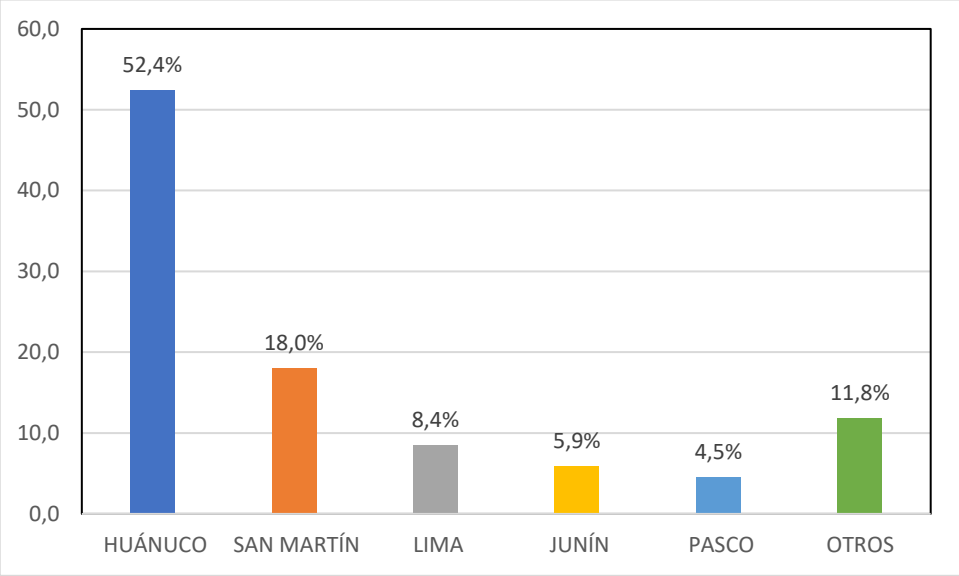
Número de estudiantes ingresantes por el departamento de procedencia

DEPARTAMENTO	Frecuencia	Porcentaje	Porcentaje acumulado
HUÁNUCO	1526	52.4	52.4
SAN MARTÍN	523	18.0	70.4
LIMA	245	8.4	78.8
JUNÍN	172	5.9	84.7
PASCO	131	4.5	89.2
OTROS	315	11.8	100.0
Total	**2912**	**100.0**	

Nota. La tabla muestra la cantidad de alumnos ingresantes de los distintos departamentos del Perú.

Figura 18

Distribución de estudiantes por departamento de procedencia



Nota. La figura muestra la distribución de estudiantes de acuerdo al lugar de procedencia.

Como se puede observar el 52,4 % son del departamento de Huánuco y 18% del departamento de san Martin.

Tabla 4

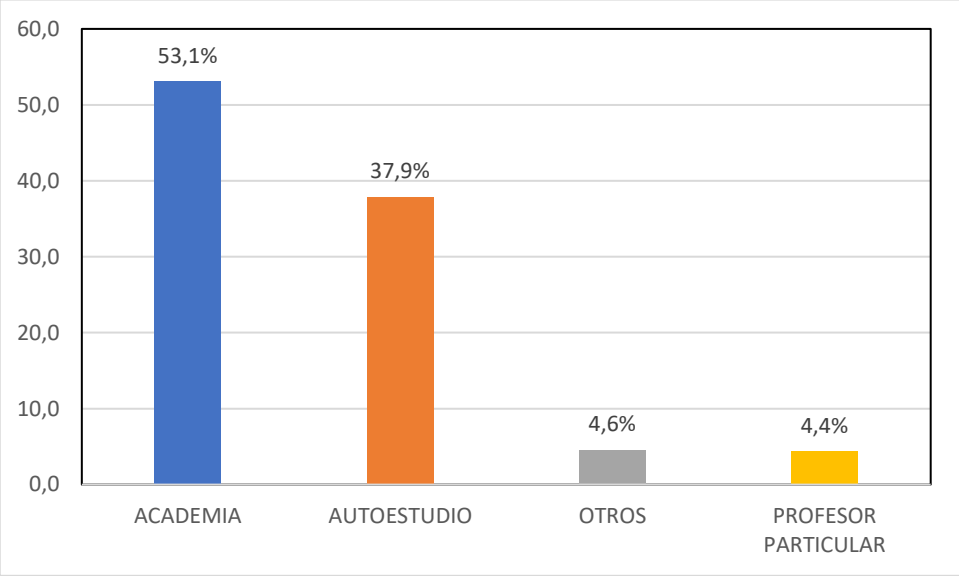
Número de alumnos de acuerdo al tipo de preparación

Tipo de Preparación	Frecuencia	Porcentaje	Porcentaje acumulado
ACADEMIA	1547	53.1	53.1
AUTOESTUDIO	1103	37.9	91.0
OTROS	133	4.6	95.6
PROFESOR	127	4.4	100.0
PARTICULAR			
6	1	0.0	100.0
Total	**2911**	**100.0**	

Nota. La tabla muestra la cantidad de alumnos ingresantes de acuerdo al tipo de preparación.

Figura 19

Distribución de estudiantes de acuerdo al tipo de preparación.



Nota. La figura muestra el porcentaje de la distribución de frecuentas de acuerdo al tipo de preparación que han tenido.

De acuerdo con el grafico el 53,1% de los estudiantes se han preparado en academias, mientras el el 37,9 % ha optado por el autoestudio.

Tabla 5

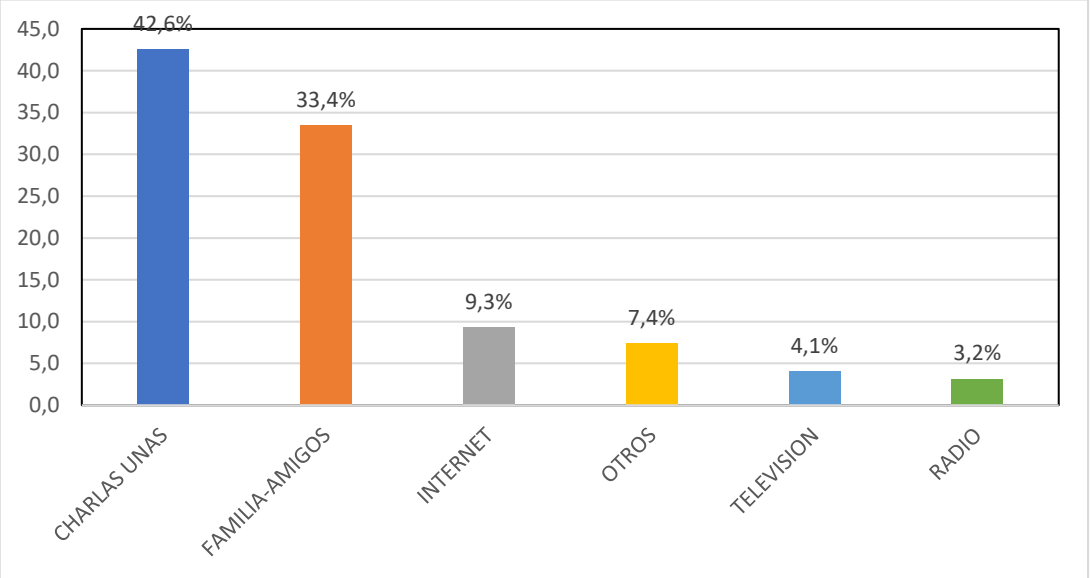
Número de alumnos de acuerdo a como se enteraron de las fechas de los concursos de admisión de UNAS

Forma de informarse	Frecuencia	Porcentaje	Porcentaje acumulado
CHARLAS UNAS	1240	42.6	42.6
FAMILIA-AMIGOS	974	33.4	76.1
INTERNET	270	9.3	85.4
OTROS	215	7.4	92.7
TELEVISION	118	4.1	96.8
RADIO	92	3.2	100.0
7	1	0.0	100.0
Total	**2910**	**99.9**	

Nota. La tabla muestra el número de alumnos que se prepararon en distintas modalidades para poder ingresar a la UNAS.

Figura 20

Distribución de estudiantes de acuerdo a como se enteraron de la UNAS



Nota. La figura muestra la distribución del porcentaje de alumnos de acuerdo a la forma como se enteraron del examen de admisión de la UNAS.

Como se observa en la figura los alumnos ingresantes en un 42,6% se información de postulación a la UNAS por charlas brindadas por la universidad, mientras que un 33,3 % lo hicieron por recomendación de familiares o amigos.

Tabla 6

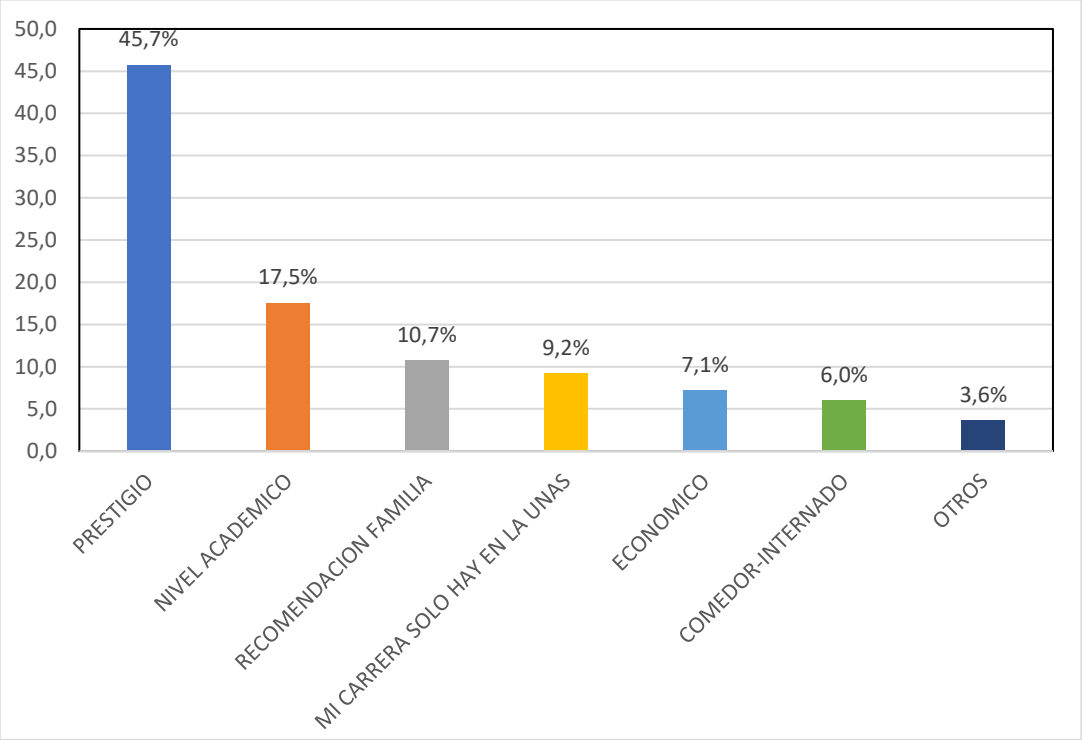
Número de alumnos de acuerdo al motivo por el que postularon a la UNAS

Motivo postulación	Frecuencia	Porcentaje	Porcentaje acumulado
PRESTIGIO	1332	45.7	45.8
NIVEL ACADEMICO	511	17.5	63.3
RECOMENDACION FAMILIA	312	10.7	74.0
MI CARRERA SOLO HAY EN LA UNAS	269	9.2	83.3
ECONOMICO	208	7.1	90.4
COMEDOR-INTERNADO	174	6.0	96.4
OTROS	105	3.6	100.0
Total	**2911**	**100.0**	

Nota. La tabla muestra el número de alumnos de acuerdo a los motivos por los cuales decidieron postular a la UNAS.

Figura 21

Distribución de alumnos de acuerdo al motivo por el cual postularon a la UNAS



Nota: La figura muestra la distribución de alumnos de acuerdo al motivo por el cual decidieron postular a la universidad.

Como se observa en la figura el 45,7% de los alumnos ingresaron a la UNAS por el prestigio de la universidad, mientras que el 17,5 % ingreso por el nivel académico de la UNAS.

Tabla 7

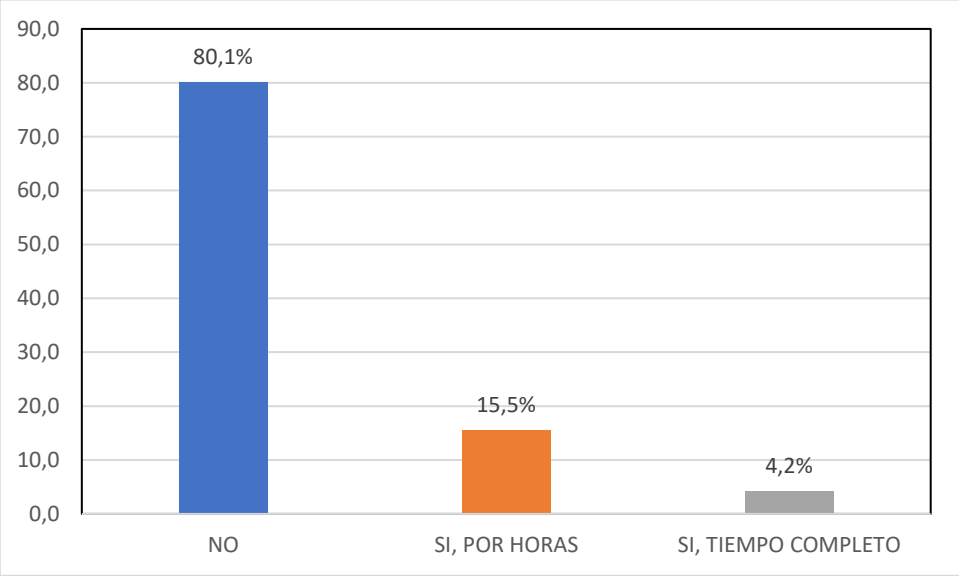
Número de alumnos de acuerdo a la pregunta ¿Trabaja?

TRABAJA	Frecuencia	Porcentaje	Porcentaje acumulado
NO	2333	80.1	80.1
SI, POR HORAS	452	15.5	95.7
SI, TIEMPO COMPLETO	122	4.2	99.9
Total	**2909**	**100.0**	

Nota. La tabla muestra el número de alumnos que trabajan y aquellos que lo hacen a tiempo parcial o a tiempo completo.

Figura 22

Distribución de alumnos de acuerdo a la pregunta ¿Trabajan?



Nota: La figura muestra el porcentaje de alumnos que no trabajan y los que si trabajan. De acuerdo con la figura el 80,1 % no trabaja mientras que el 15,5% trabaja solo por horas y un 4,2% trabaja a tiempo completo.

Tabla 8

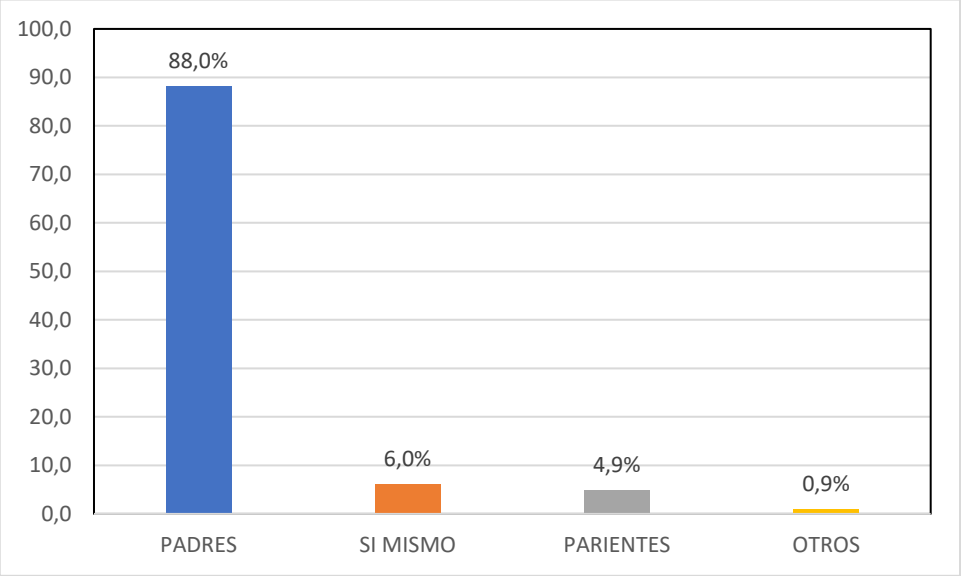
Número de alumnos de quienes dependen económicamente

Depende económicamente	Frecuencia	Porcentaje	Porcentaje acumulado
PADRES	2564	88.0	88.1
SI MISMO	176	6.0	94.2
PARIENTES	143	4.9	99.1
OTROS	27	0.9	100.0
Total	**2910**	**99.9**	

Nota. La tabla muestra de quienes dependen económicamente los alumnos ingresantes.

Figura 23

Distribución de alumnos de acuerdo a la dependencia económica



Nota. La figura muestra de quienes dependen económicamente los alumnos ingresantes a la universidad.

De acuerdo con el grafico el 88 % de los estudiantes dependen económicamente de sus padres mientras que un 6 % se solventan solos con sus gastos de la universidad.

Tabla 9

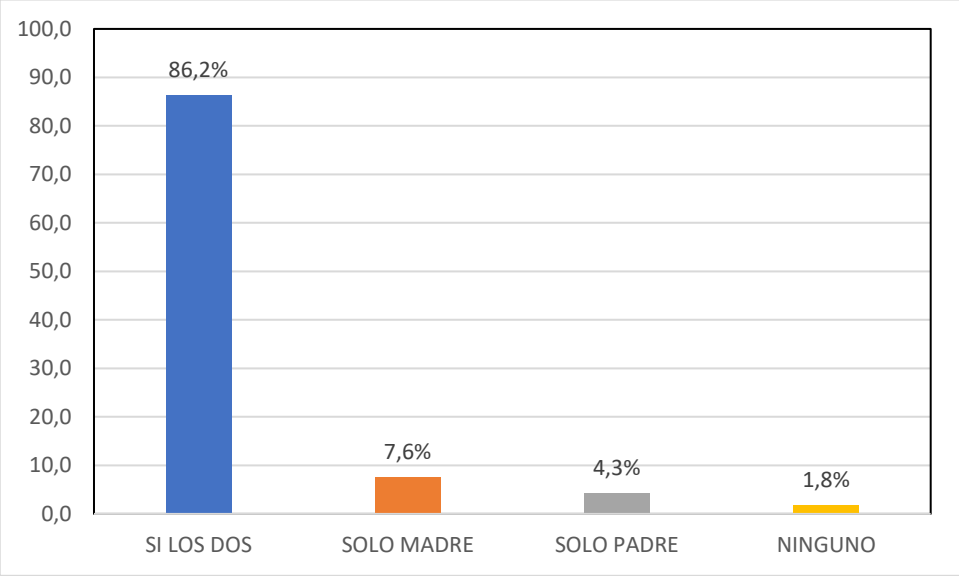
Número de alumnos de acuerdo a la pregunta ¿Viven tus padres?

Padres vivos	Frecuencia	Porcentaje	Porcentaje acumulado
SI LOS DOS	2511	86.2	86.3
SOLO MADRE	222	7.6	93.9
SOLO PADRE	125	4.3	98.2
NINGUNO	51	1.8	100.0
6	1	0.0	100.0
Total	**2910**	**99.9**	

Nota. La tabla muestra el número de alumnos que tienen vivos a sus dos padres, solo a su madre, solo a su padre o ningunos.

Figura 24

Distribución de alumnos que de acuerdo a la pregunta ¿Viven tus padres?



Nota. La figura muestra el porcentaje de alumnos que tienen vivos a sus padres, solo padre, solo madre o ninguno.

Como se observa en la figura el 86,2 % tienen vivos a sus dos padres.

Tabla 10

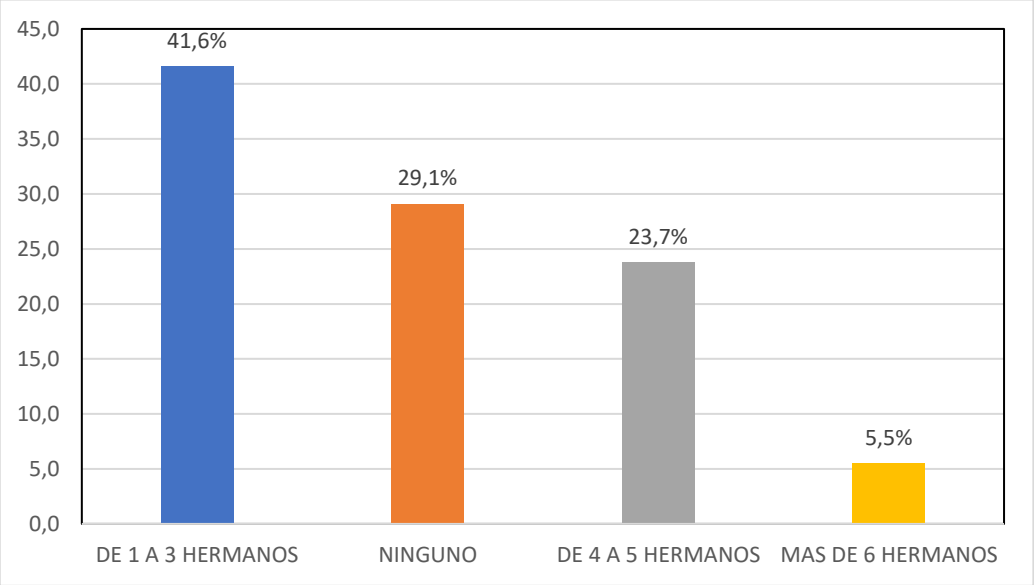
Número de alumnos de acuerdo a cuantos hermanos son en la familia

Número de hermanos	Frecuencia	Porcentaje	Porcentaje acumulado
DE 1 A 3 HERMANOS	1211	41.6	41.6
NINGUNO	847	29.1	70.7
DE 4 A 5 HERMANOS	691	23.7	94.4
MAS DE 6 HERMANOS	160	5.5	99.9
5	1	0.0	100.0
6	1	0.0	100.0
Total	**2911**	**100.0**	

Nota. La tabla muestra el número de alumnos de acuerdo a la cantidad de hermanos que tienen en su familia.

Figura 25

Distribución de alumnos de acuerdo a la cantidad de hermanos



Nota. La figura muestra el porcentaje de alumnos a acuerdo a la cantidad de hermanos que son en la familia.

Como se observa en la figura el 41,6 % tiene de 1 a 3 hermanos mientras que el 29,1 % son hijos únicos y un 23,7 % tiene de 4 a 5 hermanos.

Tabla 11

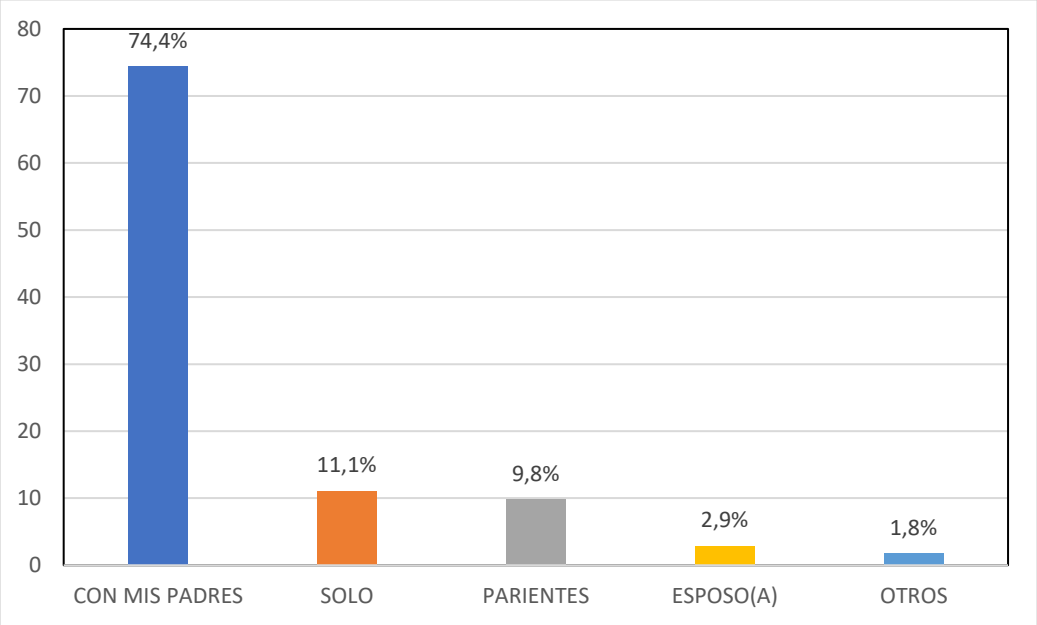
Número de alumnos de acuerdo con quien viven actualmente

Con quien vive	Frecuencia	Porcentaje	Porcentaje acumulado
CON MIS PADRES	2166	74.4	74.4
SOLO	322	11.1	85.5
PARIENTES	285	9.8	95.3
ESPOSO(A)	85	2.9	98.2
OTROS	53	1.8	100
Total	**2911**	**100**	

Nota. La tabla muestra el número de alumnos que viven o con sus padres, solos o con parientes.

Figura 26

Distribución de alumnos de acuerdo con quien viven actualmente



Nota. La figura muestra el porcentaje de alumnos que viven con sus padres, solos o parientes.

De acuerdo con la figura el 74,4 % viven con sus padres, mientras que un 11,1 % vive solo.

Tabla 12

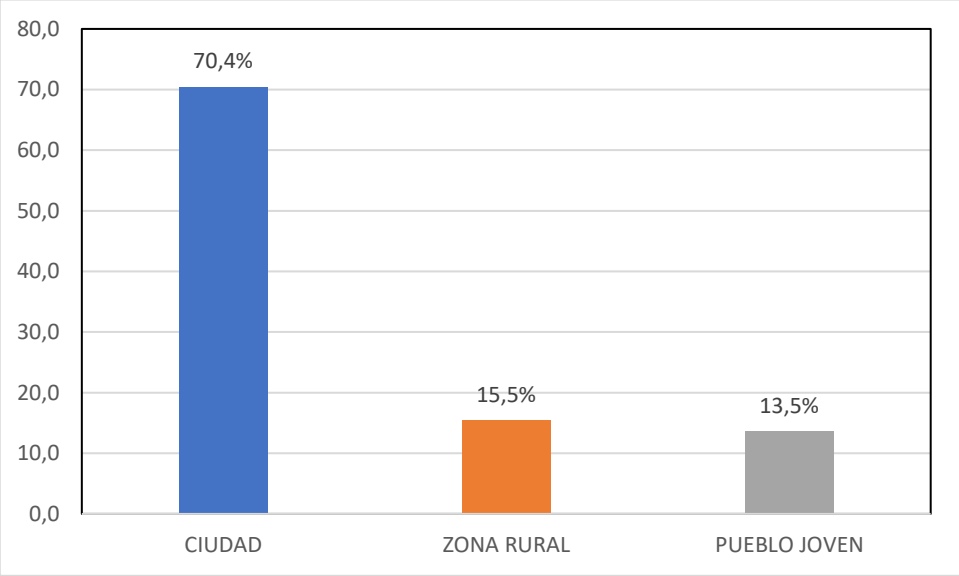
Número de alumnos donde están viviendo actualmente

Lugar donde vive	Frecuencia	Porcentaje	Porcentaje acumulado
CIUDAD	2051	70.4	70.7
ZONA RURAL	451	15.5	86.2
PUEBLO JOVEN	394	13.5	99.8
4	4	0.1	99.9
5	2	0.1	100.0
6	1	0.0	100.0
Total	**2903**	**99.7**	

Nota. La tabla muestra el número de alumnos según el lugar donde viven actualmente.

Figura 27

Distribución de alumnos de acuerdo al lugar donde vive



Nota. La figura muestra el porcentaje de alumnos de acuerdo al lugar donde viven actualmente.

En la figura ser puede observar que el 70,4 % vive dentro de la ciudad, el 15,5 % vive en la zona rural y el 13,5 % vive en un pueblo joven.

Tabla 13

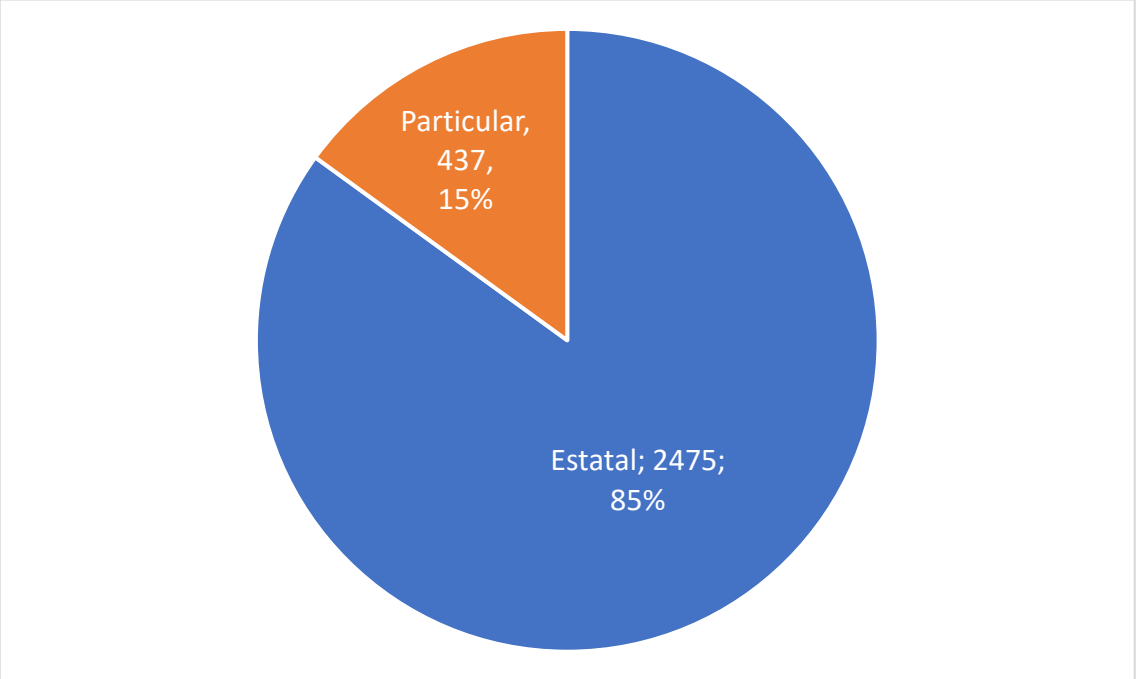
Ingresantes por tipo de colegio del 2015 - 2018

Colegio	Frecuencia	Porcentaje	Porcentaje acumulado
Estatal	2475	85.0	85.0
Particular	437	15.0	100.0
Total	**2912**	**100.0**	

Nota. La tabla muestra la cantidad de ingresantes de los colegios estatales o particulares.

Figura 28

Distribución de ingresantes por tipo de colegio del 2015-2018



Nota. La figura muestra la cantidad y porcentaje de alumnos ingresantes de los colegios estatales y particulares.

Como se observa en la figura el 85 % de los ingresantes provienen de colegios estatales, esto significa que de cada 10 alumnos de la UNAS aproximadamente 8 alumnos son de colegios públicos.

Tabla 14

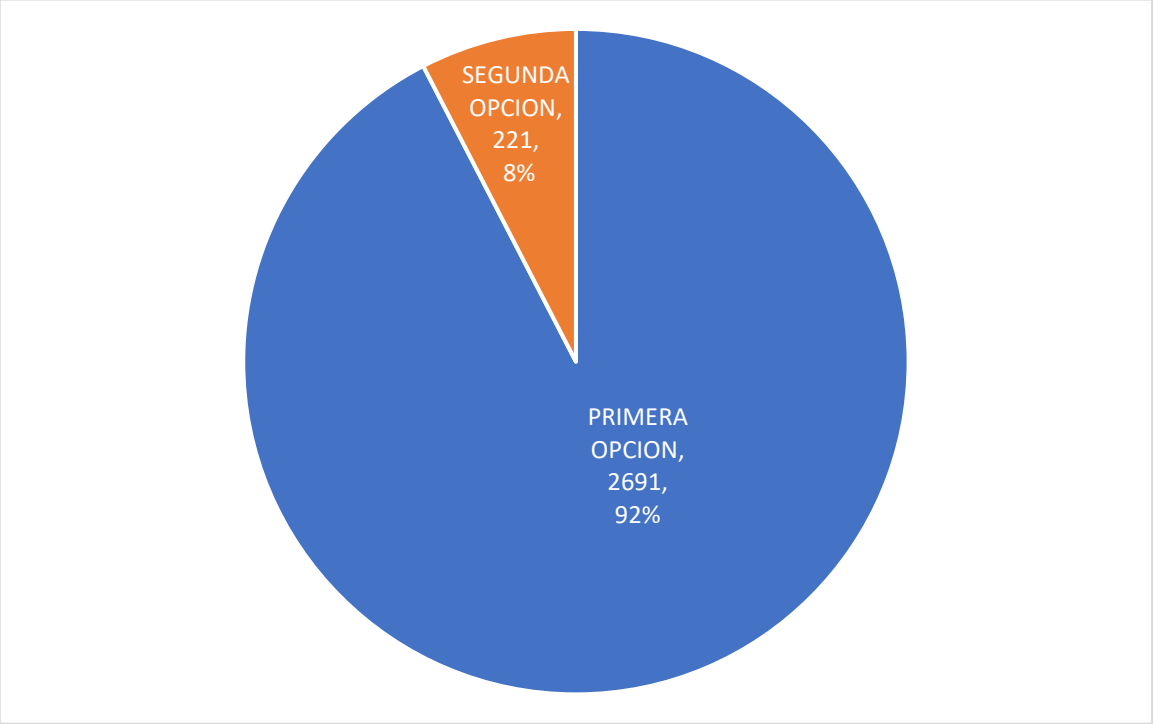
Ingresantes según la opción de ingreso

Opción de ingreso	Frecuencia	Porcentaje	Porcentaje acumulado
PRIMERA OPCION	2691	92.4	92.4
SEGUNDA OPCION	221	7.6	100.0
Total	**2912**	**100.0**	

Nota. La tabla muestra la cantidad de alumnos que ingresaron por la primera y segunda opción.

Figura 29

Distribución de ingresantes según la opción que ingresaron



Nota. La figura muestra la cantidad y porcentaje de alumnos que ingresaron por primera o segunda opción.

De acuerdo con el grafico se observa que el 92 % de los ingresantes a la UNAS lo hicieron por la primera opción, esto significa que de cada 10 alumnos 9 de ellos ingresaron por la primera opción.

Tabla 15

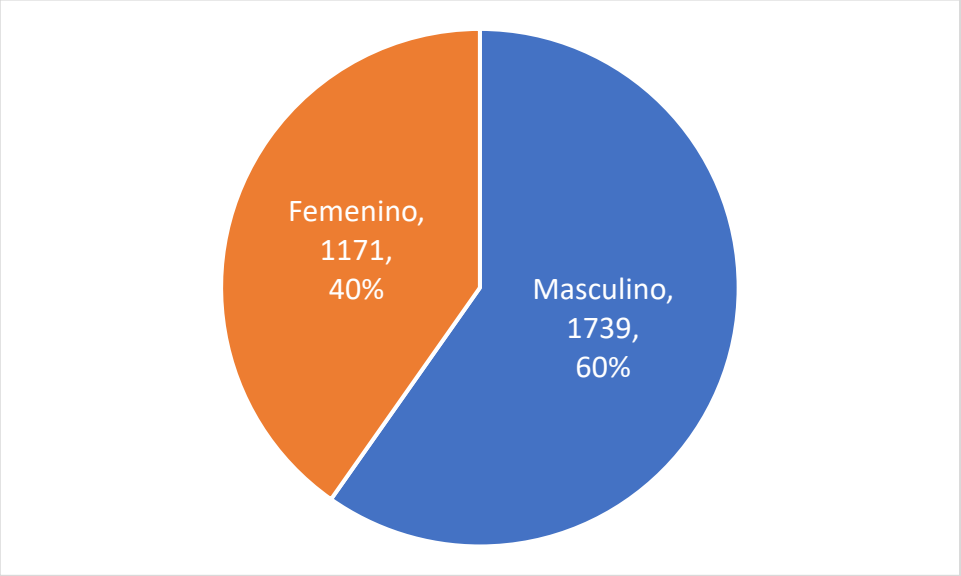
Ingresantes por genero del 2015-2018

Genero	**Frecuencia**	**Porcentaje**	**Porcentaje acumulado**
Masculino	1739	59.7	59.7
Femenino	1171	40.2	99.9
	2	0.1	100.0
Total	**2912**	**100.0**	

Nota. La tabla los ingresantes a la UNAS por género.

Figura 30

Distribución de ingresantes por genero del 2015-2018



Nota. La figura muestra la cantidad y porcentaje de ingresantes a la UNAS por género. Como se observa en la figura el 60 % de los ingresantes a las UNAS son varones y el 40 % son mujeres, por cada 10 ingresantes, 6 de ellos son varones y 4 mujeres.

5.1.2.4. ***Verificar la calidad de los datos.*** En esta fase se hace un análisis de la calidad de los datos, es decir se analiza si existen o no inconsistencias, teniendo los siguientes detalles:

- Existen datos duplicados de alumnos, ya que algunos ingresaron por el centro preuniversitario por la segunda opción, y luego dieron el examen de admisión para ingresar por su primera opción. En algunos casos ingresaron por modalidades y luego dieron el examen de admisión, logrando ingresar en ambos casos.
- Existen datos incompletos les falta la nota del promedio ponderado, muchos de los alumnos que ingresaron por modalidades no tienen nota de ingreso, por ejemplo, los alumnos que ingresaron por primeros puestos ingresaron sin dar examen en algunas facultades.
- La edad de algunos de ellos con inconsistentes ya que aparecen que tienen 115 años, esto debido a que la fecha de nacimiento fue las registrado.
- También existen registros vacíos o algunos casos pusieron opciones que no existen en la encuesta.

5.1.3 PREPARACIÓN DE DATOS

En esta etapa desarrollamos las actividades para construir la base de datos final que nos servirá para el modelamiento, estos datos se obtienen a partir de los datos brutos iniciales.

Estas tareas incluyen la selección de atributos y registros, también la transformación y la limpieza de los datos, para generar los modelos predictivos.

- Selección de datos

Primero se procedió a eliminar las variables que no se pueden usar por ética y seguridad como son la variable apellidos y nombres, el DNI del alumno, el código del alumno, al igual que las elecciones de las carreras de su primera opción y segunda opción y solo nos quedamos con la opción de ingreso y la carrera que ingreso, se calculó la edad del alumno usando las columnas fecha de inscripción y fecha de nacimiento, para luego eliminar las variables fecha de nacimiento y fecha de inscripción.

Luego la variable abigeo de procedencia se tuvo que extraer la ubicación de la región =EXTRAE([@Column5],11,2), para llenar la variable con la región de procedencia.

Como se desea conocer la condición del ingresante al finalizar el primer semestre si está con promedio ponderado mayor o igual a 11 (aprobado), caso contrario menor a 11 (desaprobado), agregamos una columna de datos condición del alumno de tipo categórico.

A continuación, se describen los atributos seleccionados:

Tabla 16

Atributos seleccionados para la minería de datos

Atributo	Tipo	Descripción
Modalidad	Cualitativa – politómica	La modalidad por la que ingresaron a la universidad
Tipo_colegio	Cualitativa – dicotómica	Establece el tipo de colegio donde termino la secundaria
Ingreso	Cualitativa – dicotómica	Indica la opción con que ingreso a la universidad si es la primera o segunda opción
Procedencia	Cualitativa - politómica	Región de procedencia del alumno ingresante
Sexo	Cualitativa - dicotómica	Genero del estudiante
NotaIngreso	Cuantitativa – continua	Es la nota con la que ingreso el estudiante a la universidad
Edad	Cuantitativa – discreta	Edad con la ingreso el estudiante
TipoPrep	Cualitativa - politómica	Tipo de preparación que realizo el alumno para ingresar
Comosente	Cualitativa - politómica	La forma como se enteró de la UNAS y postular a la misma.
MOTPOSTULAR	Cualitativa - politómica	El motivo por el cual postula a la universidad
TRABAJA	Cualitativa - politómica	SI trabaja en el tiempo que estudia
DEP_ECONOMICA	Cualitativa - politómica	Establece de quien depende económicamente
VIVEPADRES	Cualitativa - politómica	Indica si sus padres están vivos
NUM_HERMANOS	Cualitativa - politómica	Establece la cantidad de hermanos que tiene
VIVE_CON	Cualitativa - politómica	Indica con quien vive actualmente mientras estudia
DONDEVIVE	Cualitativa - politómica	Indica el lugar donde vive actualmente
ESCUELA PROFESIONAL	Cualitativa - politómica	Estable la escuela profesional a la cual ingreso
PPS	Cuantitativa – continua	Promedio ponderado semestral del estudiante
CONDICION	Cualitativa – dicotómica	Si al final aprobó o desaprobó el semestre

Nota. La tabla muestra los atributos seleccionados para el modelo predictivo.

- Limpieza de datos

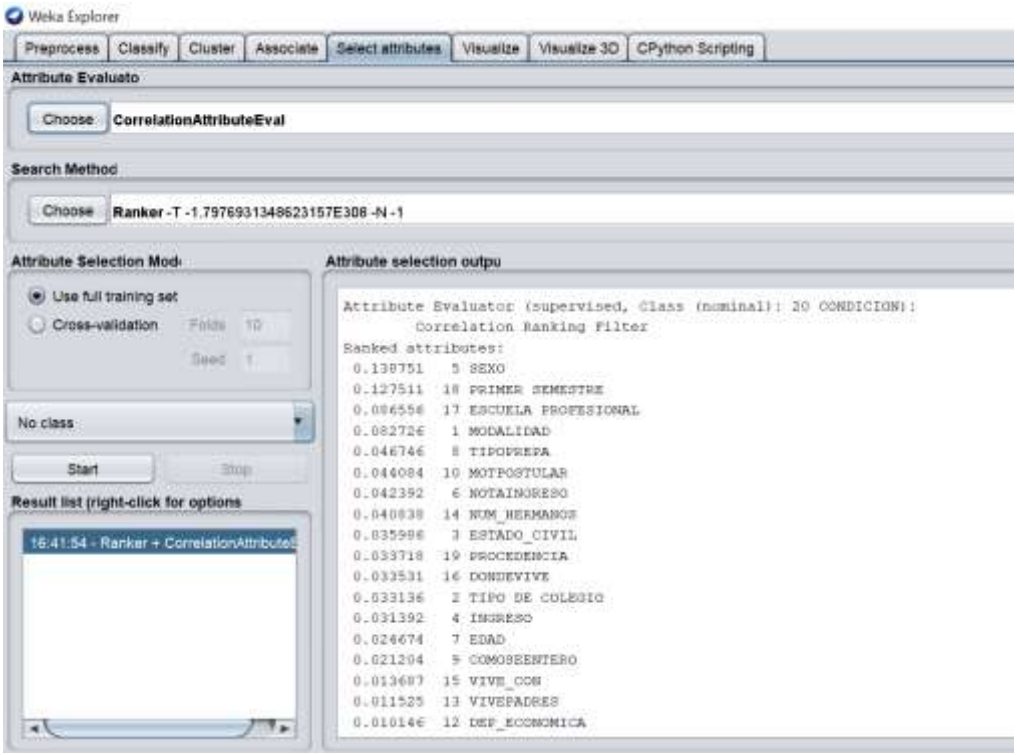
En esta fase describimos los factores más importantes que permitirán predecir el rendimiento académico, para ello primero de los 2912 registros se procedió a eliminar los datos duplicados, ya que algunos ingresaron por la modalidad preuniversitaria y luego por admisión en algunos casos a distintas carreras o la misma carrera, seguido se eliminó aquellos registros nulos en el promedio ponderado ya que ingresaron, pero no se matricularon en el ciclo correspondiente.

También se eliminaron 5 registros que no tenían las encuestas rellenadas, al igual que había datos inconsistentes como la edad de un estudiante con 118 años, o 12 años.

Para la limpieza se analizó cada uno de los atributos de la base de datos. En el software WEKA también se analizó que atributos permiten tener una mejor correlación, obteniéndose los siguientes resultados:

Figura 31

Correlación según los atributos en el software WEKA



Nota. La figura muestra la correlación de cada uno de los atributos con respecto al atributo condición.

- Estructuración de datos

Para la construcción de los nuevos datos, algunos datos se normalizo, y las carreras se construyeron nuevos datos de la siguiente manera:

- ADMINISTRACION: {0,1}

  Donde:

  0: No es ADMINISTRACION

  1: Es ADMINISTRACION
- AGRONOMIA: {0,1}

  Donde:

  0: No es AGRONOMIA

  1: Es AGRONOMIA
- CONTABILIDAD: {0,1}

  Donde:

  0: No es CONTABILIDAD

  1: Es CONTABILIDAD
- ECONOMIA: {0,1}

  Donde:

  0: No es ECONOMIA

  1: Es ECONOMIA
- AMBIENTAL: {0,1}

  Donde:

  0: No es AMBIENTAL

  1: Es AMBIENTAL
- ING SUELOS: {0,1}

  Donde:

  0: No es ING SUELOS

  1: Es ING SUELOS
- INDUSTRIAS ALIMENTARIAS: {0,1}

  Donde:

  0: No es INDUSTRIAS ALIMENTARIAS

  1: Es INDUSTRIAS ALIMENTARIAS
- SISTEMAS: {0,1}

  Donde:

  0: No es SISTEMAS

  1: Es SISTEMAS
- RNR: {0,1}

Donde:

0: No es RECURSONAS NATURALES RENOVABLES

1: Es RECURSONAS NATURALES RENOVABLES

- FORESTAL: {0,1}

  Donde:

  0: No es FORESTAL

  1: Es FORESTAL

- MECANICA: {0,1}

  Donde:

  0: No es MECANICA

  1: Es MECANICA

- ZOOTECNIA: {0,1}

  Donde:

  0: No es ZOOTECNIA

  1: Es ZOOTECNIA

- Integrar datos

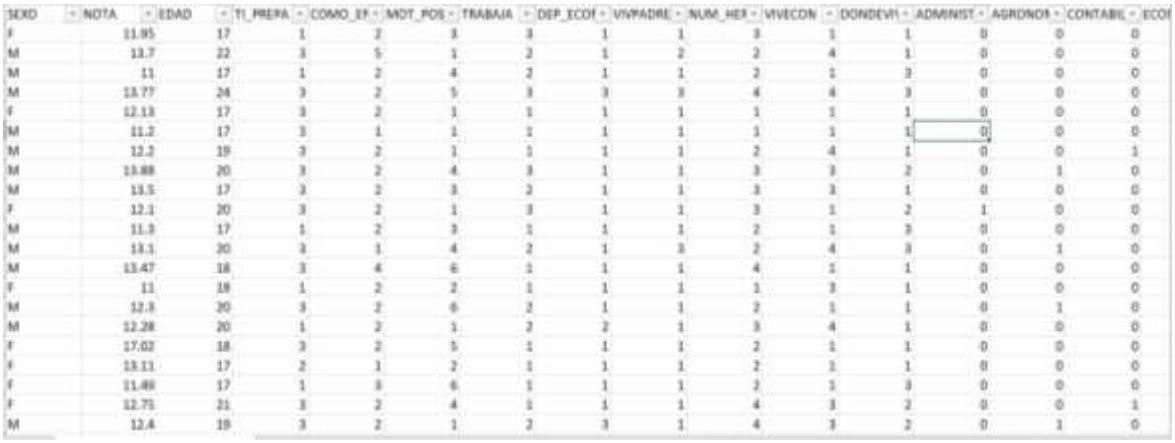
Los registros fueron integrados en una sola tabla llamada DATOS_ALUMNO_DETALLADO que es vamos a utilizar para aplicar la minería de datos las cuales fueron integrados de los registros obtenidos desde las oficinas de admisión, y de la base de datos de DICCA. Al final nos quedamos con un total de 2404 registros.

- Formatear datos

La tabla se convirtió en archivo tipo .csv para usar en el software WEKA, quedando de la siguiente manera:

Figura 32

Datos en el archivo .csv



SEXO	NOTA	EDAD	TI_PREPA	COMO_EF	MOT_POS	TRABAJA	DEP_ECOF	VIVPADRE	NUM_HER	VIVECON	DONDEVI	ADMINIST	AGRONOM	CONTABIL	ECOF
F	11.95	17	1	2	3	3	1	1	3	1	1	0	0	0	
M	13.7	22	3	5	1	2	1	2	2	4	1	0	0	0	
M	11	17	1	2	4	2	1	1	2	1	3	0	0	0	
M	13.77	24	3	2	5	3	3	3	4	4	3	0	0	0	
F	12.13	17	3	2	1	1	1	1	1	1	1	0	0	0	
M	11.2	17	3	1	1	1	1	1	1	1	1	0	0	0	
M	12.2	19	3	2	1	1	1	1	2	4	1	0	0	1	
M	13.88	20	3	2	4	3	1	1	3	3	2	0	1	0	
M	13.5	17	3	2	3	2	1	1	3	3	1	0	0	0	
F	12.1	20	3	2	1	3	1	1	3	1	2	1	0	0	
M	11.3	17	1	2	3	1	1	1	2	1	3	0	0	0	
M	13.1	20	3	1	4	2	1	3	2	4	3	0	1	0	
M	13.47	18	3	4	6	1	1	1	4	1	1	0	0	0	
F	11	18	1	2	2	1	1	1	1	3	1	0	0	0	
M	12.3	20	3	2	6	2	1	1	2	1	1	0	1	0	
M	12.28	20	1	2	1	2	2	1	3	4	1	0	0	0	
F	17.02	18	3	2	5	1	1	1	2	1	1	0	0	0	
F	13.11	17	2	1	2	1	1	1	2	1	1	0	0	0	
F	11.49	17	1	3	6	1	1	1	2	1	3	0	0	0	
F	12.75	21	3	2	4	1	1	1	4	3	2	0	0	1	
M	12.4	19	3	2	1	2	3	1	4	3	2	0	1	0	

Nota. La figura muestra los datos que están en el archivo csv.

5.1.4. MODELAMIENTO

En esta fase, seleccionamos y aplicamos las técnicas de modelado en el software WEKA 3.9, se calibran los parámetros y hiperparametros para obtener los valores óptimos para la predicción.

- Selección de la técnica de modelado

Para el modelado utilizaremos el software WEKA 3.9, usaremos 7 algoritmos, los más importantes, a continuación, se detallan cada uno de los algoritmos.

a. ALGORITMO NAIVE BAYES

En el software WEKA 3.9 seleccionados el algoritmo “naive bayes” lo cual nos arroja los siguientes resultados resumidos en las siguientes tablas:

Tabla 17

Clasificación de las instancias con el algoritmo NAIVE BAYES

ALGORITMO NAIVE BAYES		
Instancias correctamente clasificadas	1662	69.1348%
Instancias incorrectamente clasificadas	742	30.8652%
Total	**2404**	**100%**

Nota. En la tabla se muestra la cantidad y el porcentaje de la clasificación de las instancias.

Tabla 18

Matriz de confusión de acuerdo a Naive bayes

MATRIZ DE CONFUSIÓN		
	Aprobados	Desaprobados
Aprobados	1019	436
Desaprobados	306	643

Nota. En la tabla se observa las instancias correctamente clasificadas

Interpretación de los resultados con naive bayes.

Se puede observar en la tabla 19 que el 69.1348% se clasifico de manera correcta y en la matriz de confusión 1019 instancias fueron clasificados de manera correcta para los aprobados y 643 instancias fueron clasificados de manera correcta para los desaprobados.

b. ALGORITMO ARBOL DE DECISIÓN J48

En el software WEKA 3.9 seleccionados el algoritmo "árbol de decisión J48" lo cual nos arroja los siguientes resultados resumidos en las siguientes tablas:

Tabla 19

Clasificación de las instancias según algoritmo árbol de decisión J48

ALGORITMO ARBOL DE DECISIÓN J48		
Instancias correctamente clasificadas	1590	66.1398%
Instancias incorrectamente clasificadas	814	33.8602%
Total	**2404**	**100%**

Nota. En la tabla se muestra la cantidad y el porcentaje de la clasificación de las instancias.

Tabla 20

Matriz de confusión de árbol de decisión J48

MATRIZ DE CONFUSIÓN		
	Aprobados	Desaprobados
Aprobados	1085	370
Desaprobados	444	505

Nota. En la tabla se observa las instancias correctamente clasificados

Interpretación de los resultados con árbol de decisión J48

Se puede observar en la tabla 21 que el 66.1398% se clasifico de manera correcta y en la matriz de confusión 1085 instancias fueron clasificados de manera correcta para los aprobados y 505 instancias fueron clasificados de manera correcta para los desaprobados.

c. ALGORITMO ARBOL DE DECISIÓN RAMDOM FOREST

En el software WEKA 3.9 seleccionados el algoritmo "árbol de decisión ramdom forest" lo cual nos arroja los siguientes resultados resumidos en las siguientes tablas:

Tabla 21

Clasificación de las instancias según el algoritmo árbol de decisión ramdom forest

ALGORITMO ARBOL DE DECISIÓN RAMDOM FOREST		
Instancias correctamente clasificadas	1677	69.7587%
Instancias incorrectamente clasificadas	727	30.2413%
Total	**2404**	**100%**

Nota. En la tabla se muestra la cantidad y el porcentaje de la clasificación de las instancias.

Tabla 22

Matriz de confusión del árbol de decisión ramdom forest

MATRIZ DE CONFUSIÓN		
	Aprobados	Desaprobados
Aprobados	1174	281
Desaprobados	446	503

Nota. En la tabla se observa las instancias correctamente clasificadas

Se puede observar en la tabla 22 que el 69.7587% se clasifico de manera correcta y en la matriz de confusión 1174 instancias fueron clasificados de manera correcta para los aprobados y 503 instancias fueron clasificados de manera correcta para los desaprobados.

d. ALGORITMO REGRESION LOGISTICA

En el software WEKA 3.9 seleccionados el algoritmo "regresión logística" lo cual nos arroja los siguientes resultados resumidos en las siguientes tablas:

Tabla 23

Clasificación de las instancias según el algoritmo regresión logística

ALGORITMO REGRESÓN LOGISTICA		
Instancias correctamente clasificadas	1724	71.7138%
Instancias incorrectamente clasificadas	680	28.2862%
Total	**2404**	**100%**

Nota. En la tabla se muestra la cantidad y el porcentaje de la clasificación de las instancias.

Tabla 24

Matriz de confusión de la regresión logística

MATRIZ DE CONFUSIÓN		
	Aprobados	Desaprobados
Aprobados	1158	297
Desaprobados	383	566

Nota. En la tabla se observa las instancias correctamente clasificadas

Se puede observar en la tabla 25 que el 71.7138% se clasifico de manera correcta y en la matriz de confusión 1724 instancias fueron clasificados de manera correcta para los aprobados y 566 instancias fueron clasificados de manera correcta para los desaprobados.

e. ALGORITMO DE VECINOS MAS CERCANOS

En el software WEKA 3.9 seleccionados el algoritmo “vecinos más cercanos” lo cual nos arroja los siguientes resultados resumidos en las siguientes tablas:

Tabla 25

Clasificación de las instancias según el algoritmo vecinos mas cercanos

ALGORITMO VECINOS MAS CERCANOS		
Instancias correctamente clasificadas	1528	63.5607%
Instancias incorrectamente clasificadas	876	36.4393%
Total	**2404**	**100%**

Nota. En la tabla se muestra la cantidad y el porcentaje de la clasificación de las instancias.

Tabla 26

Matriz de la confusión de vecinos más cercanos

MATRIZ DE CONFUSIÓN		
	Aprobados	Desaprobados
Aprobados	1028	427
Desaprobados	449	500

Nota. En la tabla se observa las instancias correctamente clasificadas

Se puede observar en la tabla 27 que el 63.5607% se clasifico de manera correcta y en la matriz de confusión 1028 instancias fueron clasificados de manera correcta para los aprobados y 500 instancias fueron clasificados de manera correcta para los desaprobados.

f. ALGORITMOS MULTILAYER PERCEPTRÓN

En el software WEKA 3.9 seleccionados el algoritmo “multilayer perceptron” lo cual nos arroja los siguientes resultados resumidos en las siguientes tablas:

Tabla 27

Clasificación de las instancias según el algoritmo multilayer perceptron

ALGORITMO MULTILAYER PERCEPTRÓN		
Instancias correctamente clasificadas	1602	66.6389%
Instancias incorrectamente clasificadas	802	33.3611%
Total	**2404**	**100%**

Nota. En la tabla se muestra la cantidad y el porcentaje de la clasificación de las instancias.

Tabla 28

Matriz de confusión multilayer perceptrón

MATRIZ DE CONFUSIÓN		
	Aprobados	Desaprobados
Aprobados	1067	388
Desaprobados	414	535

Nota. En la tabla se observa las instancias correctamente clasificados

Se puede observar en la tabla 29 que el 66.6389% se clasifico de manera correcta y en la matriz de confusión 1067 instancias fueron clasificados de manera correcta para los aprobados y 535 instancias fueron clasificados de manera correcta para los desaprobados.

g. ALGORITMOS EMSAMBLADOS VOTE

En el software WEKA 3.9 seleccionados un Meta clasificador "vote" dentro de ellos seleccionamos el algoritmo regresión logística con clasificador principal y ramdom forest como segundo clasificador lo cual nos arroja los siguientes resultados resumidos en las siguientes tablas:

Tabla 29

Clasificación de las instancias según el algoritmo VOTE

ALGORITMO VOTE		
Instancias correctamente clasificadas	1709	71.0899%
Instancias incorrectamente clasificadas	695	28.9101%
Total	**2404**	**100%**

Nota. En la tabla se muestra la cantidad y el porcentaje de la clasificación de las instancias.

Tabla 30

Matriz de confusión ensamblado VOTE

MATRIZ DE CONFUSIÓN		
	Aprobados	Desaprobados
Aprobados	1171	284
Desaprobados	411	538

Nota. En la tabla se observa las instancias correctamente clasificadas

Se puede observar en la tabla 31 que el 71.0899% se clasifico de manera correcta y en la matriz de confusión 1171 instancias fueron clasificados de manera correcta para los aprobados y 538 instancias fueron clasificados de manera correcta para los desaprobados.

- Generación de la prueba de diseño

Las pruebas de diseños las hemos obtenido en la elección de las técnicas de modelado y la hemos clasificado en la siguiente tabla:

Tabla 31

Resumen de los algoritmos con los porcentajes de predicción y los aciertos

Nro	ALGORITMO	PORCENTAJE DE ACEIRTO	ACERTADOS APROBADOS	ACERTADOS DESAPROBADOS
1	NAIVE BAYES	69.1348%	1019	643
2	ARBOL DE DECISIÓN J48	66.1398%	1085	505
3	RAMDOM FOREST	69.7587%	1174	503
4	REGRESION LOGISTICA	71.7138%	1158	566
5	VECINOS MAS CERCANOS	63.5607%	1028	500
6	MULTILAYER PERCEPTRON	66.6389%	1067	535
7	EMSAMBLADOS VOTE	71.0899%	1171	538

Nota. La tabla muestra los porcentajes de aciertos y de las instancias aprobadas y desaprobadas de forma correcta.

Como se observa en la tabla el de regresión logística es el que mejor porcentaje de acerito tiene y con el que vamos a trabajar para realizar las predicciones del rendimiento académico de los estudiantes ingresantes de la UNAS.

- Construcción del modelo

Para la construcción del modelo tenemos dos grupos definidos los cuales usaremos con el modelo que mejor resultado hemos obtenido anteriormente.

Para el entrenamiento vamos a usar los 2404 registros que son alumnos de los primeros ciclos desde el 2015 hasta el 2018

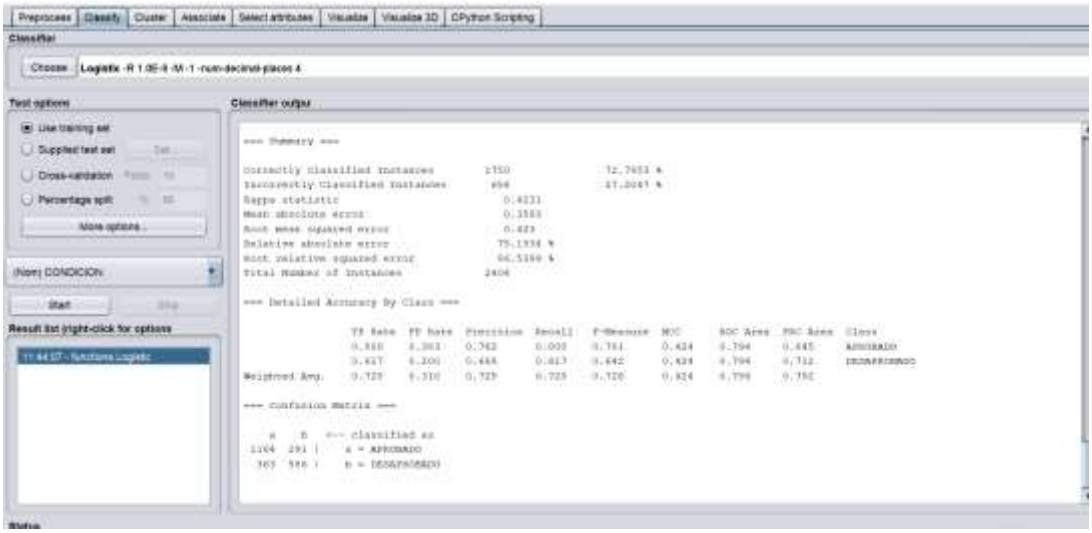
Para la predicción o grupo (test) usaremos los registros de los alumnos del primer ciclo que estudiaron durante el ciclo 2019-I que son 591 registros, el atributo de **condición**: “APROBADO” o “DESAPROBADO”, serán desconocidos, las cuales tendrán el signo de interrogación, el algoritmo será el que lo clasifique si el rendimiento del estudiante

será aprobado o desaprobado al final del semestre, lo cual será luego corroborado con los valores reales que se tienen en la “DATOS_ALUMNOS_PREDICCION”.

Se realiza un entrenamiento al modelo elegido para luego guardar el modelo. El entrenamiento del modelo se observa en la siguiente figura:

Figura 33

Entrenamiento del modelo regresión logística



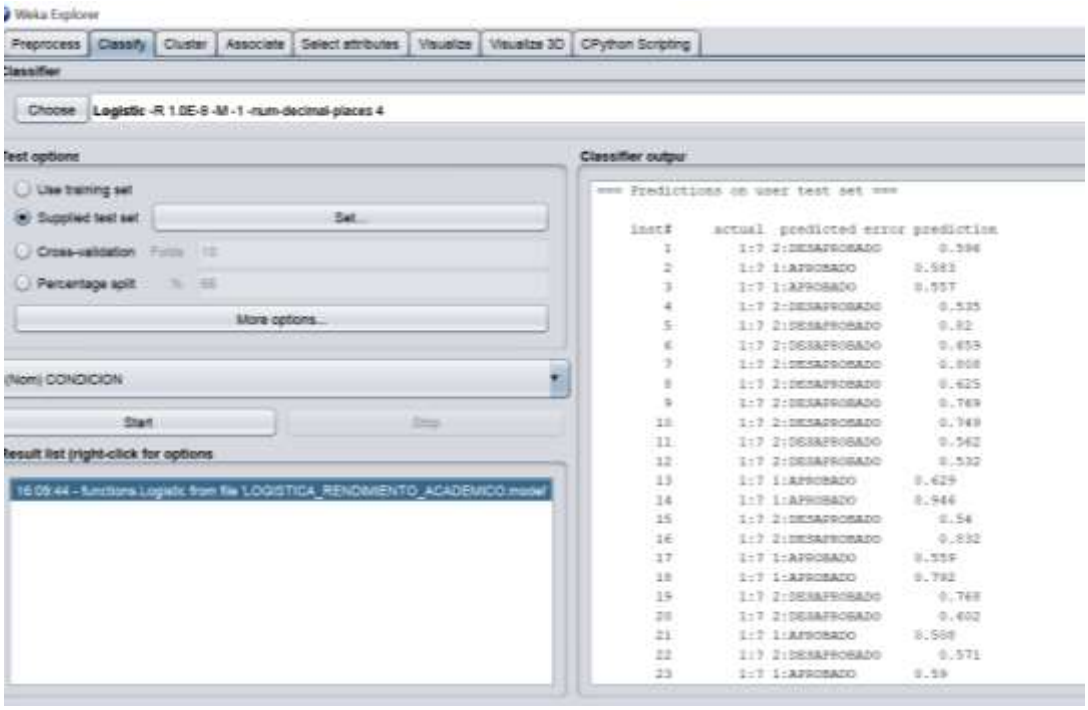
Nota. En la figura se muestra los resultados del entrenamiento del algoritmo de regresión logística.

- Evaluación del modelo

Con el modelo ya construido y entrenado en el software WEKA 3.9 se realiza el test a la base de datos DATOS_ALUMNOS_PREDICCION.csv teniendo los siguientes resultados:

Figura 34

Predicción con el modelo de regresión logística



Nota. En la figura se observa el resultado de la predicción con el porcentaje de acierto. Los demás resultados están en el anexo 04

5.1.5. Evaluación

Aquí vamos a evaluar el modelo construido en la fase anterior.

- Evaluación de los resultados

El resultado obtenido en la predicción vamos comparar con los resultados reales en una hoja Excel para determinar en las cuales las predicciones fueron CORRECTAS y en cuales INCORRECTAS.

Tabla 32

Instancias clasificadas de manera correcta e incorrecta

PREDICCIÓN	DATOS REALES	OBSERVACIÓN
APROBADO	APROBADO	CORRECTO
APROBADO	APROBADO	CORRECTO
APROBADO	APROBADO	CORRECTO
APROBADO	APROBADO	CORRECTO
DESAPROBADO	DESAPROBADO	CORRECTO
APROBADO	APROBADO	CORRECTO
APROBADO	DESAPROBADO	INCORRECTO
DESAPROBADO	DESAPROBADO	CORRECTO
DESAPROBADO	DESAPROBADO	CORRECTO
DESAPROBADO	APROBADO	INCORRECTO
APROBADO	APROBADO	CORRECTO
DESAPROBADO	DESAPROBADO	CORRECTO
APROBADO	APROBADO	CORRECTO
APROBADO	APROBADO	CORRECTO
DESAPROBADO	APROBADO	INCORRECTO
APROBADO	APROBADO	CORRECTO
APROBADO	APROBADO	CORRECTO
DESAPROBADO	DESAPROBADO	CORRECTO
APROBADO	APROBADO	CORRECTO
APROBADO	APROBADO	CORRECTO
APROBADO	DESAPROBADO	INCORRECTO
APROBADO	APROBADO	CORRECTO
APROBADO	APROBADO	CORRECTO
DESAPROBADO	APROBADO	INCORRECTO

Nota. La tabla muestra las instancias con las que se realizó la predicción con el modelo entrenado de regresión logística.

- Proceso de revisión

Se realizó la comparación de todas las instancias clasificadas con el modelo entrenado en el software WEKA 3.9

- Determinación de futuras fases

Los pasos por seguir será la predicción de los nuevos ingresantes, y cada que los datos sean mayores, la predicción tendrá un mayor porcentaje de acierto.

5.1.6. DESPLIEGUE DEL PROYECTO

Los resultados obtenidos serán entregados a las autoridades de la universidad para que tomen las medidas correctivas para mejorar el rendimiento académico de los estudiantes ingresantes a la universidad.

5.2. CONTRASTACIÓN DE HIPÓTESIS

Los resultados obtenidos en el software WEKA 3.9 utilizando algoritmos de minería de datos en la información de los estudiantes que ingresaron en el año 2109 de acuerdo con la hipótesis planteada en la investigación.

HG: Con las técnicas de la minería de datos se puede predecir el rendimiento académico de los estudiantes del primer ciclo de la Universidad Nacional Agraria de la Selva, Tingo María 2021

Las predicciones que se realizó con las distintas técnicas de minería de datos arrojan distintos niveles de exactitud. Los resultados muestran que con la minería de datos se puede realizar predicciones del rendimiento académico de los estudiantes del primer ciclo, utilizando diferentes modelos. Los resultados arrojaron que se encontraron porcentajes muy alto de acierto de 98% que tiene un acuracy de 71.7 % en la tabla mostramos los modelos con mejores resultados.

Tabla 33

Resultados de las predicciones

Nro	ALGORITMO	PORCENTAJE DE ACEIRTO	ACERTADOS APROBADOS	ACERTADOS DESAPROBADOS
1	RAMDOM FOREST	69.7587%	1174	503
2	REGRESIÓN LOGISTICA	71.7138%	1158	566
3	EMSAMBLADOS VOTE	71.0899%	1171	538

Nota. La tabla muestra los porcentajes de aciertos y de las instancias aprobadas y desaprobadas de forma correcta.

HI: El rendimiento académico en estudiantes universitarios de primer ciclo se puede predecir mediante indicadores sociales, económicos y académicos de la Universidad Nacional Agraria de la Selva.

Tabla 35

Las variables no están en la ecuación

			Chi-cuadrado	gl	Sig.
Paso 0	Variables	MODALIDAD	130.997	9	7.3975E-24
		MODALIDAD(1)	54.916	1	1.2577E-13
		MODALIDAD(2)	.026	1	0.872
		MODALIDAD(3)	11.130	1	0.001
		MODALIDAD(4)	7.781	1	0.005
		MODALIDAD(5)	2.181	1	0.140
		MODALIDAD(6)	12.137	1	0.000494
		MODALIDAD(7)	6.609	1	0.010
		MODALIDAD(8)	3.080	1	0.079
		MODALIDAD(9)	79.531	1	4.7483E-19
		PROCEDENCIA	2.099	1	0.147
		TIPOCOLEGIO	6.546	1	0.011
		UBIGEOCOLEGIO	.960	1	0.327
		ESTADO CIVIL	.235	2	0.889
		ESTADO CIVIL(1)	.189	1	0.664
		ESTADO CIVIL(2)	.046	1	0.830
		INGRESO	.805	1	0.370
		SEXO(1)	58.684	1	1.851E-14
		NOTA	9.917	1	0.002
		EDAD	2.306	1	0.129
		TI_PREPA	3.973	1	0.046
		COMO_ENTERO	3.428	1	0.064
		MOT_POSTULAR	6.886	1	0.009
		TRABAJA	.191	1	0.662

	DEP_ECONO	.027	1	0.869
	VIVPADRES	.016	1	0.898
	NUM_HERM	4.021	1	0.045
	VIVECON	2.603	1	0.107
	DONDEVIVE	2.395	1	0.122
	ADMINISTRACION	69.284	1	8.5278E-17
	AGRONOMIA	21.589	1	0.000003
	CONTABILIDAD	74.610	1	5.7351E-18
	ECONOMIA	57.180	1	0.000
	AMBIENTAL	1.914	1	0.166
	ING SUELOS	.641	1	0.423
	INDUSTRIAS ALIMENTARIAS	11.490	1	0.001
	SISTEMAS	35.957	1	2.0175E-9
	RNR	.462	1	0.497
	FORESTAL	14.981	1	0.000109
	MECANICA	12.827	1	0.000342
	ZOOTECNIA	4.609	1	0.032
	PRIMER SEMESTRE	53.041	1	3.2658E-13

Nota. La tabla muestra la significancia de la correlación de los atributos con la condición del alumno.

De los diferentes datos extraídos donde se observa indicadores sociales, económicos y académicos, se puede observar de los datos que existen atributos que son significativos cuyo nivel de confianza supera el 95% ya que se obtuvo un valor- P menor a 0.05.

Tabla 35

Pruebas ómnibus de coeficientes de modelo

	Chi-cuadrado	gl	Sig.
Paso	624.843	39	0.000
Bloque	624.843	39	0.000
Modelo	624.843	39	0.000

Nota. La tabla muestra que la significancia del modelo.

Como se puede observar e la prueba de ómnibus de regresión logística los valores son significativos para el modelo con lo cual comprobamos en los coeficientes que al menos uno de ello no es cero, esto nos demuestra que la variable rendimiento académico esta explicado por al menos una de las variables.

Tabla 36

Significancia de las variables en el rendimiento academico

	Error estándar	gl	Sig.	Exp(B)
MODALIDAD		9	6.4827E-24	
MODALIDAD(1)	0.476	1	0.062	2.431
MODALIDAD(2)	0.801	1	0.100	3.731
MODALIDAD(3)	0.683	1	0.000026	17.678
MODALIDAD(4)	1.188	1	0.353	.331
MODALIDAD(5)	0.528	1	0.056	2.750
MODALIDAD(6)	0.600	1	0.614	.739
MODALIDAD(7)	0.504	1	0.009	3.740
MODALIDAD(8)	28192.975	1	0.999	.000

MODALIDAD(9)	0.487	1	0.000009	8.696
PROCEDENCIA	0.000	1	0.969	1.000
TIPOCOLEGIO	0.138	1	0.001	1.574
UBIGEOCOLEGIO	0.000	1	0.905	1.000
ESTADO CIVIL		2	0.661	
ESTADO CIVIL(1)	1.154	1	0.364	.351
ESTADO CIVIL(2)	1.355	1	0.955	1.079
INGRESO	0.204	1	0.000134	2.179
SEXO(1)	0.106	1	7.6492E-7	1.691
NOTA	0.040	1	1.2948E-7	1.233
EDAD	0.020	1	0.009	1.053
TI_PREPA	0.051	1	0.772	.985
COMO_ENTERO	0.034	1	0.162	1.048
MOT_POSTULAR	0.030	1	0.276	1.033
TRABAJA	0.075	1	0.681	1.031
DEP_ECONO	0.100	1	0.649	.956
VIVPADRES	0.076	1	0.423	.941
NUM_HERM	0.064	1	0.519	1.042
VIVECON	0.046	1	0.074	.920
DONDEVIVE	0.068	1	0.908	1.008
ADMINISTRACION	0.274	1	8.0991E-11	5.950

AGRONOMIA	0.227	1	0.024	.599
CONTABILIDAD	0.283	1	7.5579E-10	5.692
ECONOMIA	0.271	1	6.2453E-7	3.866
AMBIENTAL	0.245	1	0.042	.607
ING SUELOS	0.235	1	0.763	1.073
INDUSTRIAS ALIMENTARIAS	0.238	1	0.016	.562
SISTEMAS	0.246	1	0.002	.465
RNR	0.243	1	0.722	1.090
FORESTAL	0.238	1	0.056	.635
MECANICA	0.333	1	0.000226	.293
PRIMER SEMESTRE	0.000	1	5.0437E-19	1.000
Constante	20.540	1	4.5321E-20	.000

Nota. La tabla muestra la significancia de las variables y cuáles son las que explican el rendimiento académico.

En el análisis los atributos de las dimensiones sociales, económicos y académicos relacionados en con rendimiento académico de los estudiantes ingresantes se observa que los indicadores académicos tienen mayor incidencia en el rendimiento académico, dentro de ellos tenemos a la modalidad de ingreso, a la opción de ingreso, la nota de ingreso y a la carrera que ingresaron todos ellos con un p-valor < 0.05.

Los indicadores sociales como sexo y la edad tienen una incidencia en el rendimiento académico ya que tienen un p-valor < 0.05. Del indicador económico el tipo de colegio tiene incidencia en el rendimiento académico el p-valor es menor a 0.05.

En la tabla observamos que la procedencia y el ubigeo de colegio no son significativos para el modelo, también el estado civil y el tipo de preparación no tienen una significancia para la predicción del rendimiento académico. El tipo de preparación (TI_PREPA) está

relacionado con el rendimiento académico, pero no es significativo para el modelo, también como se enteró del examen de la UNAS (COMO_ENTERO) no es significativo para el modelo. El motivo para postular del estudiante (MOT_POSTULAR) está relacionado con el rendimiento académico, pero no es significativo para el modelo. Los atributos: trabaja (TRABAJA), de quien depende económicamente (DEP_ECONO), si sus padres están vivos (VIVPADRES), el número de hermanos que tiene (NUM_HERM), con quien vive actualmente (VIVECON) y el lugar donde vive (DONDEVIVE) no son significativos para el modelo.

En la siguiente tabla detallamos los atributos que tienen mayor incidencia en el rendimiento académico:

HI: Los algoritmos de aprendizaje automático de la minería de datos pueden predecir el rendimiento académico de los estudiantes universitarios de primer ciclo de la Universidad Nacional Agraria de la Selva.

En la tabla 33 observamos que los algoritmos que mejor pueden predecir el rendimiento académico de los alumnos ingresantes son ramdom forest con un 29,75% también el agoritmo de regresión logística con un acuracy de 71,71%, mientras que el algoritmo emsamblados vote tiene un acuracy de 71,08%

Tabla 37

Atributos que inciden en el rendimiento académico

	Error estándar	gl	Sig.	Exp(B)
MODALIDAD		9	6.4827E-24	
MODALIDAD(3)	0.683	1	0.000026	17.678
TIPOCOLEGIO	0.138	1	0.001	1.574
INGRESO	0.204	1	0.000134	2.179
SEXO(1)	0.106	1	7.6492E-7	1.691
NOTA	0.040	1	1.2948E-7	1.233
EDAD	0.020	1	0.009	1.053
ADMINISTRACION	0.274	1	8.0991E-11	5.950
AGRONOMIA	0.227	1	0.024	.599
CONTABILIDAD	0.283	1	7.5579E-10	5.692
ECONOMIA	0.271	1	6.2453E-7	3.866
AMBIENTAL	0.245	1	0.042	.607
INDUSTRIAS ALIMENTARIAS	0.238	1	0.016	.562
SISTEMAS	0.246	1	0.002	.465
MECANICA	0.333	1	0.000226	.293
PRIMER SEMESTRE	0.000	1	5.0437E-19	1.000

Nota. La tabla muestra los atributos que tienen mayor incidencia en el rendimiento académico.

En la tabla se puede observar que de los indicadores sociales el sexo y la edad son significativos para el modelo. De los indicadores económicos el tipo de colegio es significativo. Los que mayor incidencia tienen son los indicadores académicos, la modalidad de ingreso, la nota de ingreso, tipo de colegio, la opción de ingreso son los más significativos y por lo tanto tienen mayor incidencia en el rendimiento académico. Como se observa en la tabla los alumnos que ingresan por la modalidad (3) tienen una alta probabilidad de aprobar el semestre.

5.3. DISCUSIÓN DE RESULTADOS

- La presente investigación tuvo como objetivo predecir con las las técnicas de minería de datos predecir el rendimiento académico de los alumnos ingresantes a la UNAS mediante los datos recolectados en las oficinas de admisión y DICCA. De todos los modelos evaluados el que tiene mayor exactitud (acuracy) es el modelo de regresión logística con 72,79% al momento del entrenamiento. Por otro lado Candia encontró que el algoritmo de árbol de decisión "Random Forest" tiene una mejor performance para la predicción del 69% mientras que Holgado encontró que el algoritmo que tiene mejor rendimiento es el C5.0 el cual tiene una mejor exactitud de clasificación (acuracy) de 77.8%,
- Hay que tener en cuenta que para este tipo de investigación es muy importante la calidad de la data para poder realizar las predicciones, fue necesario realizar operaciones para tener una mejor calidad en la data, así obtenemos mejores resultados y cumplir con los objetivos que es ver que indicadores tienen mayor incidencia en la predicción del rendimiento académico, se logró determinar que indicadores tienen mayor incidencia en el rendimiento académico Por una lado Yamao (2018) afirma que las variables más influyentes en el rendimiento académico son los siguientes: Nota del examen de admisión, la edad, el sexo, la distancia que se encuentran de la universidad, el colegio de excelencia. Además, Candia (2109) encontró que los factores claves son: la nota de ingreso, la escuela profesional que se estudia, el semestre, el género y la modalidad de ingreso son los que determinan en rendimiento académico de los estudiantes.
- Entre las los indicadores que mayor incidencia tienen en el rendimiento académico tenemos la ***modalidad de ingreso*** como la variable que más influencia tiene, en segundo lugar, tenemos la ***nota de ingreso***, luego teneos el sexo, la edad, el tipo de colegio, la carrera a la que ingreso y el semestre que estudia, por el contrario para Holgado (2018) la variable que más influye en el rendimiento académico es la cantidad de ***asignaturas cursadas*** seguido por la variable ***servicio de comedor universitario*** y finalmente la variable que también influye es la ***carrera profesional*** a la que ingresa. Para él autor estas tres variables son las más influyentes.

5.4. APORTE CIENTÍFICO A LA INVESTIGACIÓN

Los resultados de la investigación muestran que es posible predecir el rendimiento académico con los datos de los estudiantes recopilados al momento de la inscripción en el proceso de admisión usando técnicas de minería de datos. El principal aporte es haber encontrado indicadores sociales, económicos y académicos que nos permiten tener una predicción del rendimiento académico para una toma de decisiones oportuna por los órganos de gobierno de la universidad.

CONCLUSIONES

- El rendimiento académico de los estudiantes es un tema muy complejo y con los indicadores de ingreso como económicos, sociales o académicos, a través de la aplicación de técnicas de minería de datos y la metodología CRISP-DM, usando la aplicación de diferentes técnicas de minería de datos se logró predecir el rendimiento de los estudiantes del primer ciclo de la universidad nacional agraia de la selva.
- Se logró determinar los indicadores sociales, económicos y académicos que más influyen en la predicción del rendimiento académico de los estudiantes, tales como la nota de ingreso, la modalidad de ingreso, la edad, sexo, el tipo de colegio, la opción de ingreso.
- Se determino que los algoritmos de aprendizaje automático de las técnicas de la minería de datos que pueden predecir el rendimiento académico de los alumnos ingresantes; es la ***regresión logística*** que llega a una exactitud (acuracy) de 72.79 %. También otro algoritmo muy significativo es los ensamblados vote con una exactitud de 71.09%.

SUGERENCIAS

- A las oficinas de admisión de la UNAS, generar nuevas políticas y mejorar la obtención de la información de los postulantes para mejorar los porcentajes de predicción.
- Usar la metodología CRISP-DM, es una metodología que nos brinda una guía a través de 6 fases muy bien definidas en procesos de modelos predictivos, es una herramienta muy sencilla de entender independientemente de herramienta de minería de datos a usar y algoritmo de aprendizaje automático a usar.
- A las autoridades de la Universidad Nacional Agraria de la Selva implementar acciones para mejorar el rendimiento académico de los estudiantes poniendo énfasis a los alumnos que tienen mayor probabilidad de salir desaprobados y tomar medidas correctivas para brindarles asesorías para tener mejor calidad y éxito.
- Usar las técnicas de minería de datos y los algoritmos de aprendizaje automático de acuerdo a la data que se tiene y agregando otras variables de los datos históricos, sería una herramienta para apoyar a la toma de decisiones y a mejoras las políticas para tener un mejor rendimiento académico, además que esto permitirá tener un impacto en la sociedad y tener una mejor imagen de nuestros alumnos.
- Para futuras investigaciones se recomienda incluir variables como los cursos que se dictan, y los profesores encargados, así como su clasificación en SISTEMA DE FOCALIZACION DE HOGARES (sisfoh).

REFERENCIAS

Alania Ricaldi, F. (2018). *APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA PREDECIR LA DESERCIÓN ESTUDIANTIL DE LA FACULTAD DE INGENIERÍA DE LA UNIVERSIDAD NACIONAL DANIEL ALCIDES CARRIÓN.* PASCO: Universidad Nacional Daniel Alcides Carrión.

Aluja, T. (2001). LA MINERA DE DATOS, ENTRE LA ESTADISTICA Y LA INTELIGENCIA ARTIFICIAL. *Qüestiió: quaderns d'estadística i investigació operativa,*, 479-498. Retrieved 11 de octubre de 2021, from https://www.researchgate.net/publication/28177489_La_mineria_de_datos_entre_la_estadistica_y_la_inteligencia_artificial

Beneyto Sánchez, S. (2015). *Entorno Familiar y Rendimiento.* Alicante: ÁREA DE INNOVACIÓN Y DESARROLLO, S.L.

Brito-Jiménez, I., & Palacio-Sañudo, J. (2016). Calidad de vida, desempeño académico y variables sociodemográficas en estudiantes universitarios de Santa Marta-Colombia. *Duazary, 13(2),* 133–141.

Candia Oviedo, D. (2019). *PREDICCIÓN DEL RENDIMIENTO ACADÉMICO DE LOS ESTUDIANTES DE LA UNSAAC A PARTIR DE SUS DATOS DE INGRESO UTILIZANDO ALGORITMOS DE APRENDIZAJE AUTOMÁTICO.* Cusco: Repositorio Institucional - UNSAAC.

Carrasco Diaz, S. (2008). *MÉTODOLOGIA DE LA INVESTIGACION CIENTIFICA.* Lima: San Marcos.

Cerda, J., & VillarroelL, L. (2008). Evaluación de concordancia inter-observador en investigación pediátrica: coeficiente de Kappa. *Revista chilena de pediatría,* 54-58. Retrieved 11 de Octubre de 2021, from https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0370-41062008000100008

Charte, F. (17 de noviembre de 2020). *campusmvp*. https://www.campusmvp.es/recursos/post/el-proceso-de-extraccion-de-conocimiento-a-partir-de-bases-de-datos.aspx

Charte, F. (17 de Noviembre de 2020). *Campusmvp*. https://www.campusmvp.es/recursos/post/el-proceso-de-extraccion-de-conocimiento-a-partir-de-bases-de-datos.aspx

Díaz, M., Urquijo, P., Arias Blanco, J., Escudero Escorza, T., Rodriguez Espinar, S., & Vidal García, J. (2002). Evaluación del rendimiento en la enseñanza superior. Comparación de resultados entre alumnos procedentes de la LOGSE y del COU. *Revista de Investigación Educativa*, 357–383.

Fayyad, U., Piatetsky-Shapiro, G., & Padhraic, S. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework. *ASOCIACIÓN PARA EL AVANCE DE LA INTELIGENCIA ARTIFICIAL*, 83-84.

Fuentes Quiñonez, J., Cortez Vásquez, A., & Jara Rivas, D. (2014). *Aplicación de Minería de datos a un Sistema de Mantenimiento Predictivo de Detección de Incendios.* Lima: Universidad Nacional Mayor de San Marcos.

Gallardo Arancibia, J. (2009). *Metodología para la definición de requisitos en proyectos de data mining.* Madrid: Biblioteca de la Universidad Politécnica de Madrid.

Garcia Oritz, Y. (214). Estudiantes universitarios con bajo rendimiento académico, ¿qué hacer? *EDUMECENTRO*, 6.

García Ortiz, Y., López de Castro Machado, D., & Rivero Frutos, O. (2014). Estudiantes universitarios con bajo rendimiento académico,. *EDUMECENTRO*, 272.

González, R. (1989). *Análisis de las causas del fracaso escolar en la Universidad Politécnica de Madrid.* Madrid: España.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques.* AMSTERDAM : Morgan Kaufmann is an imprint of Elsevier.

Hernandez, J. (2004). *Introducción a la minería de datos.* Valencia: España.

Holgado Apaza, L. A. (2018). *DETECCIÓN DE PATRONES DE BAJO RENDIMIENTO ACADÉMICO MEDIANTE TÉCNICAS DE MINERÍA DE DATOS DE LOS ESTUDIANTES DE LA UNIVERSIDAD NACIONAL AMAZÓNICA DE MADRE DE DIOS 2018.* PUNO: Repositorio Institucional UNA-PUNO.

Lantz, B. (2013). *Machine learning with R.* Packt Publishing Ltd.

Márquez Vera, C. (2015). *PREDICCIÓN DEL FRACASO Y EL ABANDONO ESCOLAR MEDIANTE TÉCNICAS DE MINERÍA DE DATOS.* Córdoba: UNIVERSIDAD DE CÓRDOBA. https://core.ac.uk/download/pdf/60900921.pdf

Menacho Chiok, C. (2017). Predicción del rendimiento académico aplicando técnicas de minería de datos. *Revitas la molina*, 26.

Microsoft. (13 de setiembre de 2021). *Microsoft*. https://docs.microsoft.com/eses/sql/analysis-services/data-mining/data-mining-concepts?view=sql-server2017

Moine, J., Gordillo, S., & Haedo, A. (18 de Julio de 2012). *Análisis comparativo de metodologías para la gestión de proyectos de minería de datos*. http://sedici.unlp.edu.ar/handle/10915/18749

Mondragon, R. (2007). *EXPLORACIONES SOBRE EL SOPORTE MULTI-AGENTE BDI EN EL PROCESO DE DESCUBRIMIENTO DE CONOCIMIENTO EN BASES DE DATOS. (Tesis para Optar el Grado de Magister).* Mexico.: UNIVERSIDAD VERACRUZANA.

Moreno Garcia, M., Miguel Quintales, L. A., García Peñalvo, F. J., & Polo Martín, M. J. (2001). *APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS EN LACONSTRUCCIÓN Y VALIDACIÓN DE MODELOS PREDICTIVOS Y ASOCIATIVOS A PARTIR DE ESPECIFICACIONES DE REQUISITOS DE SOFTWARE.* Retrieved 11 de octubre de 2021, from researchgate: https://www.researchgate.net/publication/220958273_Aplicacion_de_Tecnicas_de_Mineria_de_Datos_en_la_Construccion_y_Validacion_de_Modelos_Predictivos_y_Asociativos_a_Partir_de_Especificaciones_de_Requisitos_De_Software

Nettleton, D. (2003). *Analisis de datos comerciales. Madrid: Diaz de santos S.A. Pedro Isasi Viñuela.* Madrid: PEARSON S.A.

Oñate, A. (2016). *Análisis de la Deserción y Permanencia Académica en la Educación Superior Aplicando Minería De Datos.* Bogotá: Universidad Nacional de Colombia. https://repositorio.unal.edu.co/bitstream/handle/unal/57387/alvaroagustino%c3%b1atebowen.2016.pdf?sequence=1&isAllowed=y

Ordoñes Leyva, Y., & Grass Boada, D. (2011). HERMINWEB: Herramienta de Minería de Uso de la Web Aplicado a los Registros del Proxy. *ResearchGate*.

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis. *Expert Systems with Applications*, 1432-1462.

Perez, C., & Santin, D. (2007). *Minería de datos. Técnicas y herramientas.* Madrid: Thomson.

Ramesh, V., Parkavi, P., & Ramar, K. (2014). Predicting Student Performance:. *International Journal of Computer Applications*, 5.

Reyes Tejada, Y. (2003). *RELACIÓN ENTRE EL RENDIMIENTO ACADÉMICO, LA ANSIEDAD ANTE LOS EXÁMENES,LOS RASGOS DE PERSONALIDAD, EL AUTOCONCEPTO Y LA ASERTIVIDADEN ESTUDIANTES DEL PRIMER AÑO DE PSICOLOGÍA DE LA UNMSM.* Lima: Oficina General del Sistema de Bibliotecas y Biblioteca Central. Retrieved 12 de octubre de 2021.

Rodríguez, Á., & Arenas, D. (2016). Programas de intervención para Estudiantes Universitarios con bajo rendimiento académico. *Informes Psicológicos, 16(1)*, 13-34.

Rosado Gómez, A. A., & Verjel Ibáñez, A. (2014). Minería de datos aplicada a la demanda del transporte aéreo en Ocaña, Norte de Santander. *Revista Tecnura*, 101-113. doi:http://dx.doi.org/10.14483/udistrital.jour.tecnura.2015.3.a08

S. C, D. (2008). *Métodologia de la Investigacion Cientifica.* Lima: San Marcos: Sposito.

Solano, L. (2015). *Rendimiento académico de los estudiantes de secundaria obligatoria y su relación con las aptitudes mentales y las actitudes ante el estudio.* ESPAÑA: UNED.

Tejedor, F. (1998). *Los alumnos de la Universidad de Salamanca. Características y rendimiento académico Universidad de Salamanca.* Salamanca: España.

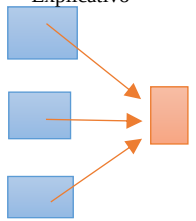
Timarán Pereira, S., Hernández Arteaga, I., Caicedo-Zambrano, S., Hidalgo Troya, A., & Alvarado Pérez, J. (2016). *El proceso de descubrimiento de conocimiento en bases de datos. En Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas.* Bogotá: Ediciones Universidad Cooperativa de Colombia. https://doi.org/http://dx.doi.org/10.16925/9789587600490

Yamao, E. (2018). *PREDICCIÓN DEL RENDIMIENTO ACADÉMICO MEDIANTE MINERÍA DE DATOS EN ESTUDIANTES DEL PRIMER CICLO DE LA ESCUELA PROFESIONAL DE INGENIERÍA DE COMPUTACIÓN Y SISTEMAS, UNIVERSIDAD DE SAN MARTÍN DE PORRES, LIMA-PERÚ.* LIMA. https://repositorio.usmp.edu.pe/bitstream/handle/20.500.12727/3555/yamao_e.pdf?sequence=3&isAllowed=y

ANEXOS

ANEXO 01. Matriz de consistencia

TITULO: RENDIMIENTO ACADÉMICO MEDIANTE TÉCNICAS DE MINERÍA DE DATOS EN ESTUDIANTES DE LA UNIVERSIDAD NACIONAL AGRARIA DE LA SELVA

PROBLEMA	OBJETIVOS	HIPÓTESIS	VARIABLES	DIMENSIONES	INDICADORES	METODOLOGÍA
PROBLEMA GENERAL ¿Cómo la minería de datos puede predecir el rendimiento académico de los estudiantes del primer ciclo de la Universidad Nacional agraria de la selva?	**OBJETIVO GENERAL** Predecir con las técnicas de la minería de datos el rendimiento académico de los estudiantes del primer ciclo de la Universidad Nacional agraria de la selva.	**Hipótesis de investigación** HI: Con las técnicas de la minería de datos se puede predecir el rendimiento académico de los estudiantes del primer ciclo de la Universidad Nacional Agraria de la Selva, Tingo María 2021 H0: Con las técnicas de la minería de datos no se puede predecir el rendimiento académico de los estudiantes del primer ciclo de la Universidad Nacional Agraria de la Selva, Tingo María 2021..	**VARIABLE INDEPENDIENTE** **Minería de datos**	Indicadores sociales Indicadores económicos Indicadores académicos	• Sexo • Edad • Provincia • Colegio de procedencia • Financiamiento de estudios • Tipo de colegio • Puntaje examen de ingreso • Facultad • Modalidad de ingreso	Tipo: Aplicada Nivel: No experimental **Diseño** Explicativo 
PROBLEMAS ESPECIFICOS 1. ¿Cuáles son los indicadores sociales, económicos y académicos que tienen mayor incidencia para predecir el rendimiento académico en estudiantes universitarios de primer ciclo de la Universidad Nacional Agraria de la Selva? 2. ¿Qué algoritmos de aprendizaje automático de la minería de datos son capaces de predecir el rendimiento académico de los estudiantes del primer semestre de la Universidad Nacional Agraria de la Selva?	**OBEJETIVOS ESPECIFICOS** 1.Estimar indicadores sociales, económicos y académicos para el rendimiento académico en estudiantes universitarios de primer ciclo de la Universidad Nacional Agraria de la Selva. 2. Determinar los algoritmos de aprendizaje automático de la minería de datos que pueden predecir el rendimiento académico en estudiantes universitarios de primer ciclo de la Universidad Nacional Agraria de la Selva.	**HIPÓTESIS ESPECIFICAS** 1.HI: El rendimiento académico en estudiantes universitarios de primer ciclo puede medirse mediante indicadores sociales, económicos y académicos de la Universidad Nacional Agraria de la Selva. H0: El rendimiento académico en estudiantes universitarios de primer ciclo no puede estimarse mediante indicadores sociales, económicos y académicos de la Universidad Nacional Agraria de la Selva 2. HI: Los algoritmos de aprendizaje automático de la minería de datos pueden predecir el rendimiento académico de los estudiantes universitarios de primer ciclo de la Universidad Nacional Agraria de la Selva. H0: Los algoritmos de aprendizaje automático de la minería de datos no pueden predecir el rendimiento académico de los estudiantes universitarios de primer ciclo de la Universidad Nacional Agraria de la Selva.	**VARIABLE DEPENDEDIENTE** Rendimiento Académico	Predicción Nivel de concordancia	• Aprobado • Desaprobado Medida de concordancia	

ANEXO 02. CONSENTIMIENTO INFORMATIVO

UNIVERSIDAD NACIONAL HERMILIO VALDIZÁN

ESCUELA DE POSGRADO

MAESTRÍA EN INGENIERÍA DE SISTEMAS, MENCIÓN EN TECNOLOGÍA DE INFORMACIÓN Y COMUNICACIÓN

DOCUMENTO DE AUTORIZACIÓN POR LA AUTORIDAD DE LA UNIVERSIDAD NACIONAL AGRARIA DE LA SELVA.

Título de la Investigación:	RENDIMIENTO ACADÉMICO MEDIANTE TÉCNICAS DE MINERÍA DE DATOS EN ESTUDIANTES DE LA UNIVERSIDAD NACIONAL AGRARIA DE LA SELVA
Objetivo:	Predecir con las técnicas de la minería de datos el rendimiento académico de los estudiantes del primer ciclo de la Universidad Nacional Agraría de la Selva.
Investigador:	PONCE GUIZABALO SANTOS VICTOR

A través del presente documento se le hace de su conocimiento a su persona que se realizara la predicción del rendimiento académico de los alumnos ingresantes a la UNAS para los cuales se usara los datos de los alumnos ingresantes en el periodo 2015 a 2019, tener en cuenta que no se usara ningún dato que afecte su afecte su identidad personal, es por ello que mi persona solicita a usted el permiso para desarrollar esta investigación sin ningún problema para tal efecto plasmara su firma en virtud que si acepta y autoriza el permiso para el acceso a los datos de los alumnos ingresantes.

Firma de la autoridad competente

ANEXO 03. INSTRUMENTOS

UNIVERSIDAD NACIONAL "HERMILIO VALDIZÁN"

ESCUELA DE POST GRADO - Maestría en Ingeniería de sistemas, mención en tecnologías de información y comunicación

FICHA DE ANÁLISIS DOCUMENTAL PARA LA TESIS:

"RENDIMIENTO ACADÉMICO MEDIANTE TÉCNICAS DE MINERÍA DE DATOS EN ESTUDIANTES DE LA UNIVERSIDAD NACIONAL AGRARIA DE LA SELVA"

UNIVERSIDAD:

_____

Datos de la aplicación:

Fecha de aplicación 1: _____ / _____ / _____

Documentos Revisados de la Universidad:

Documentos	Tiene		Se revisó	
	SI	NO*	SI	NO
Datos Sociales				
Datos Económicos				
Datos Académicos				

Los datos re recolectan a través de una encuesta a los postulantes y luego a los ingresantes, los datos académicos si aprobó o desaprobó el ciclo son registrados al finalizar el ciclo por la Dirección de Coordinación y Desarrollo Académico (DICDA). Todos los datos están en una base de datos.

1. Sobre Datos Sociales - Oficina admisión
   a) Se registra la edad del postulante Si ( ) No ( )
   b) Se registra el sexo del postulante Si ( ) No ( )
      Masculino ( ) Femenino ( )
   c) Se registra si Traja el postulante Si ( ) No ( )
      1) NO 2) Si, Tiempo Completo
      3) Si, por horas
   d) Se registra de quien depende económicamente Si ( ) No ( )
      1) sus padres 2) Parientes
      3) Si mismo 4) Otros
   e) Se registra el lugar donde vive Si ( ) No ( )
      1) Ciudad 2) Pueblo joven
      3) Zona Rural
   f) Se registra el lugar de procedencia (ubigeo) Si ( ) No ( )
      Departamento: _____
      Provincia: _____
      Distrito: _____
2. Sobre los Datos Económicos
   a) Se registra el Tipo de colegio Si ( ) No ( )
      1) Estatal 2) Particular
3. Sobre los Datos Académicos
   a) Se Registra la Facultad que ingresa Si ( ) No ( )
      - INGENIERIA AMBIENTAL ( )
      - ADMINISTRACION ( )
      - Agronomía ( )
      - CONTABILIDAD ( )
      - ECONOMIA ( )
      - INGENIERIA FORESTAL ( )
      - INGENIERIA EN INDUSTRIAS ALIMENTARIAS ( )
      - INGENIERIA MECANICA ELECTRICA ( )
      - INGENIERIA EN RECURSOS NATURALES RENOVABLES ( )
      - INGENIERIA EN INFORMATICA Y SISTEMAS ( )
      - INGENIERIA EN CONSERVACION DE SUELOS Y AGUA ( )
      - ZOOTECNIA ( )
   b) Se registra el tipo de preparación para su postulación Si ( ) No ( )
      1) AUTOESTUDIO 2) PROFESOR PARTICULAR
      3) ACADEMIA 4) OTROS

c) Se registra la modalidad de ingreso Si ( ) No ( )

EXAMEN ORDINARIO ( )
Primero y segundo puesto de educación secundaria ( )
GRADUADOS Y TITULADOS ( )
CENTRO PRE UNIVERSITARIO ( )
VICTIMAS DE TERRORISMO ( )
PERSONAS CON DISCAPACIDAD ( )
DEPORTISTAS ( )
ARTE Y CULTURA ( )
COMUNIDADES NATIVAS ( )
CONVENIOS ESPECIALES ( )
TALENTO BECA 18 ( )

d) Se registra la Puntaje de Ingreso Si ( ) No ( )
_____

e) Se registra de notal Final del I ciclo académico Si ( ) No ( )
_____

f) Se registra el colegio de Procedencia Si ( ) No ( )
_____

ANEXO 04. Validación de los instrumentos por expertos

VALIDACIÓN DE INSTRUMENTO

VALIDACIÓN DEL INSTRUMENTO

Nombre del experto: Mg. Elmer Chapoñan Saldivar

Especialidad: Especialista Ing. Sistemas.

DIMENSIÓN	ÍTEM	RELEVANCIA	COHERENCIA	SUFICIENCIA	CLARIDAD
1. Sobre Datos Sociales - Oficina admisión					
a) Se registra la edad del postulante	Si ( ) No ( )				
b) Se registra el sexo del postulante	Si ( ) No ( )	4	4	4	4
Masculino ( ) Femenino ( )					
c) Se registra si Traja el postulante	Si ( ) No ( )				
1) NO 2) Si, Tiempo Completo					
3) Si, por horas					
d) Se registra de quien depende económicamente	Si ( ) No ( )				
1) sus padres 2) Parientes					
3) Si mismo 4) Otros					
e) Se registra el lugar donde vive	Si ( ) No ( )				
1) Ciudad 2) Pueblo joven					
3) Zona Rural					
f) Se registra el lugar de procedencia (ubigeo)	Si ( ) No ( )				
Departamento: _____					
Provincia: _____					
Distrito: _____					
2. Sobre los Datos Económicos					
a) Se registra el Tipo de colegio	Si ( ) No ( )	4	4	4	4
1) Estatal 2) Particular					
3. Sobre los Datos Académicos					
a) Se Registra la Facultad que ingresa	Si ( ) No ( )				
b) Se registra el tipo de preparación para su postulación	Si ( ) No ( )	4	4	4	4
1) AUTOESTUDIO 2) PROFESOR PARTICULAR					
3) ACADEMIA 4) OTROS					
c) Se registra la modalidad de ingreso	Si ( ) No ( )				

d) Se registra la Puntaje de Ingreso Si ( ) No ( ) e) Se registra de notal Final del I ciclo académico Si ( ) No ( ) f) Se registra el colegio de Procedencia Si ( ) No ( )	4	4	4	4

¿Hay alguna dimensión o ítem que no fue evaluada? SÍ ( ) NO (X) En caso de que sí, ¿Qué dimensión falta?

Decisión del experto: El instrumento debe ser aplicado SI

VALIDACIÓN DE INSTRUMENTO

VALIDACIÓN DEL INSTRUMENTO

Nombre del experto: Dra. JESUS TOLENTINO, INES E.

Especialidad: Dra. Gestion Empresarial

DIMENSIÓN	ÍTEM	RELEVANCIA	COHERENCIA	SUFICIENCIA	CLARIDAD
1. Sobre Datos Sociales - Oficina admisión					
a) Se registra la edad del postulante	Si ( ) No ( )	4	4	4	4
b) Se registra el sexo del postulante	Si ( ) No ( )				
Masculino ( ) Femenino ( )					
c) Se registra si Traja el postulante	Si ( ) No ( )				
1) NO 2) Si, Tiempo Completo					
3) Si, por horas					
d) Se registra de quien depende económicamente	Si ( ) No ( )				
1) sus padres 2) Parientes					
3) Si mismo 4) Otros					
e) Se registra el lugar donde vive	Si ( ) No ( )				
1) Ciudad 2) Pueblo joven					
3) Zona Rural					
f) Se registra el lugar de procedencia (ubigeo)	Si ( ) No ( )				
Departamento: _____					
Provincia: _____					
Distrito: _____					
2. Sobre los Datos Económicos					
a) Se registra el Tipo de colegio	Si ( ) No ( )	4	4	4	4
1) Estatal 2) Particular					
3. Sobre los Datos Académicos					
a) Se Registra la Facultad que ingresa	Si ( ) No ( )				
b) Se registra el tipo de preparación para su postulación	Si ( ) No ( )				
1) AUTOESTUDIO 2) PROFESOR PARTICULAR					
3) ACADEMIA 4) OTROS					
c) Se registra la modalidad de ingreso	Si ( ) No ( )	4	4	4	4

d) Se registra la Puntaje de Ingreso Si ( ) No ( ) e) Se registra de notal Final del I ciclo académico Si ( ) No ( ) f) Se registra el colegio de Procedencia Si ( ) No ( )	4	4	4	4

¿Hay alguna dimensión o ítem que no fue evaluada? SÍ ( ) NO (X) En caso de que sí, ¿Qué dimensión falta?

Decisión del experto: El instrumento debe ser aplicado Si.

VALIDACIÓN DE INSTRUMENTO

VALIDACIÓN DEL INSTRUMENTO

Nombre del experto: JIMMY GLOVER FLORES VIDAL

Especialidad: MAESTRO GESTIÓN TECNOLOGICA EMPRESARIAL

DIMENSIÓN	ÍTEM	RELEVANCIA	COHERENCIA	SUFICIENCIA	CLARIDAD
1. Sobre Datos Sociales - Oficina admisión					
a) Se registra la edad del postulante	Si ( ) No ( )				
b) Se registra el sexo del postulante	Si ( ) No ( )				
Masculino ( ) Femenino ( )					
c) Se registra si Traja el postulante	Si ( ) No ( )	4	4	4	4
1) NO 2) Si, Tiempo Completo					
3) Si, por horas					
d) Se registra de quien depende económicamente	Si ( ) No ( )				
1) sus padres 2) Parientes					
3) Si mismo 4) Otros					
e) Se registra el lugar donde vive	Si ( ) No ( )				
1) Ciudad 2) Pueblo joven					
3) Zona Rural					
f) Se registra el lugar de procedencia (ubigeo)	Si ( ) No ( )				
Departamento: _____					
Provincia: _____					
Distrito: _____					
2. Sobre los Datos Económicos					
a) Se registra el Tipo de colegio	Si ( ) No ( )	4	4	4	4
1) Estatal 2) Particular					
3. Sobre los Datos Académicos					
a) Se Registra la Facultad que ingresa	Si ( ) No ( )				
b) Se registra el tipo de preparación para su postulación	Si ( ) No ( )				
1) AUTOESTUDIO 2) PROFESOR PARTICULAR					
3) ACADEMIA 4) OTROS					
c) Se registra la modalidad de ingreso	Si ( ) No ( )	4	4	4	4

d) Se registra la Puntaje de Ingreso Si ( ) No ( ) e) Se registra de notal Final del I ciclo académico Si ( ) No ( ) f) Se registra el colegio de Procedencia Si ( ) No ( )	4	4	4	4

¿Hay alguna dimensión o ítem que no fue evaluada? SÍ ( ) NO (X) En caso de que sí, ¿Qué dimensión falta?

Decisión del experto: El instrumento debe ser aplicado SI

VALIDACIÓN DE INSTRUMENTO

VALIDACIÓN DEL INSTRUMENTO

Nombre del experto: Dra. Heidy Yelsy Rivera Vidal de Sánchez

Especialidad: Ingeniería de Sistemas

DIMENSIÓN	ÍTEM	RELEVANCIA	COHERENCIA	SUFICIENCIA	CLARIDAD
1. Sobre Datos Sociales - Oficina admisión					
a) Se registra la edad del postulante	Si ( ) No ( )	4	4	4	4
b) Se registra el sexo del postulante	Si ( ) No ( )				
Masculino ( ) Femenino ( )					
c) Se registra si Traja el postulante	Si ( ) No ( )				
1) NO 2) Si, Tiempo Completo					
3) Si, por horas					
d) Se registra de quien depende económicamente	Si ( ) No ( )				
1) sus padres 2) Parientes					
3) Si mismo 4) Otros					
e) Se registra el lugar donde vive	Si ( ) No ( )				
1) Ciudad 2) Pueblo joven					
3) Zona Rural					
f) Se registra el lugar de procedencia (ubigeo)	Si ( ) No ( )				
Departamento: _____					
Provincia: _____					
Distrito: _____					
2. Sobre los Datos Económicos					
a) Se registra el Tipo de colegio	Si ( ) No ( )	4	4	4	4
1) Estatal 2) Particular					
3. Sobre los Datos Académicos					
a) Se Registra la Facultad que ingresa	Si ( ) No ( )				
b) Se registra el tipo de preparación para su postulación	Si ( ) No ( )				
1) AUTOESTUDIO 2) PROFESOR PARTICULAR		4	4	4	4
3) ACADEMIA 4) OTROS					
c) Se registra la modalidad de ingreso	Si ( ) No ( )				

d) Se registra la Puntaje de Ingreso Si ( ) No ( ) e) Se registra de notal Final del I ciclo académico Si ( ) No ( ) f) Se registra el colegio de Procedencia Si ( ) No ( )		4	4	4	4

¿Hay alguna dimensión o ítem que no fue evaluada? SÍ ( ) NO (X) En caso de que sí, ¿Qué dimensión falta?

Decisión del experto: El instrumento debe ser aplicado SI

VALIDACIÓN DE INSTRUMENTO

VALIDACIÓN DEL INSTRUMENTO

Nombre del experto: Tomasa Veronica Cajos Bravo.

Especialidad: Metodologia Investigacion

DIMENSIÓN	ÍTEM	RELEVANCIA	COHERENCIA	SUFICIENCIA	CLARIDAD
1. Sobre Datos Sociales - Oficina admisión					
a) Se registra la edad del postulante	Si ( ) No ( )	4	4	4	4
b) Se registra el sexo del postulante	Si ( ) No ( )				
Masculino ( ) Femenino ( )					
c) Se registra si Traja el postulante	Si ( ) No ( )				
1) NO 2) Si, Tiempo Completo					
3) Si, por horas					
d) Se registra de quien depende económicamente	Si ( ) No ( )				
1) sus padres 2) Parientes					
3) Si mismo 4) Otros					
e) Se registra el lugar donde vive	Si ( ) No ( )				
1) Ciudad 2) Pueblo joven					
3) Zona Rural					
f) Se registra el lugar de procedencia (ubigeo)	Si ( ) No ( )				
Departamento: _____					
Provincia: _____					
Distrito: _____					
2. Sobre los Datos Económicos					
a) Se registra el Tipo de colegio	Si ( ) No ( )	4	4	4	4
1) Estatal 2) Particular					
3. Sobre los Datos Académicos					
a) Se Registra la Facultad que ingresa	Si ( ) No ( )		4	4	4
b) Se registra el tipo de preparación para su postulación	Si ( ) No ( )	4			
1) AUTOESTUDIO 2) PROFESOR PARTICULAR					
3) ACADEMIA 4) OTROS					
c) Se registra la modalidad de ingreso	Si ( ) No ( )				

d) Se registra la Puntaje de Ingreso	Si ( )	No ( )	4	4	4	4
e) Se registra de notal Final del I ciclo académico	Si ( )	No ( )				
f) Se registra el colegio de Procedencia	Si ( )	No ( )				

¿Hay alguna dimensión o ítem que no fue evaluada? SÍ ( ) NO (X) En caso de que sí, ¿Qué dimensión falta?

Decisión del experto: El instrumento debe ser aplicado SI

ANEXO 04. OTROS

VARIABLES EN EL ARCHIVO RDFF DE WEKA

@relation 'DATOS SIN MODIFICAR SIN DICOTOMIZAR SEPARADOS CARRERAS1-weka.filters.unsupervised.attribute.Remove-R1-2,4,36-weka.filters.unsupervised.attribute.Remove-R7'

@attribute MODALIDAD {Preuniversitario,Admision,'Exonerados Primeros Puestos','Convenios Especiales','Victimas del Terrorismo','Deportista Calificado','Exonerados Bachiller/Profesional','Arte y Cultura','Personas con Discapacidad','Exonerados Traslado Externo','Beca 18 - ME Huallaga','Comunidades Nativas','Exonerados Traslado Interno'}
@attribute PROCEDENCIA numeric
@attribute TIPOCOLEGIO numeric
@attribute UBIGEOCOLEGIO numeric
@attribute 'ESTADO CIVIL' {'S ',S,CO,C}
@attribute INGRESO numeric
@attribute SEXO {M,F}
@attribute NOTA numeric
@attribute EDAD numeric
@attribute TI_PREPA numeric
@attribute COMO_ENTERO numeric
@attribute MOT_POSTULAR numeric
@attribute TRABAJA numeric
@attribute DEP_ECONO numeric
@attribute VIVPADRES numeric
@attribute NUM_HERM numeric
@attribute VIVECON numeric
@attribute DONDEVIVE numeric
@attribute ADMINISTRACION numeric
@attribute AGRONOMIA numeric
@attribute CONTABILIDAD numeric
@attribute ECONOMIA numeric
@attribute AMBIENTAL numeric
@attribute 'ING SUELOS' numeric
@attribute 'INDUSTRIAS ALIMENTARIAS' numeric
@attribute SISTEMAS numeric
@attribute RNR numeric
@attribute FORESTAL numeric
@attribute MECANICA numeric
@attribute ZOOTECNIA numeric
@attribute 'PRIMER SEMESTRE' {01/01/2015,01/01/2017,01/01/2016,01/01/2018}
@attribute CONDICION {APROBADO,DESAPROBADO}

@data
Preuniversitario,100101,2,100601,'S ',1,M,11.19,24,1,6,3,1,2,1,3,1,1,0,0,0,0,0,0,0,0,1,0,0,0,01/01/2015,APROBADO
Admision,100601,1,100601,S,1,M,12.95,18,3,1,2,1,1,1,2,1,1,0,0,0,0,0,0,0,0,1,0,0,0,0,01/01/2015,DESAPROBADO

PREDICCION EN EL SOFTWARE WEKA

1	1:?	2:DESAPROBADO	0.596
2	1:?	1:APROBADO	0.583
3	1:?	1:APROBADO	0.557
4	1:?	2:DESAPROBADO	0.535
5	1:?	2:DESAPROBADO	0.82
6	1:?	2:DESAPROBADO	0.659
7	1:?	2:DESAPROBADO	0.808
8	1:?	2:DESAPROBADO	0.625
9	1:?	2:DESAPROBADO	0.769
10	1:?	2:DESAPROBADO	0.749
11	1:?	2:DESAPROBADO	0.562
12	1:?	2:DESAPROBADO	0.532
13	1:?	1:APROBADO	0.629
14	1:?	1:APROBADO	0.946
15	1:?	2:DESAPROBADO	0.54
16	1:?	2:DESAPROBADO	0.832
17	1:?	1:APROBADO	0.559
18	1:?	1:APROBADO	0.792
19	1:?	2:DESAPROBADO	0.768
20	1:?	2:DESAPROBADO	0.602
21	1:?	1:APROBADO	0.508
22	1:?	2:DESAPROBADO	0.571
23	1:?	1:APROBADO	0.59
24	1:?	1:APROBADO	0.587
25	1:?	1:APROBADO	0.625
26	1:?	1:APROBADO	0.801
27	1:?	1:APROBADO	0.939
28	1:?	1:APROBADO	0.894
29	1:?	2:DESAPROBADO	0.624
30	1:?	1:APROBADO	0.66
31	1:?	1:APROBADO	0.786
32	1:?	2:DESAPROBADO	0.608
33	1:?	2:DESAPROBADO	0.535
34	1:?	2:DESAPROBADO	0.577
35	1:?	1:APROBADO	0.619
36	1:?	2:DESAPROBADO	0.68
37	1:?	1:APROBADO	0.739
38	1:?	1:APROBADO	0.962
39	1:?	2:DESAPROBADO	0.875
40	1:?	1:APROBADO	0.703
41	1:?	1:APROBADO	0.521
42	1:?	2:DESAPROBADO	0.666
43	1:?	1:APROBADO	0.691
44	1:?	1:APROBADO	0.569
45	1:?	1:APROBADO	0.806
46	1:?	1:APROBADO	0.88
47	1:?	1:APROBADO	0.583
48	1:?	2:DESAPROBADO	0.86
49	1:?	2:DESAPROBADO	0.503

50	1:?	1:APROBADO	0.585
51	1:?	2:DESAPROBADO	0.606
52	1:?	1:APROBADO	0.813
53	1:?	1:APROBADO	0.553
54	1:?	2:DESAPROBADO	0.704
55	1:?	2:DESAPROBADO	0.608
56	1:?	1:APROBADO	0.77
57	1:?	2:DESAPROBADO	0.722
58	1:?	2:DESAPROBADO	0.814
59	1:?	1:APROBADO	0.902
60	1:?	1:APROBADO	0.918
61	1:?	1:APROBADO	0.612
62	1:?	2:DESAPROBADO	0.565
63	1:?	1:APROBADO	0.949
64	1:?	1:APROBADO	0.526
65	1:?	2:DESAPROBADO	0.875
66	1:?	1:APROBADO	0.939
67	1:?	2:DESAPROBADO	0.745
68	1:?	2:DESAPROBADO	0.799
69	1:?	1:APROBADO	0.782
70	1:?	1:APROBADO	0.621
71	1:?	2:DESAPROBADO	0.661
72	1:?	1:APROBADO	0.914
73	1:?	1:APROBADO	0.771
74	1:?	1:APROBADO	0.616
75	1:?	1:APROBADO	0.933
76	1:?	1:APROBADO	0.963
77	1:?	2:DESAPROBADO	0.598
78	1:?	1:APROBADO	0.775
79	1:?	2:DESAPROBADO	0.749
80	1:?	1:APROBADO	0.588
81	1:?	1:APROBADO	0.916
82	1:?	2:DESAPROBADO	0.522
83	1:?	2:DESAPROBADO	0.667
84	1:?	1:APROBADO	0.829
85	1:?	2:DESAPROBADO	0.594
86	1:?	1:APROBADO	0.563
87	1:?	2:DESAPROBADO	0.659
88	1:?	1:APROBADO	0.742
89	1:?	2:DESAPROBADO	0.665
90	1:?	2:DESAPROBADO	0.681
91	1:?	1:APROBADO	0.816
92	1:?	2:DESAPROBADO	0.741
93	1:?	1:APROBADO	0.872
94	1:?	1:APROBADO	0.585
95	1:?	2:DESAPROBADO	0.57
96	1:?	1:APROBADO	0.756
97	1:?	1:APROBADO	0.777
98	1:?	1:APROBADO	0.743
99	1:?	2:DESAPROBADO	0.61

100	1:?	2:DESAPROBADO	0.584
101	1:?	1:APROBADO	0.683
102	1:?	1:APROBADO	0.945
103	1:?	1:APROBADO	0.888
104	1:?	1:APROBADO	0.534
105	1:?	1:APROBADO	0.618
106	1:?	1:APROBADO	0.667
107	1:?	1:APROBADO	0.675
108	1:?	2:DESAPROBADO	0.713
109	1:?	2:DESAPROBADO	0.736
110	1:?	1:APROBADO	0.704
111	1:?	2:DESAPROBADO	0.51
112	1:?	1:APROBADO	0.732
113	1:?	2:DESAPROBADO	0.513
114	1:?	1:APROBADO	0.649
115	1:?	1:APROBADO	0.753
116	1:?	1:APROBADO	0.844
117	1:?	1:APROBADO	0.869
118	1:?	1:APROBADO	0.538
119	1:?	1:APROBADO	0.53
120	1:?	1:APROBADO	0.558
121	1:?	1:APROBADO	0.613
122	1:?	2:DESAPROBADO	0.743
123	1:?	1:APROBADO	0.551
124	1:?	2:DESAPROBADO	0.843
125	1:?	2:DESAPROBADO	0.674
126	1:?	2:DESAPROBADO	0.509
127	1:?	2:DESAPROBADO	0.575
128	1:?	1:APROBADO	0.978
129	1:?	2:DESAPROBADO	0.82
130	1:?	1:APROBADO	0.955
131	1:?	1:APROBADO	0.555
132	1:?	2:DESAPROBADO	0.635
133	1:?	1:APROBADO	0.942
134	1:?	1:APROBADO	0.633
135	1:?	2:DESAPROBADO	0.717
136	1:?	1:APROBADO	0.657
137	1:?	1:APROBADO	0.862
138	1:?	2:DESAPROBADO	0.514
139	1:?	2:DESAPROBADO	0.542
140	1:?	2:DESAPROBADO	0.74
141	1:?	1:APROBADO	0.822
142	1:?	2:DESAPROBADO	0.741
143	1:?	2:DESAPROBADO	0.582
144	1:?	2:DESAPROBADO	0.603
145	1:?	1:APROBADO	0.941
146	1:?	1:APROBADO	0.951
147	1:?	1:APROBADO	0.749
148	1:?	2:DESAPROBADO	0.954
149	1:?	2:DESAPROBADO	0.828

150	1:?	1:APROBADO	0.787
151	1:?	1:APROBADO	0.666
152	1:?	1:APROBADO	0.65
153	1:?	1:APROBADO	0.65
154	1:?	1:APROBADO	0.67
155	1:?	1:APROBADO	0.944
156	1:?	2:DESAPROBADO	0.543
157	1:?	1:APROBADO	0.702
158	1:?	1:APROBADO	0.894
159	1:?	2:DESAPROBADO	0.746
160	1:?	1:APROBADO	0.666
161	1:?	1:APROBADO	0.948
162	1:?	2:DESAPROBADO	0.601
163	1:?	1:APROBADO	0.615
164	1:?	1:APROBADO	0.688
165	1:?	1:APROBADO	0.951
166	1:?	2:DESAPROBADO	0.526
167	1:?	2:DESAPROBADO	0.544
168	1:?	2:DESAPROBADO	0.511
169	1:?	2:DESAPROBADO	0.98
170	1:?	1:APROBADO	0.947
171	1:?	2:DESAPROBADO	0.688
172	1:?	1:APROBADO	0.911
173	1:?	1:APROBADO	0.717
174	1:?	1:APROBADO	0.719
175	1:?	1:APROBADO	0.954
176	1:?	1:APROBADO	0.732
177	1:?	2:DESAPROBADO	0.873
178	1:?	2:DESAPROBADO	0.596
179	1:?	2:DESAPROBADO	0.56
180	1:?	1:APROBADO	0.509
181	1:?	2:DESAPROBADO	0.632
182	1:?	2:DESAPROBADO	0.612
183	1:?	1:APROBADO	0.613
184	1:?	1:APROBADO	0.564
185	1:?	1:APROBADO	0.953
186	1:?	1:APROBADO	0.578
187	1:?	1:APROBADO	0.944
188	1:?	2:DESAPROBADO	0.522
189	1:?	2:DESAPROBADO	0.546
190	1:?	1:APROBADO	0.917
191	1:?	2:DESAPROBADO	0.747
192	1:?	1:APROBADO	0.941
193	1:?	1:APROBADO	0.867
194	1:?	2:DESAPROBADO	0.645
195	1:?	1:APROBADO	0.929
196	1:?	1:APROBADO	0.743
197	1:?	1:APROBADO	0.599
198	1:?	1:APROBADO	0.655
199	1:?	2:DESAPROBADO	0.61

NOTA BIOGRÁFICA

Santos Victor Ponce Guizabalo nacido en el distrito de Urpay, Provincia de Pataz y departamento de La libertad el 12 de abril de 1984 hijo de Sr. Wilfredo Ponce Lozano y la Sra. Juana Guizabalo Piundo. Sus estudios de primer a tercero de primaria los realizo en la escuela Comunitaria Tupac Amaru de Tocache. El cuarto de primaria en la I.E 0414 de Tocache, su primaria lo culmino en el caserío San Antonio del Distrito de Pólvora de la Provincia de Tocache, sus estudios secundarios los culmino en la I.E 0412 Alejandrina Morales Amasifuén de la Provincia de Tocache. Sus estudios Universitarios los realizo en la Universidad Nacional Agraria de la Selva (UNAS) en la ciudad de Tingo María en la carrea de ingeniería en informática y Sistemas, se desempeño como docente en la filial de centro preuniversitario UNAS en la provincia de Chanchamayo, luego se desempeño como docente en la I.E particular El Samaritano en el distrito de Aucayacu, trabajo como docente en el centro preuniversitario en la UNAS en la ciudad de Tingo María, fundador, coordinador y profesor en la academia preuniversitaria Poncelet en la Ciudad de Tingo María donde actualmente labora.

"Año de la Unidad, la Paz y el Desarrollo"
UNIVERSIDAD NACIONAL HERMILIO VALDIZÁN
HUÁNUCO - PERÚ
LICENCIADA CON RESOLUCIÓN DEL CONSEJO DIRECTIVO N° 099-2019-SUNEDU/CD
ESCUELA DE POSGRADO

ACTA DE DEFENSA DE TESIS PARA OPTAR EL GRADO DE MAESTRO

En la Plataforma Microsoft Teams de la Escuela de Posgrado, siendo las 19:00 h, del día miércoles 06 DE DICIEMBRE 2023 ante los Jurados de Tesis constituido por los siguientes docentes:

Dr. Edwin Roger ESTEBAN RIVERA	Presidente
Dr. Manuel MARIN MOZOMBITE	Secretario
Dr. Orlando ASCAYO LEON	Vocal

Asesor (a) de tesis: Mg. Alexander Frank PASQUEL CAJAS (Resolución N° 03763-2021-UNHEVAL/EPG-D)

El aspirante al Grado de Maestro en Ingeniería de Sistemas, mención en Tecnología de Información y Comunicación, Don Santos Victor PONCE GUIZABALO.

Procedió al acto de Defensa:

Con la exposición de la Tesis titulado: "RENDIMIENTO ACADÉMICO MEDIANTE TÉCNICAS DE MINERÍA DE DATOS EN ESTUDIANTES DE LA UNIVERSIDAD NACIONAL AGRARIA DE LA SELVA".

Respondiendo las preguntas formuladas por los miembros del Jurado.

Concluido el acto de defensa, cada miembro del Jurado procedió a la evaluación del aspirante al Grado de Maestro, teniendo presente los criterios siguientes:

a) Presentación personal.
b) Exposición: el problema a resolver, hipótesis, objetivos, resultados, conclusiones, los aportes, contribución a la ciencia y/o solución a un problema social y recomendaciones.
c) Grado de convicción y sustento bibliográfico utilizados para las respuestas a las interrogantes del Jurado.
d) Dicción y dominio de escenario.

Así mismo, el Jurado plantea a la tesis **las observaciones** siguientes:

.....

Obteniendo en consecuencia el Maestrista la Nota de Quince (15)
Equivalente a Bueno, por lo que se declara Aprobado

(Aprobado o desaprobado)

Los miembros del Jurado firman el presente ACTA en señal de conformidad, en Huánuco, siendo las 20:20 horas del día miércoles 06 DE DICIEMBRE 2023.

PRESIDENTE
DNI N°.....

SECRETARIO
DNI N°.....

VOCAL
DNI N°.....

Leyenda:
19 a 20: Excelente
17 a 18: Muy Bueno
14 a 16: Bueno

(Resolución N° 01481-2023-UNHEVAL/EPG)

UNIVERSIDAD NACIONAL HERMILIO VALDIZÁN

ESCUELA DE POSGRADO

CONSTANCIA DE ORIGINALIDAD N° 049-2023-SOFTWARE ANTIPLAGIO TURNITIN-UNHEVAL-EPG

La que suscribe, emite la presente constancia de Antiplagio, aplicando el software TURNITIN, la cual reporta un ***20%*** *de originalidad, correspondiente a* ***Santos Victor PONCE GUIZABALO****, de la Maestría en Ingeniería de Sistemas, mención en Tecnología de Información y Comunicación, de la tesis titulada:* ***RENDIMIENTO ACADÉMICO MEDIANTE TÉCNICAS DE MÍNERIA DE DATOS EN ESTUDIANTES DE LA UNIVERSIDAD NACIONAL AGRARIA DE LA SELVA,*** *considerado como asesor al Mg. Alexander Frank PASQUEL CAJAS.*

DECLARANDO (APTO)

Se expide la presente, para los trámites pertinentes.

Pillco Marca, 14 de noviembre de 2023.

UNIVERSIDAD NACIONAL "HERMILIO VALDIZÁN" ESCUELA DE POSGRADO DIRECTOR HUÁNUCO

Dra. Digna Amabilia Manrique de Lara Suarez
DIRECTORA DE LA ESCUELA DE POSGRADO
UNHEVAL

Reporte de similitud

NOMBRE DEL TRABAJO

RENDIMIENTO ACADÉMICO MEDIANTE TÉCNICAS DE MÍNERIA DE DATOS EN ES TUDIANTES DE LA UNIVERSIDAD NACIO NAL AGRARIA DE LA SELVA

AUTOR

SANTOS VICTOR PONCE GUIZABALO

RECUENTO DE PALABRAS

21534 Words

RECUENTO DE CARACTERES

116312 Characters

RECUENTO DE PÁGINAS

95 Pages

TAMAÑO DEL ARCHIVO

4.9MB

FECHA DE ENTREGA

Nov 14, 2023 12:16 PM GMT-5

FECHA DEL INFORME

Nov 14, 2023 12:18 PM GMT-5

20% de similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para cada base

- 20% Base de datos de Internet
- 2% Base de datos de publicaciones
- Base de datos de Crossref
- Base de datos de contenido publicado de Cros
- 11% Base de datos de trabajos entregados

Excluir del Reporte de Similitud

- Material bibliográfico
- Material citado
- Coincidencia baja (menos de 15 palabras)

AUTORIZACIÓN DE PUBLICACIÓN DIGITAL Y DECLARACIÓN JURADA DEL TRABAJO DE INVESTIGACIÓN, TESIS, TRABAJO DE SUFIENCIA PROFESIONAL O TRABAJO ACADÉMICO PARA OPTAR UN GRADO O TÍTULO PROFESIONAL

1. Autorización de Publicación: *(Marque con una "X" según corresponda)*

Bachiller		Título Profesional		Segunda Especialidad		Maestro	X	Doctor	

Ingrese los datos según corresponda.

Facultad/Escuela	
Escuela/Carrera Profesional	
Programa	INGENIERÍA DE SISTEMAS, MENCIÓN EN TECNOLOGÍA DE INFORMACIÓN Y COMUNICACIÓN
Grado que otorga	MAESTRO EN INGENIERÍA DE SISTEMAS, MENCIÓN EN TECNOLOGÍA DE INFORMACIÓN Y COMUNICACIÓN
Título que otorga	

2. Datos del (los) Autor(es): *(Ingrese los datos según corresponda)*

Apellidos y Nombres:	PONCE GUIZABALO SANTOS VICTOR							
Tipo de Documento:	**DNI**	X	**Pasaporte**		**C.E.**		**N° de Documento:**	42915935
Correo Electrónico:	svponce12@gmail.com							
Apellidos y Nombres:								
Tipo de Documento:	**DNI**		**Pasaporte**		**C.E.**		**N° de documento:**	
Correo Electrónico:								
Apellidos y Nombres:								
Tipo de Documento:	**DNI**		**Pasaporte**		**C.E.**		**N° de Documento:**	
Correo Electrónico:								

3. Datos del Asesor: *(Ingrese los datos según corresponda)*

Apellidos y Nombres:	PASQUEL CAJAS ALEXANDER FRANK							
Tipo de Documento:	**DNI**	X	**Pasaporte**		**C.E.**		**N° de Documento:**	46084104
ORCID ID:	0000-0002-0603-0329							

4. Datos de los Jurados: *(Ingrese los datos según corresponda, primero apellidos luego nombres)*

Presidente	ESTEBAN RIVERA EDWIN ROGER
Secretario	MARIN MOZOMBITE MANUEL
Vocal	ASCAYO LEON ORLANDO
Vocal	
Vocal	
Accesitario	

5. Datos del Documento Digital a Publicar: *(Ingrese los datos y marque con una "X" según corresponda)*

Ingrese solo el año en el que sustentó su Trabajo de Investigación: *(Verifique la información en el Acta de Sustentación)*							**2023**		
Modalidad de obtención del Grado Académico o Título Profesional: *(Marque con X según corresponda)*	**Trabajo de Investigación**		**Tesis**	x	**Trabajo Académico**		**Trabajo de Suficiencia Profesional**		
Palabras claves	**MINERÍA DE DATOS**			**RENDIMIENTO ACADÉMICO**		**REGRESIÓN LOGÍSTICA**			
Tipo de acceso: *(Marque con X según corresponda)*	**Abierto**	X	**Cerrado**		**Restringido***		**Periodo de Embargo**		
(*) Sustentar razón:									

UNHEVAL UNIVERSIDAD NACIONAL HERMILIO VALDIZÁN | VICERRECTORADO DE INVESTIGACIÓN | DIRECCIÓN DE INVESTIGACIÓN | DI UNHEVAL

6. Declaración Jurada: (Ingrese todos los datos requeridos completos)

Soy Autor (a) (es) del Trabajo de Investigación Titulado: *(Ingrese el título tal y como está registrado en el **Acta de Sustentación**)*
RENDIMIENTO ACADÉMICO MEDIANTE TÉCNICAS DE MINERÍA DE DATOS EN ESTUDIANTES DE LA UNIVERSIDAD NACIONAL AGRARIA DE LA SELVA
Mediante la presente asumo frente a la Universidad Nacional Hermilio Valdizán (en adelante LA UNIVERSIDAD), cualquier responsabilidad que pueda derivarse por la autoría, originalidad y veracidad del contenido del trabajo de investigación, así como por los derechos de la obra y/o invención presentada. En consecuencia, me hago responsable frente a LA UNIVERSIDAD y frente a terceros de cualquier daño que pudiera ocasionar a LA UNIVERSIDAD o a terceros, por el incumplimiento de lo declarado o que pudiera encontrar causas en los trabajos de investigación presentado, asumiendo toda la carga pecuniaria que pudiera derivarse de ello. Asimismo, por la presente me comprometo a asumir además todas las cargas pecuniarias que pudiera derivar para LA UNIVERSIDAD en favor de terceros con motivos de acciones, reclamaciones o conflictos derivados del incumplimiento de lo declarado o las que encontraren causa en el contenido del Trabajo de Investigación. De identificarse fraude, piratería, plagio, falsificación o que el trabajo haya sido publicado anteriormente; asumo las consecuencias y sanciones que de mis acciones se deriven, sometiéndome a las acciones legales y administrativas vigentes.

7. Autorización de Publicación Digital:

A través de la presente autorizo de manera gratuita a la Universidad Nacional Hermilio Valdizan a publicar la versión digital de este trabajo de investigación en su biblioteca virtual, repositorio institucional y base de datos, por plazo indefinido, consintiendo que con dicha autorización cualquier tercero podrá acceder a dichas paginas de manera gratuita pudiendo revisarla, imprimirla o grabarla siempre y cuando se respete la autoría y sea citada correctamente.			
Apellidos y Nombres	PONCE GUIZABALO SANTOS VICTOR	**Firma**	
Apellidos y Nombres		**Firma**	
Apellidos y Nombres		**Firma**	

FECHA: Huánuco, 03 de Junio del 2024

Nota:

- ✓ No modificar los textos preestablecidos, conservar la estructura del documento.
- ✓ Marque con una X en el recuadro que corresponde.
- ✓ Llenar este formato de forma digital, con tipo de letra calibri, tamaño de fuente 09, manteniendo la alineación del texto que observa en el modelo, sin errores gramaticales (recuerde las mayúsculas también se tildan si corresponde).
- ✓ La información que escriba en este formato debe coincidir con la información registrada en los demás archivos y/o formatos que presente, tales como: DNI, Acta de Sustentación, Trabajo de Investigación (PDF), Constancia de Similitud, Reporte de Similitud.
- ✓ Cada uno de los datos requeridos en este formato, es de carácter obligatorio según corresponda.
- ✓ Se debe de imprimir, firmar y luego escanear el documento (legible).