UNIVERSIDAD NACIONAL DE SAN ANTONIO ABAD DEL CUSCO

ESCUELA DE POSGRADO

MAESTRÍA EN CIENCIAS MENCIÓN INFORMÁTICA

TESIS

"PREDICCIÓN DEL RENDIMIENTO ACADÉMICO DE LOS ESTUDIANTES DE LA UNSAAC A PARTIR DE SUS DATOS DE INGRESO UTILIZANDO ALGORITMOS DE APRENDIZAJE AUTOMÁTICO"

Presentada por el Bachiller en Ingeniería Informática y de Sistemas

BR. DENNIS IVÁN CANDIA OVIEDO

PARA OPTAR AL GRADO ACADÉMICO DE:

MAESTRO EN INFORMÁTICA

ASESOR:
DR. LAURO ENCISO RODAS

Tesis auspiciada por el convenio ARES-UNSAAC

Cusco – Perú

2019

Dedicatoria

Dedico este trabajo de tesis a mi familia.

A mis hijas Dayana y Mayerly, motor y motivo para alcanzar mis metas, todo lo que hago es por ustedes y disculpen por haberles quitado vuestro tiempo.

A mi esposa Ana Carmen, tu afecto, cariño y comprensión, nos permitirá llegar muy lejos.

Las Amo.

Agradecimientos

Agradezco a la Universidad Nacional de San Antonio Abad del Cusco, por permitirme crecer académicamente, ser profesional, darme la oportunidad de desempeñarme laboralmente y ser útil a la sociedad.

Gracias UNSAAC iii.

RESUMEN

El presente trabajo de tesis tiene como propósito fundamental predecir el rendimiento académico de los estudiantes de la Universidad Nacional de San Antonio Abad del Cusco (UNSAAC) en el primer semestre a partir de sus datos del proceso de ingreso o de admisión a la institución, considerando que sería muy importante para una institución de formación universitaria saber de manera anticipada el posible rendimiento académico de sus estudiantes, es decir el éxito o fracaso en su primer semestre en la universidad, el cual redundará en los semestres posteriores, para de esta forma plantear estrategias que le permitan no solamente a la institución, sino también a los docentes y al mismo estudiante mejorar sus actividades del proceso enseñanza aprendizaje.

Para lograr el propósito de la predicción se utiliza aprendizaje automático (*machine learning*), que es una rama de la inteligencia artificial, la metodología utilizada para generar los modelos predictivos es CRISP-DM, que es una metodología bastante utilizada para este tipo de proyectos, y WEKA como plataforma de software para el despliegue y comparación del desempeño de los diferentes algoritmos de aprendizaje automático supervisado, el mismo que es software libre.

Los algoritmos supervisados de clasificación utilizados en la presente tesis son: Arboles de decisión J48, Random Forest, Vecinos más cercanos (KNN), Función de Regresión Logística y Perceptrón multicapa, de estos 5 algoritmos el que tuvo la mejor performance, es decir el que tuvo la mejor predicción fue el algoritmo Random Forest también conocido como Bosques Aleatorios, logrando predecir hasta un 69.35%, el segundo mejor predictor fue la función de Regresión Logística con un 68.33%, es importante mencionar que los factores que influyen en el rendimiento académico son la

nota de ingreso a la institución, la Escuela Profesional donde estudia, el semestre, la modalidad de ingreso y la cantidad de cursos matriculados.

PALABRAS CLAVES

Rendimiento académico, aprendizaje automático, metodología CRISP-DM, WEKA, algoritmos de clasificación.

RESUMO

O objetivo principal desta tese é predizer o desempenho acadêmico dos estudantes da Universidade Nacional de San Antonio Abad de Cusco (UNSAAC) no primeiro semestre a partir de seus dados do processo de admissão ou admissão à instituição, considerando que sería muito importante para uma instituição universitária conhecer antecipadamente o possível desempenho acadêmico de seus alunos, ou seja, o sucesso ou o fracasso no primeiro semestre da universidade, o que resultará em semestres subseqüentes, a fim de propor estratégias. que permitem não somente à instituição, señao também aos professores e ao mesmo aluno melhorar suas atividades de ensino-aprendizagem.

Para alcançar o objetivo da previsão, é usada o aprendizagem automática (*machine learning*), que é um ramo da inteligência artificial. A metodologia utilizada para gerar modelos preditivos é CRISP-DM, que é uma metodologia amplamente utilizada para este tipo de projeto. e WEKA como uma plataforma de software para a implantação e comparação do desempenho dos diferentes algoritmos de aprendizado de máquina supervisionados, o mesmo que é software livre.

Os algoritmos supervisionados de classificação utilizados nesta tese são: arboles de decisión J48, Random Forest, vecinos más cercanos (KNN), Función de Regresión Logística e Perceptron Multicapa, destes 5 algoritmos o que teve o melhor desempenho, ou seja, o que teve a melhor previsão foi o algoritmo Random Forest também conhecida como Bosques Aleatorios, prevendo até 69,35%, o segundo melhor preditor foi Función de Regresión Logística com 68,33%, é importante mencionar que os fatores que

influenciam o desempenho acadêmico são a nota de entrada para a instituição. , a escola profissional onde estuda, o semestre, o tipo de internação e o número de cursos matriculados.

PALAVRAS CHAVES

Desempenho acadêmico, aprendizado de máquina, metodología CRISP-DM, WEKA, algoritmos de classificação.

ÍNDICE GENERAL

Índice de cuadros

Índice de Gráficos

Índice de Figuras

INTRODUCCIÓN

Las organizaciones en la actualidad están generando inmensas cantidades de información que muchas veces no son procesadas adecuadamente, es decir las tareas de recolección, extracción, almacenamiento y análisis de la información no está siendo realizada y utilizada adecuadamente, considerando que existen tecnologías que nos permiten realizar estas tareas. La Universidad Nacional de San Antonio Abad del Cusco (UNSAAC), genera información diaria respecto de todas sus actividades tanto académicas como administrativas, y es precisamente en la parte académica, donde se realiza el presente trabajo de tesis, para predecir el rendimiento académico de los estudiantes en su primer semestre a partir de la información generada en el proceso de admisión o ingreso a la Universidad.

Respecto del rendimiento académico y según muchos autores es un término bastante complejo de definir, en vista que tiene muchas aristas que nos permitan medirla, sin embargo la mayoría coincide en que el rendimiento académico está asociado a las calificaciones que el estudiante obtiene en las diferentes evaluaciones que este rinde, el presente trabajo busca predecir el rendimiento académico de los estudiantes en el primer semestre, considerando la información de los ingresantes y la consignada en el momento de la postulación, para determinar a través de una predicción con datos históricos desde el semestre 2014-I, hasta el 2018-1, si un estudiante ingresante tendrá éxito o fracasará.

Sería importante determinar de manera anticipada sí los estudiantes tendrán éxito o fracaso en el primer semestre de su vida académica universitaria, lo cual permitiría a los encargados del proceso enseñanza aprendizaje u otros encargados, tomar las previsiones necesarias para que los estudiantes puedan, si es el caso, mejorar su rendimiento

académico a través de tutorías personalizadas, de apoyo psicológico, incentivos o charlas, este aspecto sería beneficioso no solo para la Universidad sino también para el estudiante y los docentes.

En este sentido el capítulo I, describe el planteamiento del problema, cuyo problema general está planteado de la siguiente manera: "Será posible predecir el rendimiento académico de los estudiantes de la UNSAAC en el primer semestre a partir de sus datos de ingreso utilizando algoritmos de aprendizaje automático"

El capítulo II, describe el marco teórico, dividido en las bases teóricas y marco conceptual respecto del rendimiento académico, aprendizaje automático, metodología CRISP-DM, WEKA, y los algoritmos supervisados de *machine learning*.

El capítulo III, contiene la hipótesis y variables, teniendo como variable dependiente al "rendimiento académico" y como variables independientes a los "factores socio demográficos" y "factores socio educativos", los cuales fueron obtenidos a partir de la encuesta web a los postulantes a la UNSAAC y de la información proporcionada por la Unidad de Centro de Computo.

El capítulo IV, describe la metodología de la tesis, el cual es una investigación correlacional, la población de estudio y la muestra está en base a los estudiantes ingresantes a la UNSAAC, desde el semestre 2014-I hasta el semestre 2018-I.

El capítulo V, describe los resultados y discusiones, en el mismo que utilizamos la metodología CRISP-DM, en cada una de sus 6 fases para comprender el contexto del trabajo, la comprensión y preparación de los datos, la generación de modelos predictivos, comparación y selección del algoritmo con mayor desempeño de predicción, evaluación y despliegue, para lograr el objetivo de la predicción.

Finalmente se describen las conclusiones, recomendaciones, bibliografía y anexos.

Capítulo 1

PLANTEAMIENTO DEL PROBLEMA

1.1. Situación Problemática

Las Universidades son agentes de cambio y de transformación de personas y de sus capacidades para poder insertarlos al mercado laboral competitivo y a nuestra sociedad, sin embargo en la actualidad las instituciones de educación superior cuentan con escasas políticas de monitoreo del rendimiento académico de sus estudiantes durante su estancia en la institución y existe desconocimiento de los niveles de deserción estudiantil, de sus causas y/o factores, de la misma manera existen muy pocas investigaciones que intente predecir el éxito o el fracaso en la vida académica del estudiante a partir de sus datos de ingreso o de admisión a la Universidad. El rendimiento académico de los estudiantes es bastante complejo definirlo, pero en concreto es el proceso de aprendizaje de los estudiantes considerando una infinidad de factores, la evaluación de dicho aprendizaje generalmente es medido como aprobado o desaprobado.

El problema objeto de estudio de la presente investigación nace a partir de la identificación del perfil académico de los estudiantes ingresantes, el cual es desconocido para la Universidad, para después buscar patrones y/o factores que puedan ser analizados a partir de la información proporcionada por los datos de ingreso de los postulantes en cada uno de los exámenes de admisión desde el semestre 2014-I hasta el 2018-I, para tal fin se utilizó algoritmos de aprendizaje automático específicamente las técnicas de clasificación para plantear un modelo predictivo que permita determinar el rendimiento académico de los estudiantes en

el primer semestre y, también determinar el mejor algoritmo predictor, esta información generada podría ser utilizada por los encargados del proceso enseñanza aprendizaje de la UNSAAC para tomar las previsiones que consideren necesarias, tanto a nivel de toda la Universidad como también por Escuela Profesional.

El presente trabajo determina también la relación entre los factores socio demográficos y socioeducativos (datos proporcionados por los postulantes a la UNSAAC), y el rendimiento académico, para intentar dar respuestas a las preguntas ¿Por qué del bajo rendimiento?, ¿Por qué desertan de la universidad?, del mismo modo ¿Por qué son estudiantes de alto rendimiento académico?, la deserción estudiantil es también un problema bastante complejo que afectan no solamente a las instituciones universitarias, sino también es un problema que tiene repercusiones en el estudiante, en la familia y por ende en la sociedad, además genera también problemas económicos al estado.

Tomando en consideración las limitaciones de los recursos económicos estatales para la educación superior pública, son imprescindibles investigaciones en el campo del rendimiento académico, considerando además que el gasto público anual en el Perú por alumno universitario, según un boletín de El Comercio (2016) es de 1100 dólares americanos, este es un problema que las autoridades universitarias deberían evaluar y buscar políticas y estrategias para que sus estudiantes puedan egresar en los semestres establecidos por las Escuelas Profesionales y así evitar gastos innecesarios en la formación de sus estudiantes, este aspecto se lograría si se identificaría de manera anticipada el posible rendimiento académico de los estudiantes como se pretende en este trabajo de investigación.

Las universidades de nuestra región y en especial la UNSAAC vienen generando políticas para poder garantizar la calidad en la formación de sus estudiantes, buscando la acreditación de sus escuelas profesionales, esta tesis podría proporcionar a los encargados del proceso enseñanza-aprendizaje de la UNSAAC conocer de manera anticipada el posible rendimiento de los ingresantes para de esta manera reducir probabilidades de bajo rendimiento, deserción y/o abandono, o en el otro escenario (estudiantes con posible buen rendimiento), generar políticas de incentivos, motivación u otros similares.

Respecto al rendimiento o desempeño académico de los estudiantes se ha podido investigar que existen diversas publicaciones, artículos y libros, que describen las causas del mismo, algunos por ejemplo describen hasta 4 factores (Factores relacionados a los estudiantes, la institución, los docentes y a la familia), en este trabajo de tesis analizaremos los factores relacionados al rendimiento académico de los estudiantes en el primer semestre es decir de los estudiantes ingresantes a partir de la información generada en el proceso de admisión o de ingreso a la UNSAAC, obtenidos a partir de la encuesta a los postulantes y de la información solicitada a la Unidad de Centro de Computo.

En resumen, se pretende desarrollar un modelo predictivo del rendimiento académico de los estudiantes de la UNSAAC del primer semestre (ingresantes) utilizando algoritmos de aprendizaje automático a partir de la información de sus datos de ingreso a la UNSAAC, así mismo, conocer y evaluar el desempeño de los algoritmos de aprendizaje automático (*machine learning*) para determinar el algoritmo más óptimo o con el mayor grado de predicción, se espera contribuir a intentar dar una respuesta sobre el rendimiento académico expresado como: Bueno o Malo de los estudiantes de la UNSAAC en su primer semestre, la información

para realizar dicho modelo de predicción, ha sido solicitado a la Unidad del Centro de Computo de la Institución, la cual ha sido utilizada después de una análisis, tratamiento y limpieza de los datos.

1.2. Formulación del Problema

1.2.1. Problema general

¿Es posible predecir el rendimiento académico de los estudiantes de la UNSAAC en el primer semestre a partir de sus datos de ingreso utilizando algoritmos de aprendizaje automático?

1.2.2. Problemas específicos

1.- ¿Cuáles son los factores claves de los datos de ingreso que determinan el rendimiento académico de los estudiantes de la UNSAAC en su primer semestre?

2.- ¿Cuál es el algoritmo de aprendizaje automático más eficiente que predice el rendimiento académico de los estudiantes de la UNSAAC a partir de sus datos de ingreso en el primer semestre?

1.3. Justificación de la Investigación

Siendo políticas de las instituciones de carácter universitario la calidad de enseñanza, la presente investigación tiene relevancia social y pedagógica considerando que la predicción del rendimiento académico es un factor que influye directamente en la mejora de la calidad de la educación superior para universidades públicas y privadas en base a la información generada en el proceso de ingreso o de admisión a las universidades. Para la UNSAAC poder predecir el rendimiento académico de sus ingresantes a partir de los datos generados en el proceso de ingreso, le servirá para mejorar la toma de decisiones de acuerdo a la predicción de la presente tesis, para de esta manera generar políticas de mejora académica como tutorías personalizadas, apoyo psicológico, incentivos, becas, recomendaciones u otros.

La relevancia práctica se describe desde el punto de vista de los actores más importantes del proceso de enseñanza aprendizaje es decir docentes y estudiantes, para un profesor que regenta una determinada asignatura, conocer el posible rendimiento académico de sus estudiantes le permitirá poder aplicar y/o mejorar sus técnicas de enseñanza aprendizaje, buscando alcanzar de la mejor manera los objetivos educacionales de su escuela profesional. Para los estudiantes saber su rendimiento académico antes de iniciar sus estudios Universitarios, le permitirá conocer y tener conciencia de sus fortalezas y debilidades con las que enfrenta su vida académica dentro de la Universidad.

La presente tesis, tiene también implicancia profesional y personal, puesto que para el investigador existe un constante interés de aprender temas nuevos de la profesión y de la docencia universitaria y que sobre todo tengan aplicación a nuestro entorno para resolver problemas de situaciones reales con la intención de generar mejoras.

Finalmente se justifica esta investigación porque apoyaría de manera sustancial a los indicadores de acreditación y licenciamiento respecto de la utilización de las TIC’s en las escuelas profesionales, en los aspectos de las tutorías y seguimiento académico de los estudiantes.

1.4. Objetivos de la Investigación

1.4.1. Objetivo General

Predecir el rendimiento académico de los estudiantes de la UNSAAC del primer semestre a partir de sus datos de ingreso utilizando algoritmos de aprendizaje automático.

1.4.2. Objetivos Específicos

1. Determinar los factores claves de los datos de ingreso a la UNSAAC, que permiten la predicción del rendimiento académico de los estudiantes en el primer semestre.

2. Determinar el algoritmo de aprendizaje automático más eficiente, para predecir el rendimiento académico de los estudiantes de la UNSAAC del primer semestre a partir de sus datos de ingreso a la Universidad.

Capítulo 2

MARCO TEÓRICO CONCEPTUAL

2.1. Bases Teóricas

2.1.1. Rendimiento Académico

Blanco, L (1989), el tema de rendimiento académico o el fracaso escolar universitario hay que estudiarlo dentro de un contexto sociocultural-económico-político a la vez que familiar, personal y académico, ya que los factores que influyen en el mismo son numerosos y se encuentran muy interrelacionados. Exigirá, por tanto, un tratamiento interdisciplinar con una metodología común y análisis multivariados.

Frente a esta complejidad de posibles factores que determinan de alguna manera, el rendimiento de los alumnos universitarios, los estudios realizados no profundizan, en general, en las verdaderas causas de fracaso universitario, sino que más bien analizan de forma aislada algunos de estos factores.

Son estudios realizados desde una perspectiva fundamentalmente sociológica; los datos se analizan de un modo sencillo y con propósitos puramente descriptivos. No se abordan aspectos como las relaciones entre distintos sectores que forman la Universidad.

Blanco, L (1989), describía que los estudios realizados sobre el rendimiento académico son fundamentalmente sociológicos, hemos podido investigar que en los últimos años ya existen investigaciones y trabajos de tesis que no son puramente descriptivos sino que más bien y con el avance de la tecnología y el auge de la Inteligencia artificial, minería de datos, algoritmos de predicción, entre otros, se

empezaron a desarrollar investigaciones que relacionan los factores y las causas del éxito o fracaso de los estudiantes en su vida académica, y hasta se intenta predecir posibles resultado. Sin embargo, el autor describe algo muy importante y que hasta la fecha no se ha podido encontrar investigaciones que describan o relacionen el rendimiento académico de los estudiantes en la universidad con el puesto de trabajo que ocupará fuera de ella, es decir relacionar el rendimiento académico de un estudiante con el éxito o fracaso en la vida ocupacional del individuo.

Según Solano, L. (2015), El estudio del rendimiento escolar constituye, hoy en día, un tema destacado en la investigación educativa. Y es que, para nuestra sociedad actual, caracterizada por el bombardeo continuado de información desde distintos medios, el gran desafío de la educación es transformar esa gran cantidad de información en conocimiento personal valido para poder desenvolverse con eficacia en la vida. De ahí que lograr el éxito o el fracaso en los estudios es de vital importancia para el futuro profesional individual.

No obstante, aunque binomio éxito-fracaso hace referencia a una normativa general sin tener en cuenta, a veces, el proceso evolutivo y las diferentes individualidades de cada alumno, lo cierto es que un buen rendimiento académico conllevara al éxito escolar mientras que un rendimiento escolar deficiente al fracaso.

Díaz, M. & Apodaca, P. & Arias, J. & Escudero, T. & Rodríguez, S. & Vidal, J. (2002) el rendimiento académico se divide en 2 partes desde el punto de vista práctico, la tendencia más habitual es identificar rendimiento con resultados, distinguiendo entre estas 2 categorías: inmediatos y diferidos. En el caso de la enseñanza superior los primeros estarían determinados por las calificaciones que obtienen los alumnos a lo largo de los estudios hasta obtener la titulación

correspondiente. Los segundos hacen referencia al impacto que las formaciones recibidas por los titulados tienen en la vida social; es decir la utilidad que dichos estudios tienen en el proceso de incorporación al mundo laboral de los graduados universitarios. Ambos criterios también denominados rendimiento interno y externo constituyen los parámetros de referencia que se emplean con mayor frecuencia para evaluar el rendimiento académico de la enseñanza superior.

El documento refiere a que el rendimiento académico se podría medir a partir de las notas obtenidas por los estudiantes a lo largo de su vida académica universitaria al cual lo llama categoría “inmediata”, y la otra al éxito o fracaso en la incorporación al mundo laboral del individuo, al cual lo denomina “diferidos”

2.1.1.1. Factores a tener en cuenta para determinar el rendimiento académico

Gonzales, R (1989), describe en su investigación tres factores:

1. Factores inherentes al alumno

a) Falta de preparación para acceder a estudios superiores o niveles de conocimientos no adecuados a las exigencias de la Universidad
b) Desarrollo inadecuado de aptitudes especificas acordes con el tipo de carrera profesional elegida.
c) Aspectos de índole actitudinal
d) Falta de métodos de estudio o técnicas de trabajo intelectual
e) Estilos de aprendizaje no acordes con la carrera profesional elegida

2. Factores inherentes con el profesor

a) Deficiencias pedagógicas
b) Falta de tratamiento individualizado
c) Falta de mayor dedicación

3. Factores inherentes a la organización académica universitaria

a) Ausencia de objetivos claramente definidos
b) Falta de coordinación entre distintas materias
c) Sistemas de selección utilizados
d) Criterios objetivos para la evaluación.

En conclusión y respecto del trabajo de Gonzales, R. (1989), el rendimiento académico de los estudiantes depende estrictamente de la universidad donde estudia, de sus docentes y sobre todo de las capacidades de los alumnos, y de este

último es todavía más complejo puesto que existen una serie de factores que permitirán determinar el éxito o fracaso de un alumno, en la presente investigación nos centraremos básicamente en la información relacionada al alumno considerando únicamente sus datos de ingreso o de admisión a la universidad, como se describirá más adelante.

Tejedor (1998) define dos tipos de rendimiento académico: por una parte, el rendimiento en sentido estricto, medido a través de la presentación a exámenes o éxito en las pruebas (calificaciones); por otra el rendimiento en sentido amplio, medido a través del éxito (finalización puntual), el retraso o el abandono en los estudios. También se habla de "regularidad académica" cuando el concepto de rendimiento académico se operacionaliza mediante las tasas de presentación o no a las convocatorias de examen.

No obstante, hay que señalar que el rendimiento académico, en sentido estricto, se analiza en muy pocos estudios a nivel universitario, dándose más importancia en los mismos a otros criterios de medición. Ello parece lógico, si tenemos en cuenta que cuanto más bajos son los niveles de escolarización, menos relevancia tiene el problema de la deserción y más relevancia tiene el tema de las calificaciones escolares para determinar el rendimiento académico de los estudiantes.

Los conceptos o definiciones del rendimiento académico no son fáciles de definir como se ha podido evidenciar en la búsqueda de información para el presente trabajo, sin embargo, la mayoría de autores y trabajos de investigación coinciden en afirmar que el rendimiento académico es multidimensional y está basado en diferentes factores que los estudiantes asumen como retos en su proceso de

enseñanza aprendizaje, asociados generalmente a las calificaciones que estos obtienen.

Considerando toda esta complejidad y polémica que significa para determinar el rendimiento académico de los estudiantes, podemos afirmar que el rendimiento académico es la medida de las capacidades de los alumnos expresadas en el éxito o fracaso en su vida académica o en las calificaciones que estos obtienen luego del proceso de enseñanza-aprendizaje.

2.1.2. Aprendizaje automático

Kelleher, J. & Mac, B. & D'Arcy, A. (2017), El aprendizaje automático es una ciencia que le permite a las maquinas desarrollar técnicas para poder aprender. Este aprendizaje es un proceso automatizado que extrae patrones de los datos para la construcción de modelos que permitan realizar la predicción utilizando algoritmos supervisados, relacionando características descriptivas actuales con características de destino basadas en un conjunto de ejemplos históricos o instancias, consta de 2 pasos, que se muestran en las figuras 1 y 2.

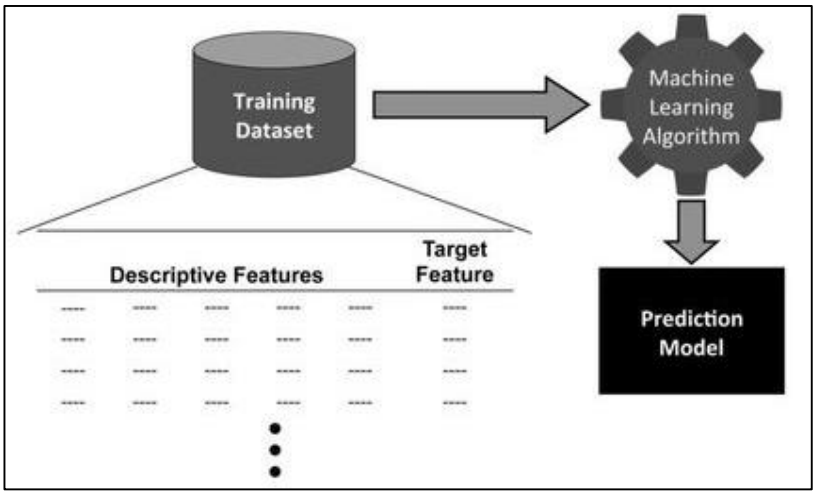

Figura 1.- Aprendiendo un modelo a partir de un conjunto de instancias históricas

Fuente: Kelleher, J. & Mac, B. & D'Arcy, A. (2017)

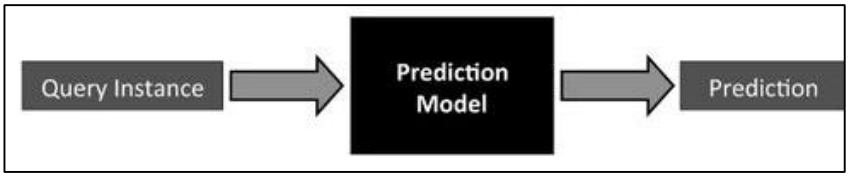

Figura 2.- Usando un modelo para hacer predicciones

Fuente: Kelleher, J. & Mac, B. & D'Arcy, A. (2017)

Mitchell, T. (1997), El aprendizaje automático se basa en ideas de un conjunto diverso de disciplinas, incluyendo la Inteligencia artificial, probabilidad y estadística, complejidad computacional, teoría de la información, psicología y neurobiología, teoría del control y filosofía.

Moreno, et al (1994), el aprendizaje se refiere a un amplio espectro de situaciones en las cuales el aprendiz incrementa su conocimiento o sus habilidades para cumplir una tarea. El aprendizaje aplica inferencias a determinada información para construir una representación apropiada de algún aspecto relevante de la realidad o de algún proceso.

Se dice que un sistema que aprende de forma automatizada (o aprendiz), es un artefacto (o un conjunto de algoritmos) que, para resolver problemas, toma decisiones basadas en la experiencia acumulada. Estos sistemas deben ser capaces de trabajar con un rango muy amplio de tipos de datos de entrada, que pueden incluir datos incompletos, inciertos, ruido, inconsistencias.

La caracterización del proceso de aprendizaje automático es:

Aprendizaje=Selección + Adaptación

Como se puede apreciar los autores identifican 2 etapas para el aprendizaje automático, la primera que hace referencia a la selección de las características más importantes y relevantes de objeto en estudio, comparándolas con otras características conocidas, y cuando estas diferencias son bastante significativas utiliza la segunda etapa que es la adaptación.

Para los autores de la revista Management solutions (2018) *Machine learning* es una pieza clave en la transformación de los modelos de negocios, las técnicas de aprendizaje automático (o *machine Learning*) están experimentando un auge sin precedentes en diversos ámbitos, tanto en el mundo académico como en el empresarial, y constituyen una importante palanca de transformación.

Bajo esta definición el concepto de aprendizaje automático lleva existiendo al menos desde los años 50, periodo en el que se descubrieron y redefinieron diversos métodos estadísticos y se aplicaron al aprendizaje automático a través de algoritmos simple, aunque circunscritos casi exclusivamente al ámbito académico.

Frente a las técnicas estadísticas clásicas, la introducción de técnicas de *machine learning*, permiten mejorar el proceso de estimación de modelos, no solo con relación al aumento del poder predictivo a través de nuevas metodologías y técnicas de selección de variables, sino también en la mejora de la eficiencia de los procesos a través de la automatización.

Russell, S. & Norvig, P. (2004), el computador debería poseer las siguientes capacidades:

- **Procesamiento de lenguaje natural. -** Que le permita comunicarse satisfactoriamente en inglés.

- **Representación del conocimiento**. – P-ara almacenar lo que se conoce o siente.
- **Razonamiento automático**. - Para utilizar la información almacenada para responder a preguntas y extraer nuevas conclusiones
- **Aprendizaje automático**. - Para adaptarse a nuevas circunstancias y para detectar y extrapolar patrones

Para poder decir que un programa dado piensa como un humano, es necesario contar con un mecanismo para determinar cómo piensan los humanos. Es necesario penetrar en el funcionamiento de las mentes humanas. Hay dos formas de hacerlo: mediante introspección (intentando atrapar nuestros propios pensamientos y conforme estos van apareciendo) y mediante experimentos psicológicos.

2.1.2.1. Tipos de aprendizaje en *machine learning*

Zambrano, J. (2018) el *machine learning* desarrolla algoritmos que hacen que las máquinas puedan aprender por su cuenta y responder a determinadas preguntas con bastante certeza. Para desarrollar estos algoritmos, existen dos modalidades: aprendizaje supervisado y no supervisado.

1. Aprendizaje supervisado

La primera modalidad de aprendizaje que tiene el ***machine learning*** es la de aprendizaje supervisado. Usándola, se entrena al algoritmo otorgándole las preguntas, denominadas características, y las respuestas, denominadas etiquetas.

Esto se hace con la finalidad de que el algoritmo las combine y pueda hacer predicciones.

Existen, a su vez, dos tipos de aprendizaje supervisado:

Regresión: Es uno de los algoritmos del aprendizaje supervisado, el cual es utilizado en aprendizaje automático y en la estadística, el cual consiste en dibujar una recta la cual indicará la tendencia de un grupo de datos continuos (números), y si fuesen discretos (cadenas de texto), se utilizaría regresión logística.

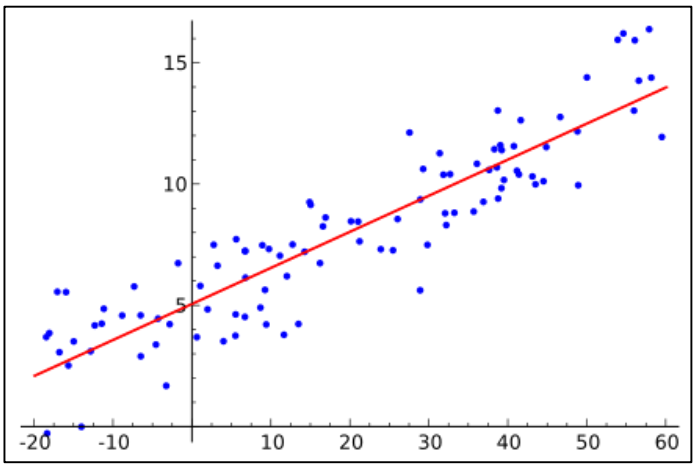

Figura 3.- Aprendizaje automático: Regresión

Fuente: Zambrano, J. (2018). ¿Aprendizaje supervisado o no supervisado?

Clasificación: Los algoritmos de clasificación como su nombre lo indica busca o encuentra patrones que luego le permitirá clasificar a los elementos y determinar a qué grupos o clases pertenecen, se debe mencionar que los valores para estos algoritmos deben ser valores discretos

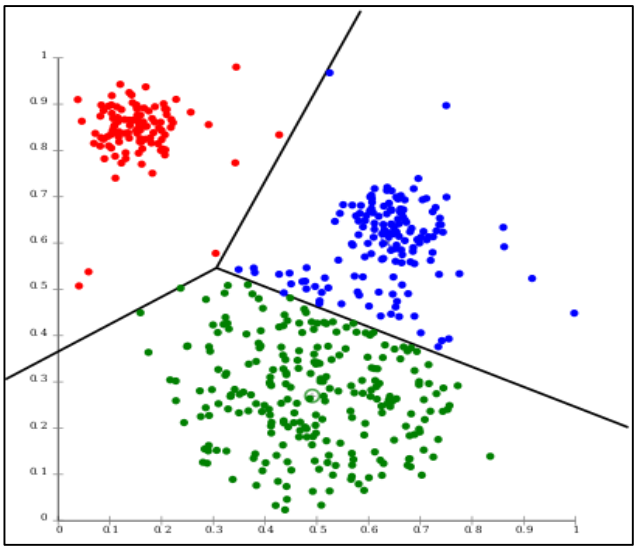

Figura 4.- Aprendizaje automático: Clasificación

Fuente: Zambrano, J. (2018). ¿Aprendizaje supervisado o no supervisado?

El algoritmo no está en capacidad de determinar a qué grupo pertenece un valor o cuál es el resultado de una operación. Solamente logra relacionar características con etiquetas y así obtener un resultado.

2.- Aprendizaje no supervisado

Peláez, N. (2012) El aprendizaje no supervisado consiste en que la red descubra por si misma características, regularidades, correlaciones o categorías en los datos de entrada y se obtengan de forma codificada en la salida. En algunos casos, la salida representa el grado de similitud entre la información que se le está presentado en la entrada y la que se le ha mostrado en el pasado. En otro caso podría realizar un *clustering* o establecimiento de categorías, indicando con la salida de la red a qué categoría pertenece la información presentada como entrada, siendo la propia red quien deba encontrar las categorías apropiadas a partir de correlaciones en las entradas presentadas.

Calvo, D. (2017) describe al Aprendizaje No Supervisado, como el conjunto de técnicas que permiten inferir modelos para extraer conocimiento de conjuntos de datos donde a priori se desconoce.

Para Gonzalo, A (2018) Se llama no supervisado porque, contrariamente al supervisado, tiende a ser más subjetivo ya que no tiene respuestas correctas. Los algoritmos sirven para descubrir y presentar estructuras interesantes en los datos. El objetivo del aprendizaje no supervisado es modelizar la estructura o distribución de los datos para aprender más sobre ellos. Sirve tanto para entender como para resumir un conjunto de datos.

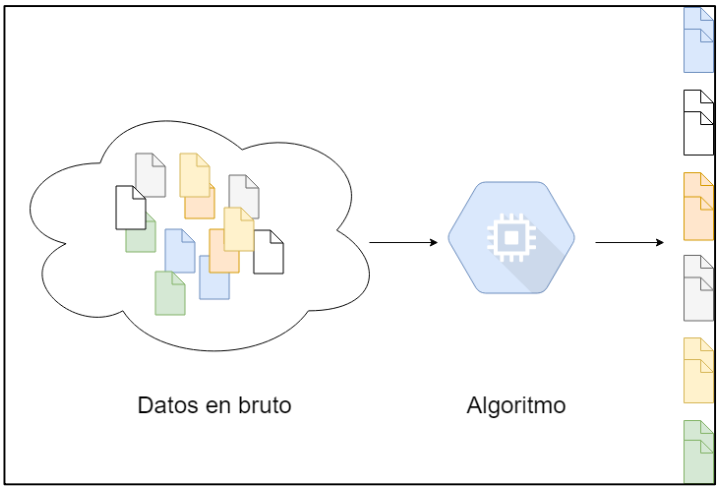

Figura 5.- Esquema de un modelo de aprendizaje no supervisado clustering o agrupamiento

Fuente: recuperado de http://machinelearningparatodos.com/tipos-de-aprendizaje-automatico/

Como lo describe Zambrano (2018) existen 2 técnicas de aprendizaje automático: el aprendizaje supervisado y el no supervisado, los cuales a su vez tienen sus propias técnicas, como se nuestra en la figura:

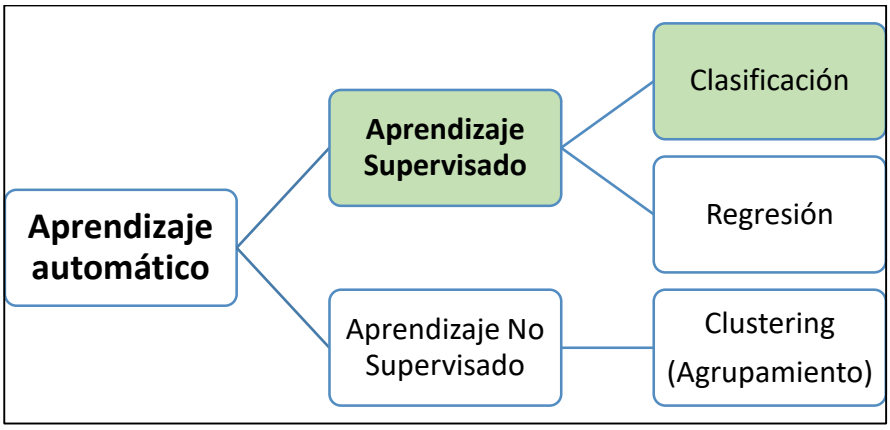

Figura 6.- Técnicas de aprendizaje automático

Fuente: Elaboración propia

2.1.2.2. Técnicas de Aprendizaje automático con algoritmos supervisados

Para el presente trabajo de tesis y los modelos predictivos utilizaremos los algoritmos supervisados, porque conocemos los datos de entrada y las respuestas a los mismos, es decir, conocemos las variables independientes (datos de ingreso), y las respuestas que están expresadas en el éxito o fracaso del rendimiento académico de los estudiantes (Bueno o Malo) que vendría a ser la variable dependiente

El aprendizaje supervisado como ya se mencionó incluye dos categorías de algoritmos:

- **Clasificación**: para valores de respuesta categóricos, en los que los datos se pueden separar en "clases" específicas
- **Regresión**: para valores de respuesta continua (Valores numéricos)

De lo anterior considerando que nuestras respuestas serán valores categóricos (Bueno o Malo) y no valores numéricos, utilizaremos los Algoritmos de clasificación.

2.1.2.3. Algoritmos de Clasificación

Para González, L. (2018). Los algoritmos de clasificación intentan etiquetar cada ejemplo eligiendo entre dos o más clases diferentes. Estos algoritmos crean modelos predictivos a partir de datos de capacitación que tienen características y etiquetas de clase. Estos modelos predictivos, a su vez usan las características aprendidas de los datos de capacitación sobre datos nuevos, no vistos previamente, para predecir sus etiquetas de clase. Elegir entre dos clases se denomina clasificación binaria, como predecir si alguien aprobará o desaprobará un curso. Elegir entre más de dos clases se denomina clasificación multiclase.

Los tipos de algoritmos de clasificación incluyen:

- Regresión logística
- Vecinos más cercanos
- Máquinas de vectores de soportes
- Árboles de decisión clasificación
- Bosques aleatorios clasificación

- **Regresión logística**

González, L. (2018). La regresión logística o *Logistic Regression* es un algoritmo de clasificación que se utiliza para predecir la probabilidad de una variable dependiente categórica. En la regresión logística, la variable dependiente es una variable binaria que contiene datos codificados como 1 – 0, sí – no, abierto – cerrado, etc.

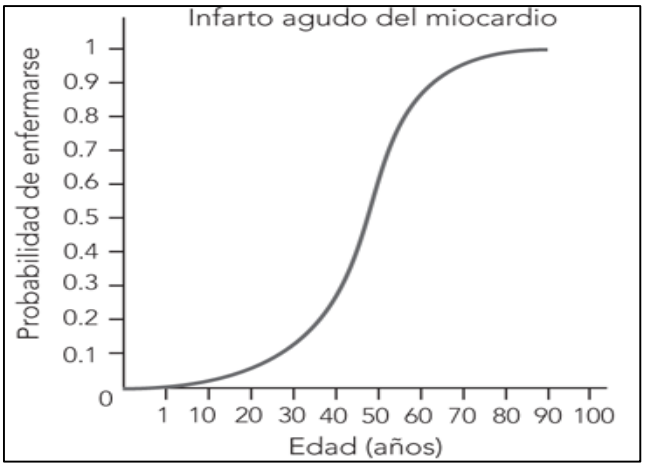

Figura 7.- Regresión Logística

Fuente: imagen extraída de: https://accessmedicina.mhmedical.com/content.aspx?bookid=1721&sectionid=115931846

Este modelo logístico binario se utiliza para estimar la probabilidad de una respuesta binaria basada en una o más variables predictoras o independientes. Como todos los análisis de regresión, la regresión logística es un análisis predictivo. Se usa para describir datos y explicar la relación entre una variable binaria dependiente y una o más variables independientes nominales, ordinales, de intervalo o de nivel de razón. La regresión logística requiere tamaños de muestra bastante grandes.

La razón por la cual la regresión logística es ampliamente utilizada, a pesar de los algoritmos avanzados como redes neuronales profunda, es porque es muy eficiente y no requiere demasiados recursos computacionales que hacen que sea asequibles ejecutar la producción.

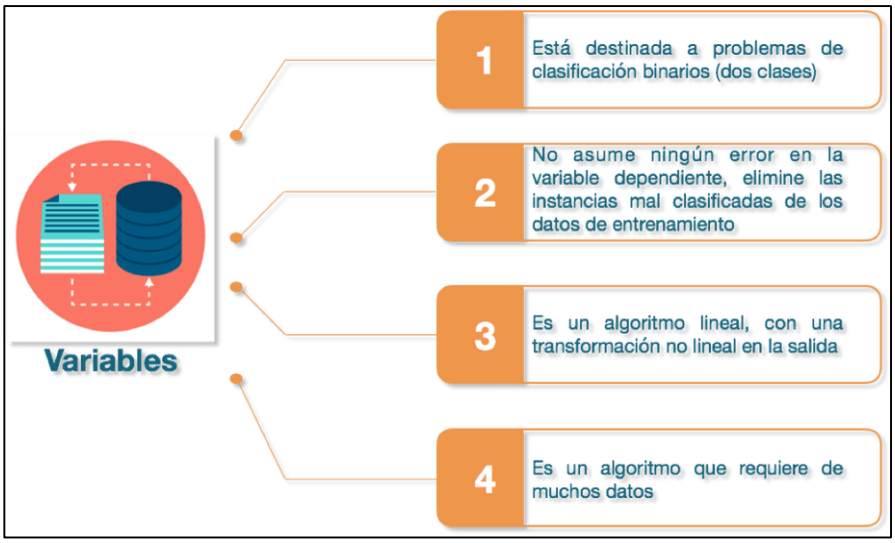

Figura 8.- Características de los algoritmos de regresión logística

Fuente: Imagen extraída de: http://ligdigonzalez.com/aprendizaje-supervisado-logistic-regression/

- **Vecinos más cercanos**

González, L. (2018). El algoritmo KNN es uno de los algoritmos de clasificación más simples, incluso con tal simplicidad puede dar resultados altamente competitivos. Pertenece al dominio de aprendizaje supervisado y puede ser utilizado para el reconocimiento de patrones, extracción de datos y detección de intrusos.

Es un clasificador robusto y versátil que a menudo se usa como un punto de referencia para clasificadores más complejos como las redes neuronales artificiales y vectores de soporte (SVM). A pesar de su simplicidad, KNN puede superar a los clasificadores más potentes y se usa en una variedad de aplicaciones tales como pronósticos económicos, compresión de datos y genética.

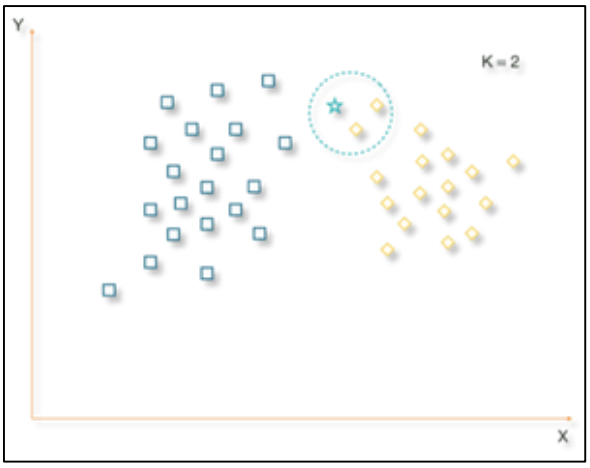

Figura 9.- Vecinos más cercanos

Fuente: Imagen extraída de: http://ligdigonzalez.com/aprendizaje-supervisado-k-nearest-neighbors/

Este algoritmo consiste en seleccionar un valor de K. Al momento del análisis los K datos más cercanos al valor que se desea predecir será la solución.

Acá lo importante es seleccionar un valor de K acorde a los datos para tener una mayor precisión en la predicción.

Figura 10.- Características de los algoritmos KNN

Fuente: Imagen extraída de: http://ligdigonzalez.com/aprendizaje-supervisado-k-nearest-neighbors/

- **Máquinas de vectores de soportes**

González, L. (2018). El algoritmo de vectores de soporte o *Support Vector Machine* es un clasificador discriminatorio definido formalmente por un hiperplano de separación. En otras palabras, dados los datos de entrenamiento etiquetados el algoritmo genera un hiperplano óptimo que clasifica los nuevos ejemplos en dos espacios dimensionales, este hiperplano es una línea que divide un plano en dos partes donde en cada clase se encuentra en cada lado.

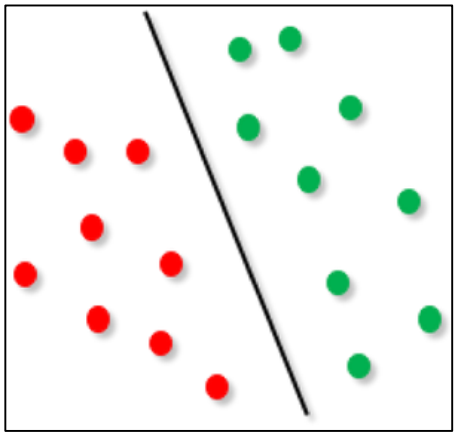

Figura 11.- Máquinas de vectores de soportes

Fuente: Imagen extraída de: http://ligdigonzalez.com/aprendizaje-supervisado-support-vector-machine/

Los vectores de soportes se basan en el concepto de planos de decisión que definen los límites de decisión. Un ejemplo de esto se muestra a continuación, los objetos pertenecen a la clase *verde* o *rojo*, la línea de separación define un límite en el lado derecho del cual todos los objetos son *verdes*, y a la izquierda de los cuales todos los objetos son *rojos*. Cualquier objeto nuevo que caiga hacia la derecha está clasificado como *verde*, o clasificado como *rojo* si cae a la izquierda de la línea de separación.

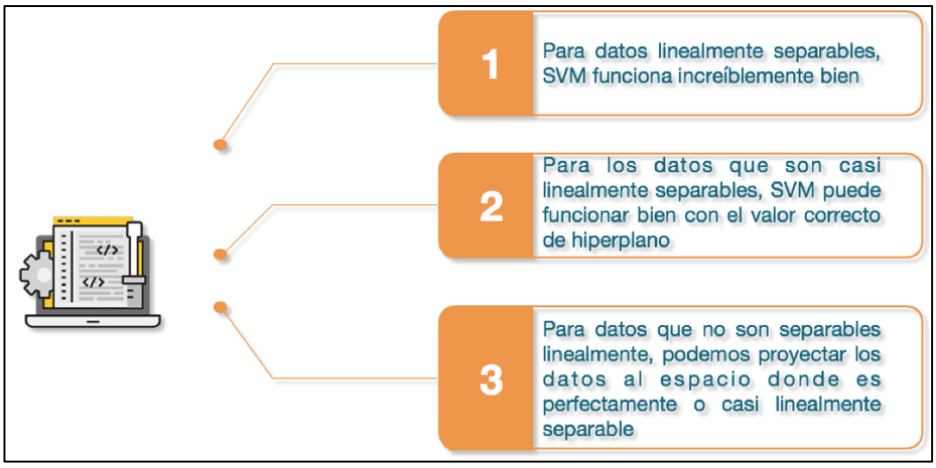

Figura 12.- Características de las Máquinas de vectores de soportes

Fuente: Imagen extraída de: http://ligdigonzalez.com/aprendizaje-supervisado-support-vector-machine/

- **Árboles de decisión**

González, L. (2018). Árbol de decisión o *Decisión Tree Classification* es un tipo de algoritmo de aprendizaje supervisado que se utiliza principalmente en problemas de clasificación, aunque funciona para variables de entrada y salida categóricas como continuas. En esta técnica, dividimos la data en dos o más conjuntos homogéneos basados en el diferenciador más significativo en las variables de entrada. El árbol de decisión identifica la variable más significativa y su valor que proporciona los mejores conjuntos homogéneos de población. Todas las variables de entrada y todos los puntos de división posibles se evalúan y se elige la que tenga mejor resultado.

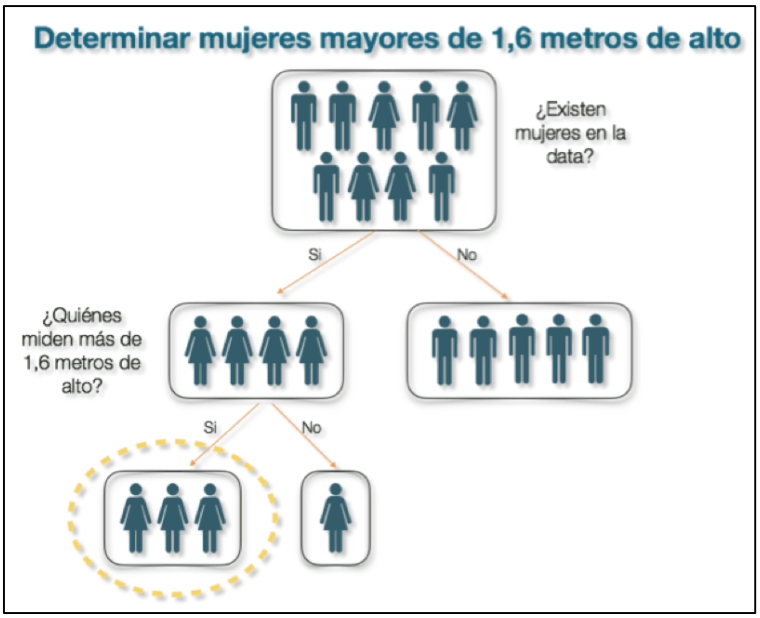

Figura 13.- Arboles de decisión

Fuente: imagen extraída de: http://ligdigonzalez.com/aprendizaje-supervisado-decision-tree-classification/

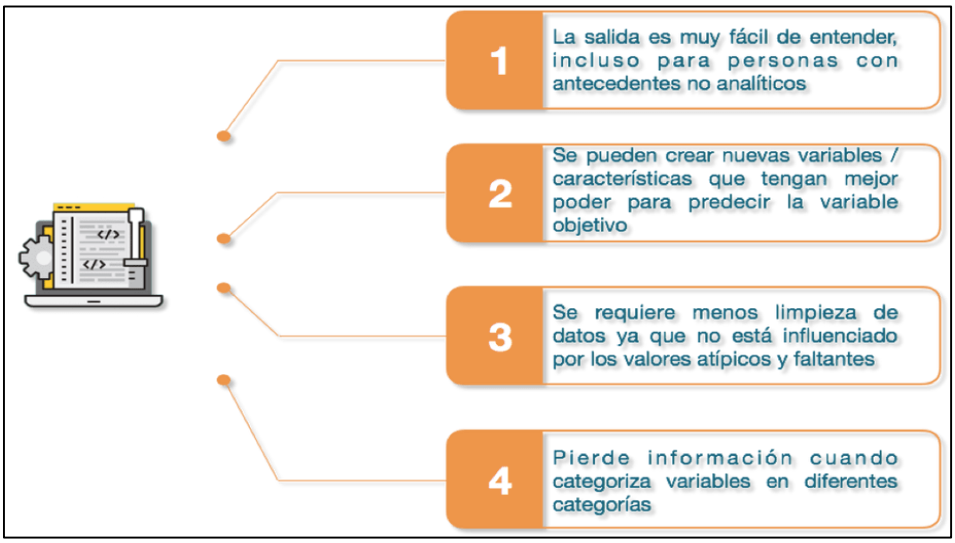

Figura 14.- Características de los árboles de decisión

Fuente: imagen extraída de: http://ligdigonzalez.com/aprendizaje-supervisado-decision-tree-classification/

- **Bosques aleatorios clasificación (Random Forest)**

González, L. (2018). *Random Forest* es un método versátil de aprendizaje automático capaz de realizar tanto tareas de regresión como de clasificación. También lleva a cabo métodos de reducción dimensional, trata valores perdidos, valores atípicos y otros pasos esenciales de exploración de datos. Es un tipo de método de aprendizaje por conjuntos, donde un grupo de modelos débiles se combinan para formar un modelo poderoso**.**

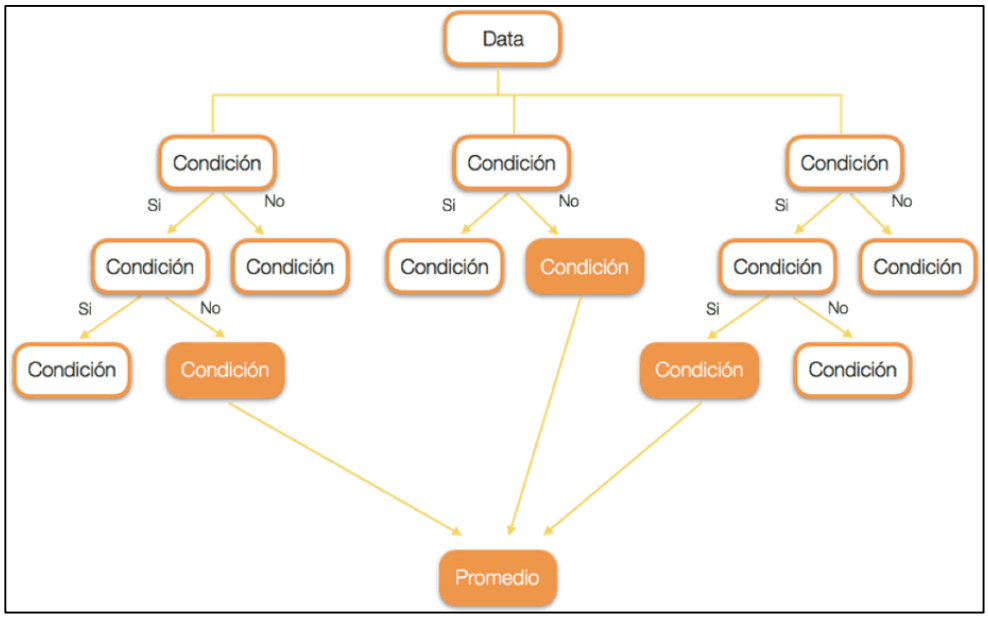

Figura 15.- Random Forest

Fuente: imagen extraída de: http://ligdigonzalez.com/aprendizaje-supervisado-random-forest-classification/

En Random Forest se ejecutan varios algoritmos de árbol de decisiones en lugar de uno solo. Para clasificar un nuevo objeto basado en atributos, cada árbol de decisión da una clasificación y finalmente la decisión con mayor “votos” es la predicción del algoritmo.

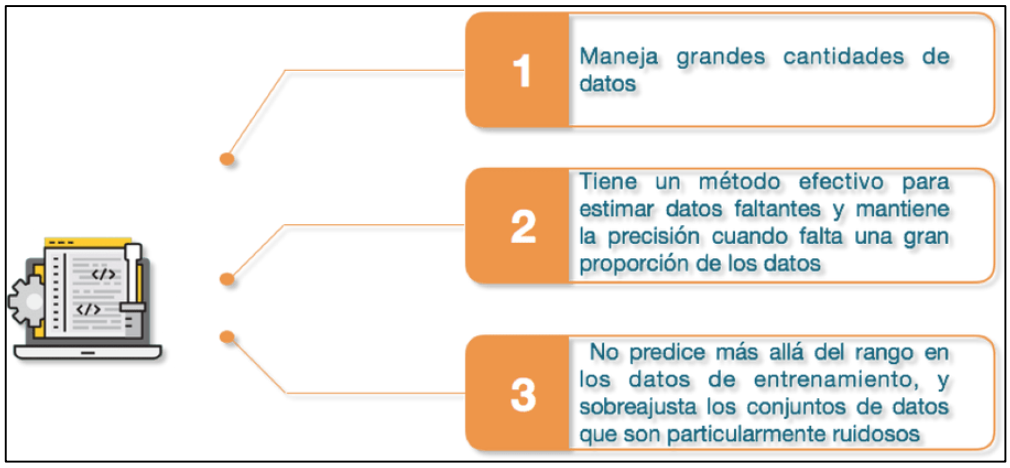

Figura 16.- Características de los algoritmos Random Forest

Fuente: imagen extraída de: http://ligdigonzalez.com/aprendizaje-supervisado-random-forest-classification/

2.1.3. Minería de datos

Como se describió, el aprendizaje automático es también un tipo de inteligencia artificial que le da a las computadoras capacidades para "aprender", este proceso de aprendizaje automático es muy similar al de la minería de datos (*data minino*), es decir ambas formas buscan entre los datos encontrar patrones que puedan ser analizados para intentar generar modelos para predicción, en tal sentido describiré algunas definiciones de minería de datos dada por autores importantes, con la intención de utilizar luego en la presente tesis la metodología de minería de datos CRISP-DM (*Cross Industry Standard Process for Data Mining*) Proceso estándar para la minería de datos.

Según Pérez, C. & Gonzales, S. (2007), "la minería de datos puede definirse como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos".

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

Hernández, J. (2004). “La minería de datos es un término relativamente moderno que integra numerosas técnicas de análisis de datos y extracción de modelos. Es capaz también de extraer patrones, de describir tendencias y regularidades, de predecir comportamientos y, en general, de sacar partido de la información computarizada que nos rodea hoy en día, generalmente heterogénea y en grandes cantidades permite a los individuos y a las organizaciones comprender y modelar de una manera más eficiente y precisa el contexto en que deben actuar y tomar decisiones”.

Marques M, (2014) define la minería de datos como “un conjunto de técnicas encaminadas al descubrimiento de la información contenida en grandes conjuntos de datos. Se trata de analizar comportamientos, patrones, tendencias, asociaciones y otras características del conocimiento inmerso en los datos. Actualmente se dispone de grandes cantidades de datos y es más necesario que nunca poder analizarlos ordenadamente para extraer de un modo automatizado la inteligencia contenida en ellos utilizando técnicas especializadas apoyadas en herramientas informáticas. Estas técnicas constituyen la minería de datos”.

2.1.4. Metodología CRISP-DM

En los años noventa se desarrollaron una serie de metodologías para la minería de datos, en ingles *Cross Industry Standard Process for Data Mining* (CRISP-DM), es una de las metodologías que más se difundió, se trata de un modelo estándar abierto del proceso que describe los enfoques comunes que utilizan los expertos en minería de datos.

Según Mayorga (2016). "Esta metodología provee algunos pasos que se requieren para un proyecto de minería de procesos, como son el entendimiento de los objetivos del negocio, recopilar, describir, analizar y limpiar datos. Sin embargo, en la etapa de modelación, los pasos son poco específicos y no dan un respaldo adecuado, para lo que se requieren el diagnóstico y análisis con minería de procesos".

Según el Manual CRISP-DM de IBM SPSS Modeler, CRISP-DM, "Es un método probado para orientar sus trabajos de minería de datos. Como metodología, incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas. Como modelo de proceso, CRISP-DM ofrece un resumen del ciclo vital de minería de datos".

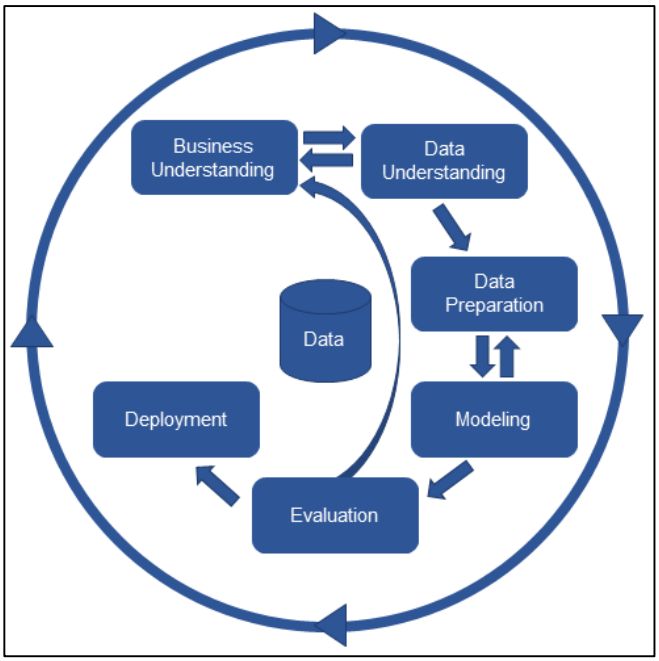

Figura 17.- Etapas de la Metodología CRISP – DM

Fuente: Manual CRISP-DM de IBM SPSS Modeler (2012)

El ciclo vital del modelo contiene seis fases con flechas que indican las dependencias más importantes y frecuentes entre fases. La secuencia de las fases no es estricta. De hecho, la mayoría de los proyectos avanzan y retroceden entre fases si es necesario.

Villena, J. (2016) describe cada una de las fases de CRISP-DM de la siguiente manera:

Fase I. *Business Understanding. Definición de necesidades del cliente* (comprensión del negocio)

Esta fase se enfoca en la comprensión de los objetivos de proyecto.

Esta comprensión del contexto en el que nos encontramos permitirá generar un conocimiento de los datos para la definición del problema de aprendizaje

automático y nos permitirá generar un modelo preliminar para alcanzar nuestros objetivos.

Fase II. *Data Understanding*. Estudio y comprensión de los datos

La fase de entendimiento de datos comienza con la colección de datos inicial y continúa con las actividades que permiten familiarizarse con los datos, identificar los problemas de calidad, descubrir conocimiento preliminar sobre los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.

Fase III. *Data Preparation*. Análisis de los datos y selección de características

La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto final de datos (los datos que se utilizarán en las herramientas de modelado) a partir de los datos en bruto iniciales.

Estas tareas implican la limpieza y transformación de la información respecto de las instancias (filas) y los atributos (columnas) para su posterior procesamiento con las herramientas de modelado.

Fase IV. *Modeling*. Modelado

El modelado es una fase en la que seleccionaremos las técnicas de modelado pertinentes al problema de predicción, y se definir o pulir los parámetros a valores válidos y óptimos. Algunas técnicas tienen requerimientos específicos sobre la forma de los datos. Por lo tanto, casi siempre en cualquier proyecto se acaba volviendo a la fase de preparación de datos.

Fase V. *Evaluation*. Evaluación (obtención de resultados)

En esta etapa en el proyecto, se han construido uno o varios modelos que parecen alcanzar calidad suficiente desde la una perspectiva de análisis de datos.

Antes de proceder al despliegue final del modelo, es importante evaluarlo a fondo y revisar los pasos ejecutados para crearlo, comparar el modelo obtenido con los objetivos de negocio. Un objetivo clave es determinar si hay alguna cuestión importante de negocio que no haya sido considerada suficientemente. Al final de esta fase, se debería obtener una decisión sobre la aplicación de los resultados del proceso de análisis de datos.

Fase VI. *Deployment*. Despliegue (puesta en producción)

Generalmente, la creación del modelo no es el final del proyecto. Incluso si el objetivo del modelo es de aumentar el conocimiento de los datos, el conocimiento obtenido tendrá que organizarse y presentarse para que el cliente pueda usarlo. Dependiendo de los requisitos, la fase de desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización periódica y quizás automatizada de un proceso de análisis de datos en la organización.

En figura 18 se presenta una guía visual y completa de todas las fases, listando las tareas a realizar en cada una de las fases, así como las conexiones entre ellas y las iteraciones que pueden llevarse a cabo, los cuales serán utilizadas para el desarrollo del presente trabajo en los capítulos siguientes.

Esta figura ha sido tomada de «a visual guide to CRISP-DM methodology», (2009)

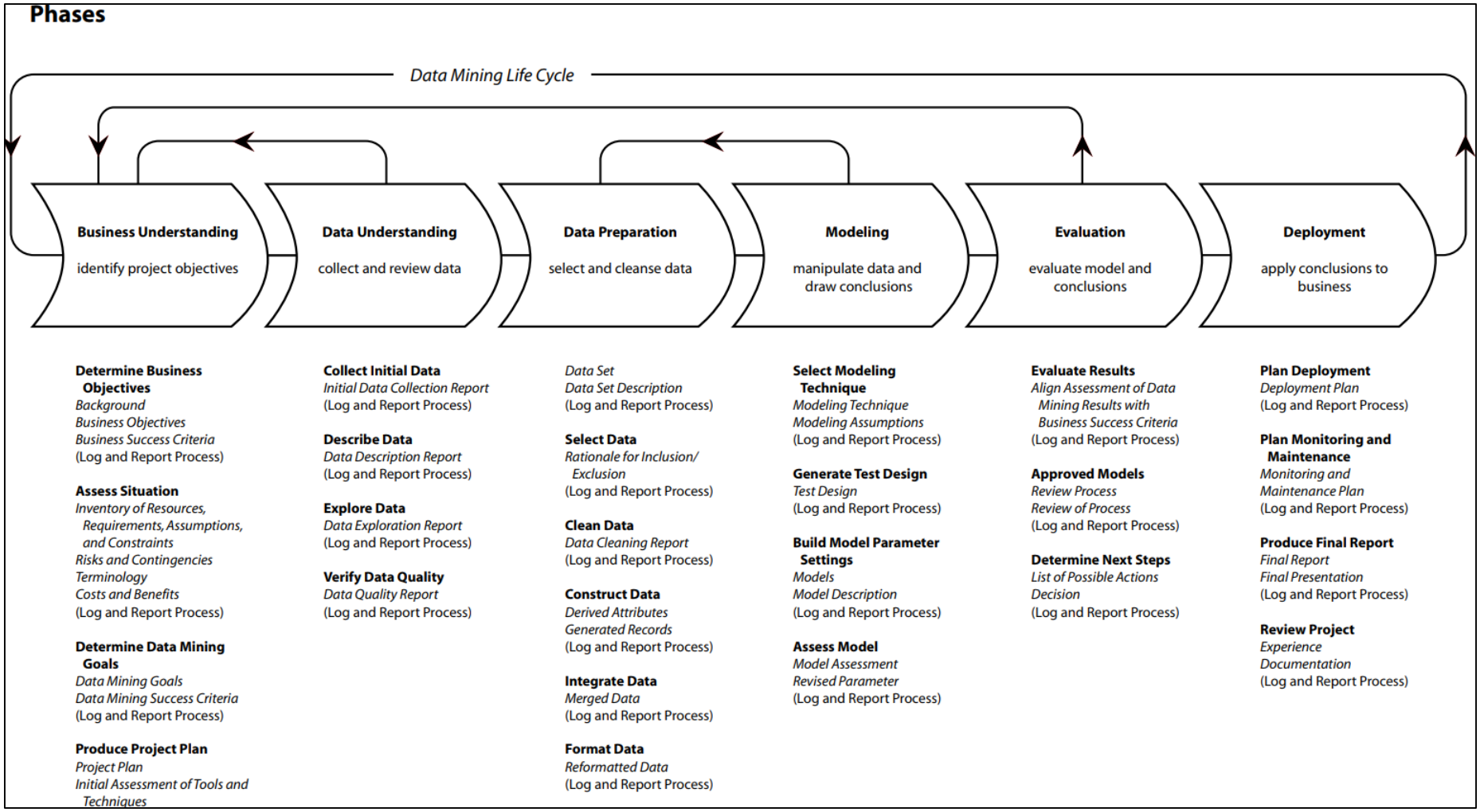

Figura 18.- Fases y tareas de la metodología CRISP DM

Fuente: Esta imagen ha sido tomada de: https://exde.wordpress.com/2009/03/13/a-visual-guide-to-crisp-dm-methodology/

A partir de las descripciones anteriores podemos manifestar que la metodología CRISP-DM (*Cross Industry Standard Process for Data Mining*) es una metodología completa porque tiene en cuenta el entorno del negocio y la parte de la secuencia de las fases no es rígida, y nos permitirá desarrollar el presente proyecto en cada una de sus fases descritas a continuación:

Fase 1: Comprensión del Negocio

- ✓ Comprender del contexto y determinar objetivos
- ✓ Evaluación de la situación.
- ✓ Determinación de objetivos de la Minería de Datos.
- ✓ Desarrollo del plan del proyecto.

Fase 2: Comprensión de los datos

- ✓ Recopilación de datos iniciales.
- ✓ Descripción los datos.
- ✓ Exploración de los datos.
- ✓ Verificación de la calidad de los datos.

Fase 3: Preparación de los datos

- ✓ Selección de datos más relevantes.
- ✓ Limpieza de datos.
- ✓ Construcción de nuevos datos (atributos).
- ✓ Integración de datos.

Fase 4: Modelado

- ✓ Selección de técnicas de modelado.

- ✓ Generación de un diseño de comprobación.
- ✓ Generación del modelo.
- ✓ Evaluación y comprobación del modelo.

Fase 5: Evaluación

- ✓ Evaluación de resultados.
- ✓ Proceso de revisión.
- ✓ Determinar los pasos siguientes a base de los resultados.

Fase 6: Despliegue

- ✓ Planificación de distribución.
- ✓ Creación del informe final.
- ✓ Revisión final del proyecto.

2.1.5. WEKA

Para Córdoba, L. (2011). WEKA "es una herramienta de tipo software para el aprendizaje automático y minería de datos diseñado a base de Java y desarrollado en la universidad de Waikato en Nueva Zelanda en el año 1993, esta herramienta por su nombre en inglés (*Waikato Environment for Knowledge Analysis*) además es una herramienta de distribución de licencia GNU-GLP o software libre".

WEKA contiene una colección de algoritmos para realizar análisis de datos y modelado predictivo, también tiene herramientas para la visualización de estos datos, además provee una interfaz gráfica que unifica las herramientas para que estén a una mejor disposición.

Sus características más importantes son:

- Es una herramienta muy versátil que soporta muchas tareas estándar de la minería de datos en especial tareas de procesamiento de datos, regresión, clasificación, *clustering* entre otras, así mismo permite la visualización y la selección de los datos.
- Todas las técnicas en WEKA están basadas en la sunción de datos que están disponibles en un fichero plano o una relación, en donde cada registro de datos esta descrito por un número fijo de atributos nominales o numéricos.
- Permite el acceso a otras instancias de bases de datos por medio de SQL, gracias al JDBC, además puede procesar un resultado generado a base de una consulta hecha a una base de datos

Figura 19.- Pantalla de inicio de WEKA 3.8

Fuente: Propia, Imagen capturada de WEKA 3.8

La herramienta WEKA, está compuesta por herramientas gráficas de visualización tiene a su vez diferentes algoritmos para el análisis de datos y modelado predictivo.

La interfaz gráfica de WEKA facilita al usuario el acceso a sus múltiples funcionalidades.

Esta potente herramienta de minería de datos se encuentra libremente disponible bajo la licencia pública general de GNU, implementada en Java.

La interfaz gráfica de WEKA cuenta con 4 formas de acceso a las diferentes funcionalidades de la aplicación (*Explorer, Experimenter, KnowledgeFlow y Workbench*) a continuación describo las 2 primeras, ya que son las herramientas que se utilizan en el presente proyecto.

- ***Explorer***, es la opción más intuitiva para el usuario, pues dispone de varios paneles que dan acceso a las principales características del programa

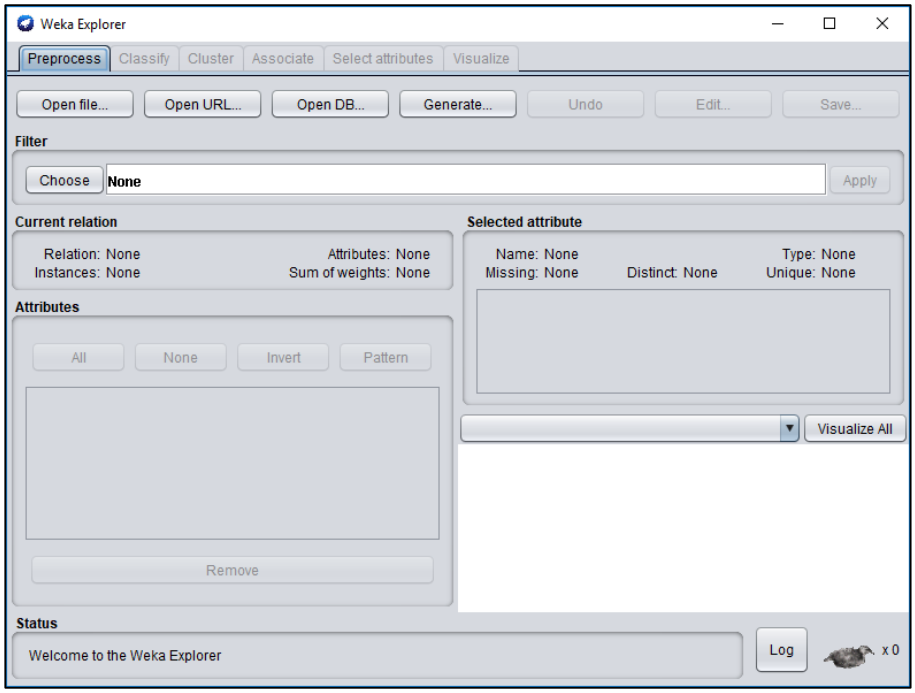

Figura 20.- WEKA Explorer

Fuente: Propia, Imagen capturada de WEKA 3.8.

- ***Experimenter***, esta herramienta nos permite la comparación de los algoritmos predictivos de WEKA sobre un conjunto de datos a partir de WEKA *Explorer*

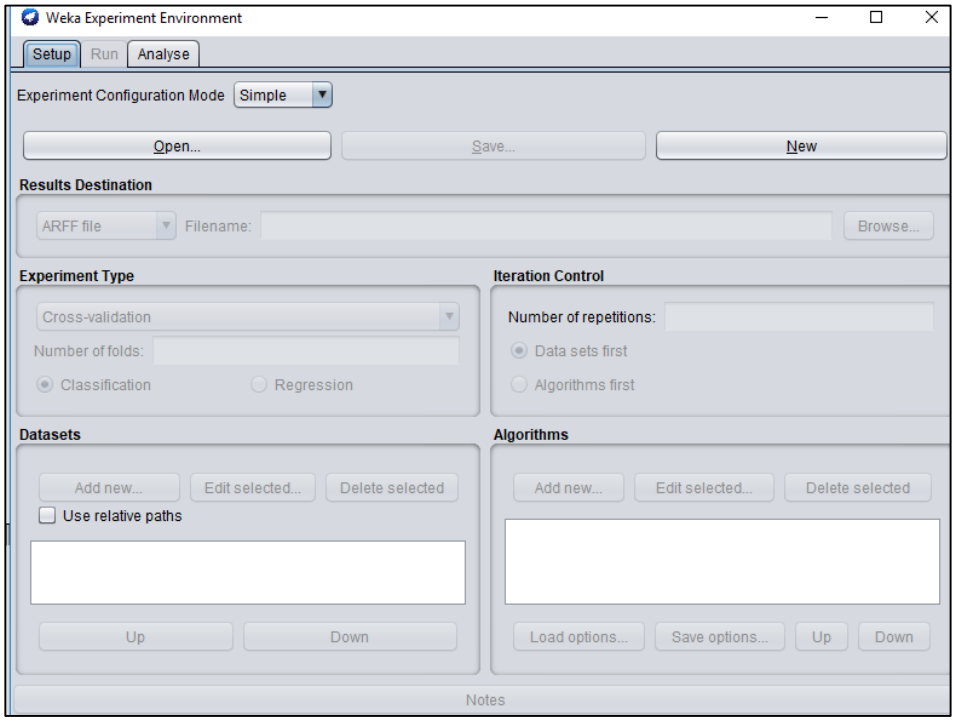

Figura 21.- WEKA Experimenter

Fuente: Propia, Imagen capturada de WEKA 3.8

2.1.5.1. Preparación de los datos con WEKA 3.8

Los datos deberán estar creados como un archivo ".arff", y los atributos separados por comas, esta información que contiene los atributos (Columnas) e instancias (Filas), son cargadas utilizando la herramienta "*Explorer*", y nos mostrara lo siguiente:

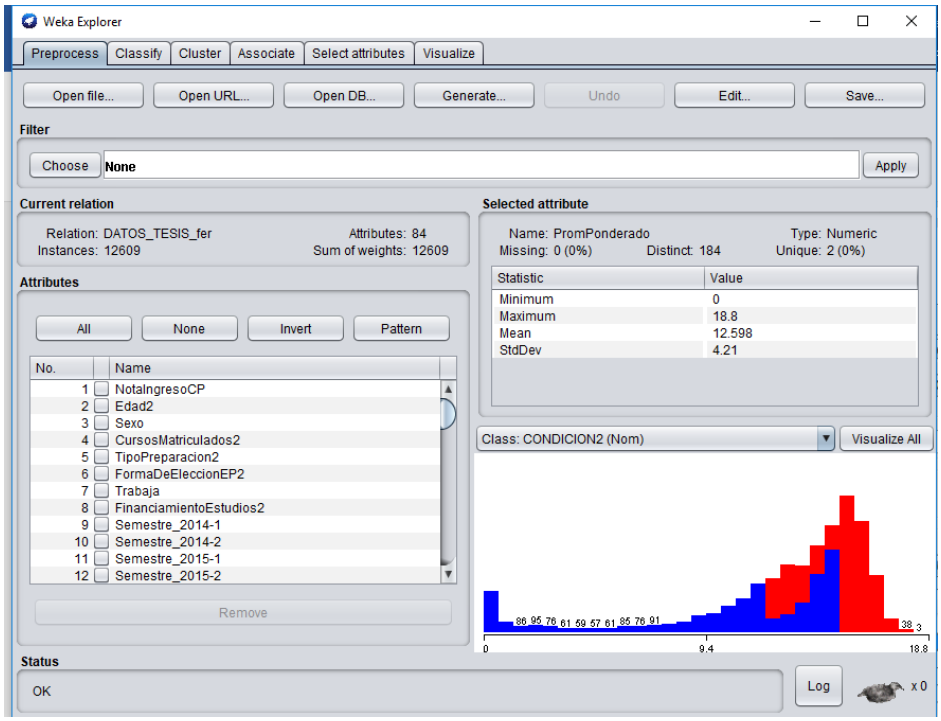

Figura 22.- Carga del Archivo a procesar sobre WEKA 3.8

Fuente: Propia, Imagen capturada de WEKA 3.8

Como podemos observar en la figura 22, WEKA reconoce todos los atributos que forman nuestro origen de datos. Automáticamente, asocia cada atributo de tipo nominal o numérico, según el contenido de los datos.

Adicionalmente, nos muestra información relevante para cada atributo, si vamos seleccionándolos uno a uno, nos muestra en los cuadros de la derecha varios datos:

- Nombre del atributo
- Valores máximos
- Valores mínimos
- Media
- Desviación estándar
- Un histograma que muestra una distribución de los valores para este determinado atributo

2.1.6. Pre procesamiento de datos para el aprendizaje automático

El pre procesamiento y la limpieza de datos según Kaufmann, M. (2011) son tareas importantes que normalmente se deben llevar a cabo para que el conjunto de datos se pueda utilizar de forma eficaz para el aprendizaje automático. Los datos sin procesar son a menudo ruidosos no confiables y es posible que les falten valores.

Las tareas de procesamiento previo y de limpieza, como la tarea de exploración de datos, se pueden llevar a cabo en una amplia variedad de entornos, como SQL y/o con diversas herramientas y lenguajes, como R, Python o WEKA.

2.1.6.1. Procesamiento y limpieza de datos

Kaufmann, M. (2011), Se recopilan datos del mundo real de varios orígenes y procesos y pueden contener irregularidades o datos dañados que comprometen la calidad del conjunto de datos. Los problemas de calidad de datos más habituales que surgen son:

- **Incompletos**: en los datos no hay atributos o contienen valores que faltan.
- **Con ruido**: los datos contienen registros erróneos o valores atípicos.
- **Incoherentes**: los datos contienen discrepancias o registros en conflicto.

Los datos de calidad son un requisito previo para los modelos predictivos de calidad. Para evitar la "entrada y salida de elementos no utilizados" y mejorar la calidad de los datos y, por tanto, el rendimiento del modelo, es fundamental llevar a cabo una pantalla de mantenimiento de datos para detectar problemas de

datos al principio y decidir acerca de los pasos de limpieza y pre procesamiento de datos correspondientes.

2.1.6.2. Tareas principales de pre procesamiento de datos

Kaufmann, M. (2011), considera las siguientes tareas:

- **Limpieza de datos**: rellene los valores que faltan, detecte y quite los valores atípicos y los datos con ruido.
- **Transformación de datos**: normalice los datos para reducir el ruido y las dimensiones.
- **Reducción de datos**: registros de datos de ejemplo o atributos para un control de datos más sencillo.
- **Discretización de datos**: convierta atributos continuos en atributos de categorías para facilitar su uso con determinados métodos de aprendizaje automático.
- **Limpieza de texto**: quite caracteres incrustados que puedan ocasionar errores en la alineación de los datos, por ejemplo, pestañas incrustadas en un archivo de datos separado por tabulaciones, nuevas líneas incrustadas que pueden dividirse en registros, etc.

2.1.6.3. Tratamiento de valores faltantes

Kaufmann, M. (2011). Para tratar los valores que faltan, es mejor identificar el motivo por el que faltan los valores para controlar mejor el problema. Los métodos de control de valores que faltan típicos son:

- **Eliminación**: quite los registros con los valores que faltan

- **Sustitución ficticia**: reemplace los valores que faltan por un valor ficticio; por ejemplo, *desconocido* para categorías o 0 para valores numéricos.
- **Sustitución media**: si los datos que faltan son numéricos, reemplace los valores que faltan por la media.
- **Sustitución frecuente**: si los datos que faltan son de categoría, cambie los valores que faltan por el elemento más frecuente
- **Sustitución de regresión**: utilice un método de regresión para reemplazar los valores que faltan por valores con regresión.

2.1.6.4. Normalización de datos

Kaufmann, M. (2011). La normalización de datos escala los valores numéricos a un intervalo especificado. Entre los métodos de normalización de datos más conocidos se incluyen:

- **Normalización mínimo-máximo**: transforme linealmente los datos a un intervalo, por ejemplo, entre 0 y 1, donde el valor mínimo se escala a 0 y el máximo a 1.
- **Normalización de puntuación Z**: escale los datos en función de la desviación estándar y media: divida la diferencia entre los datos y la media por la desviación estándar.
- **Escalado decimal**: escale los datos moviendo la coma decimal del valor del atributo.

2.1.6.5. Discretización de datos

Kaufmann, M. (2011). Los datos se pueden discretizar mediante la conversión de valores continuos en intervalos o atributos nominales. Algunas formas de hacerlo son las siguientes:

- **Discretización del mismo ancho**: divida el intervalo de todos los valores posibles de un atributo en N grupos del mismo tamaño y asigne los valores que se encuentran en una ubicación con el número de ubicación.
- **Discretización del mismo alto**: divida el intervalo de todos los valores posibles de un atributo en N grupos, cada uno de los cuales contiene el mismo número de instancias y, a continuación, asigne los valores que se encuentran en una ubicación con el número de ubicación.

2.1.7. Matriz de confusión

Una matriz de confusión en el campo de la Inteligencia Artificial es una herramienta que permite visualizar el desempeño de un algoritmo de aprendizaje supervisado

Para Sierra, B. (2006), la Matriz de Confusión nos permite ver, mediante una tabla de contingencia la distribución de los errores cometidos por un clasificador a lo largo de las distintas categorías del problema. En dicha tabla de contingencia se cruza la variable derivada de la clasificación predicha por el clasificador con la variable que guarda la verdadera clasificación.

Así podremos comprobar si el algoritmo está clasificando mal las clases y en qué medida.

En la tabla 1 se muestra una matriz de confusión con un clasificador de 2 clases:

Tabla 1.- *Matriz de confusión para un caso de estudio de dos clases*

		Clasificador	
		Negativos	***Positivos***
Valores	***Negativos***	a	b
Reales	***Positivos***	c	d

Fuente: Elaboración Propia

Donde:

- ***a*** es el número de predicciones correctas de que un caso es negativo.
- ***b*** es el número de predicciones incorrectas de que un caso es positivo, o sea la predicción es positiva cuando realmente el valor tendría que ser negativo. A estos casos también se les denomina errores de tipo I.
- ***c*** es el número de predicciones incorrectas de que un caso es negativo, o sea la predicción es negativa cuando realmente el valor tendría que ser positivo. A estos casos también se les denomina errores de tipo II.
- ***d*** es el número de predicciones correctas que un caso es positivo.

De una matriz de confusión también se puede extraer los siguientes conceptos que enriquecen a la hora de comprender la distribución y los errores cometidos por los clasificadores:

- **Sensibilidad** (*Se*, del inglés *Sensitivity*) como sinónimo de *TPrate* porque es la capacidad del clasificador de ser "sensible" a los casos positivos. Es la proporción de verdaderos positivos

- **Especifidad** (*Sp*, del inglés *Specificity*) como sinónimo de *TNrate*, porque puede dar una medida de la especificidad del test para marcar los casos positivos. Es la proporción de verdaderos negativos.

Otros conceptos importantes según EcuRed (2014) son:

- **La Exactitud** (*Ac*, del inglés *Accuracy*) es la proporción del número total de predicciones que fueron correctas:

$$Ac = \frac{a + d}{a + b + c + d}$$

- **La Razón de Verdaderos Positivos** (*TPrate*, del inglés *True Positive Rate*), a veces también denominada *Recall*, es la proporción de casos positivos que fueron correctamente identificados:

$$TPrate = \frac{d}{c + d}$$

- **La Razón de Falsos Positivos** (*FPrate*, del inglés *False Positive Rate*) es la proporción de casos negativos que han sido incorrectamente clasificados como positivos:

$$FPrate = \frac{b}{a + b}$$

- **La Razón de Verdaderos Negativos** (*TNrate*, del inglés *True Negative Rate*) es la proporción de casos negativos que han sido correctamente clasificados

$$TNrate = \frac{a}{a + b}$$

- **La Razón de Falsos Negativos** (*FNrate*, del inglés *False Negative Rate*) es la proporción de casos positivos que fueron incorrectamente clasificados como negativos:

$$FNrate = \frac{c}{c + d}$$

- **La precisión** (P, en inglés, también *Precision*) es la proporción de casos predichos positivos que fueron correctos

$$P = \frac{d}{b + d}$$

2.2. Antecedentes empíricos de la Investigación

A continuación, se describirán los trabajos e investigaciones que se asemejan al presente trabajo, con el objetivo de describir y conocer los aportes y las conclusiones a las que pudieron arribar cada uno de los autores.

2.2.1. Antecedentes Internacionales

A. García, D (2015), *Construcción de un modelo para determinar el rendimiento académico de los estudiantes basado en Learning AnalyTIC's (Análisis del aprendizaje), mediante el uso de técnicas multivariantes*, Universidad de Sevilla, España.

La tesis doctoral, desarrolla una revisión de dos metodologías que se adaptan a las características de los datos en la educación superior y se consideran como las más adecuadas, un análisis multinivel para hacer mediciones desde un ámbito cuantitativo y un análisis logístico bivariante que permite hacer mediciones desde un ámbito cualitativo, entre las principales conclusiones arribadas están:

- La investigación sobre el rendimiento académico es amplia y muy actual en el ámbito universitario, dada la necesidad de mejorar y optimizar los resultados de la inversión de recursos (económicos, humanos y sociales). En consecuencia, es esencial la búsqueda de los factores que inciden en la mejora de los mismos y, por tanto, se puede y se debe realizar otros trabajos para la consecución de variables idóneas.
- La variable género no es estadísticamente significativa en ninguna de las estimaciones,

- La variable edad indica que los estudiantes jóvenes tienen menos ventaja (entendiendo "ventaja" como la razón entre las probabilidades de éxito y fracaso) de conseguir un rendimiento académico óptimo con respecto a los estudiantes en edad adulta,
- En cuanto a la región de procedencia no se presenta ningún efecto en el rendimiento académico.
- Todas las variables del enfoque "*learning analyTIC's*" tienen una relación positiva con el rendimiento académico, siendo la participación en chat, foro y video-colaboración las que ocasionan el mayor impacto ya que provocan un incremento de entre 1 y 2 puntos en el rendimiento académico.

La investigación desarrollada desde 2 puntos de vista cuantitativa y cualitativa, describe en detalle las variables de éxito o fracaso en el rendimiento académico de los estudiantes, considerando principalmente que "learning analyTIC's" o analítica de aprendizaje, tiene una relación directa con el rendimiento académico positivo de los estudiantes, del mismo modo indica que los estudiantes de menor edad tienen menor ventaja de obtener un mejor rendimiento académico, y la procedencia y el género no influyen directamente.

B. Márquez, C (2015), *Predicción del fracaso y el abandono escolar mediante técnicas de minería de datos*, Universidad de Córdoba, España.

Tesis doctoral, que propone una metodología basada en técnicas de Minería de Datos para predecir lo más temprano posible en el periodo escolar a aquéllos que estén en riesgo de suspender o abandonar, es decir, sentar las bases para que se pueda implementar un Sistema de Alerta Temprana, para una vez detectados poder tomar decisiones en cuanto a qué tipo de apoyo o

intervención requiere cada uno de ellos para en lo posible impedir el fracaso, o bien, reducirlo y retener a los estudiantes en la institución educativa, sus principales conclusiones son:

- La tarea de predecir el fracaso escolar de los estudiantes es una tarea muy difícil de conseguir, principalmente por dos causas: la primera es que son muchos los factores que pueden influir en los estudiantes para que reprueben o abandonen sus estudios; y la segunda es que generalmente la información con la que se trabaja para predecir a estos estudiantes está desbalanceada, es decir, no hay igual número de alumnos que aprueban y pasan curso que de alumnos que suspenden, no pasan y/o abandona el curso. Para solventar estas dos dificultades en esta tesis se ha propuesto la utilización de técnicas de DM que desde hace un tiempo atrás se han empezado a utilizar de manera creciente y con éxito, en lugar de las tradicionales técnicas estadísticas.
- La metodología propuesta para predecir a los estudiantes en riesgo de fracaso está basada en la utilización de diferentes técnicas de DM. Durante el pre-procesado se ha realizado un análisis y selección de los mejores atributos para reducir la alta dimensionalidad de los datos. Y se han obtenido mejores resultados usando solamente los mejores atributos en lugar de todos los disponibles.
- La predicción se ha realizado mediante la utilización de algoritmos de clasificación ya que la clase a predecir era de tipo nominal o categórica. Este método ha obtenido muy buenos resultados de clasificación con un conjunto de datos reales que estaban desbalanceados. Además, la clase

minoritaria, que corresponde a los estudiantes en riesgo, es la que precisamente más interesa en este trabajo.

- Este trabajo explica cómo se puede predecir a los estudiantes en riesgo de reprobar o abandonar sus estudios. Ya se mostró que esto se consigue fundamentalmente a través de recoger información de los alumnos, sin embargo, es imposible desarrollar un método infalible, que pueda predecir a la totalidad de los estudiantes que fracasan; siempre hay situaciones que no se pueden predecir y que pueden llevar a los estudiantes a su fracaso.

  La predicción fue realizada utilizando algoritmos de clasificación, considerando que los datos son de tipo nominal y categoría, similares a la presente investigación, también manifiesta el autor que sería imposible desarrollar un método que sea infalible, pero que se podría medir con porcentajes de aproximación para la predicción del rendimiento académico.

C. Quintana, V & Yagual, S (2017) *Propuesta de aplicación predictiva de aprobación de una asignatura con flujo previo a través de algoritmos basados en software WEKA para estudiantes del último semestre de la Carrera de Ingeniería en Sistemas Computacionales de la Universidad de Guayaquil*, Guayaquil, Ecuador

La propuesta busca inferir el rendimiento académico de los estudiantes de la Carrera de Ingeniería de Sistemas Computacionales de la Universidad de Guayaquil en una asignatura con flujo previo establecido, para el análisis y tratamiento de datos dentro de la Minería de Datos, la herramienta utilizada es el software WEKA. Con esto se busca cumplir los objetivos establecidos permitiendo a los estudiantes puedan mejorar su rendimiento académico

conllevando a la aprobación de una determinada asignatura, las conclusiones de este trabajo son:

- Es posible la extracción de datos utilizando una muestra de los registros existentes en la base de datos de la Carrera Profesional, que nos fue otorgada por las autoridades competentes, a su vez procedimos a realizar un análisis con los registros de número de veces cursados en las asignaturas que guardan una relación previa.
- Con respecto a la herramienta WEKA, que fue utilizada para la realización del proyecto se concluye que su utilización es factible para la realización de una inferencia sobre el rendimiento académico de los estudiantes, consiguiendo analizar y preparar los datos obtenidos para luego proceder a evaluar con los datos finales.
- Se logró una predicción de la cantidad de estudiantes que pueden aprobar o reprobar la asignatura, para ayudar a que los alumnos estén preparados y así no llegar afectar en el proceso de finalización de malla curricular.
- Este proyecto ayuda a que los alumnos se incentiven a tratar de mejorar su nivel académico en conjunto con sus docentes los cuales deberán influenciar en su rendimiento para que los estudiantes puedan tener éxito en su carrera profesional.
- La minería de datos es una herramienta muy importante que permite tomar decisiones sobre el comportamiento de los datos. No existe ninguna herramienta que asegure una confianza del 100% en lo que respecta al ámbito de la minería de datos para ejercer predicciones sobre registros ya existentes que son estimados para una determinada evaluación. WEKA puede ser considerada como herramienta óptima para ejercer este proceso

de Minería de Datos debido a la variedad de algoritmos y filtros que contiene.

La tesis describe la utilización de WEKA para la predicción, el cual también será utilizado en el presente proyecto, el autor considera que WEKA es una herramienta óptima para la inferencia de la predicción del rendimiento académico por la variedad de algoritmos y filtros que esta tiene, del mismo modo manifiestan que no existen herramientas que aseguren una predicción al 100%.

D. Galán, V (2015) *Aplicación de la metodología CRISP-DM a un proyecto de minería de datos en el entorno universitario***,** Universidad Carlos III de Madrid, España

Este proyecto de fin de carreara, muestra la utilización de la metodología CRISP-DM, a un entorno universitario, utilizando también para el procesamiento de información SQL-Server, entre las principales conclusiones están:

- El uso de la metodología CRISP-DM en este Proyecto ha permitido encontrar un comportamiento predictivo a la hora de estimar la duración de la carrera de los alumnos y la nota media de los mismos. Se ha podido encontrar un plan de extracción, normalización, y codificación de datos para la realización de procesos de minería de datos cuatrimestrales.
- De los tres objetivos de minería de datos iniciales que se habían fijado se han podido alcanzar dos de ellos (objetivos 1 y 2). Además, al margen de estos objetivos, se han sacado otras conclusiones a partir de los datos estudiados, concretamente se han identificado las asignaturas más problemáticas para los alumnos de cada una de las titulaciones estudiadas.
- Se realizaron consultas significativas en SQL para tener muestras representativas de los datos, y sacar más conclusiones al margen de los objetivos iniciales de la minería.
- El lado positivo de haber creado nosotros mismos la base de datos a nuestro antojo es que la fase de preparación de los datos fue mucho más sencilla ya que no hizo falta apenas hacer una limpieza de los datos, conversiones o formateo de los mismos.

- A continuación, se realizó la elección de las técnicas de modelado y la ejecución de dichas técnicas sobre los datos empleando la herramienta escogida para ello (*Oracle Data Miner*). Esta herramienta facilitó por completo la aplicación de los modelos ya que nos permitió ver de manera muy intuitiva y visual cuales eran las técnicas más adecuadas para nuestra base de datos.
- Por último, una vez obtenidos los modelos, se analizaron para determinar la adecuación o no de los mismos. En este caso determinamos que los modelos 1 y 2 podrían ser válidos para nuestros objetivos y se descartó el 3 por no ser lo suficientemente fiable.

Este trabajo nos muestra una guía de cómo se podría utilizar la metodología CRISP-DM a proyectos similares, y es precisamente la metodología que se utilizará para el desarrollo de la presente tesis en cada una de sus fases.

E. Zambrano C & Rojas D & Carvajal K & Acuña G (2011) *Análisis de rendimiento académico estudiantil usando data warehouse y redes neuronales*, Universidad de Atacama, Arica – Chile

En este trabajo se ha implementado un Data Wherehouse (DW) en base a información obtenida de un sistema de base de datos no relacional (basado en archivos o también llamado sistema heredado). El DW se ha diseñado para el análisis del comportamiento de aprobación y avance en una malla curricular con datos reales de los currículos de los estudiantes de la carrera de Ingeniería Civil en Computación e Informática de la Universidad de Atacama. El DW no está enfocado sólo en el análisis de comportamientos históricos de los

estudiantes, sino que también ha sido pensado como una arquitectura base para la predicción de tendencias futuras a través de técnicas de RNA.

Principales conclusiones:

- Se ha realizado la implementación de un Data Warehouse y la implementación de una arquitectura de Red Neuronal Artificial para el análisis y la predicción de rendimiento académico de los estudiantes de Ingeniería Civil en Computación e Informática de la Universidad de Atacama.
- La principal ventaja en la utilización de un DW radica en la posibilidad de cruzar distintas dimensiones de análisis de forma simple y rápida, con tal de realizar un análisis exploratorio de los datos para la creación de reportes. Se puede destacar que el proceso de extracción, transformación y carga (ETL) es el que más tiempo y recursos demandó, debido principalmente a que la información debe ser cruzada desde distintas fuentes.
- Es preciso agregar que la utilización de un modelo conceptual multidimensional para generar el esquema conceptual del DW se convierte en una gran herramienta que, independiente de las plataformas, permite acotar el dominio de análisis y dar claridad al proceso posterior.

Los autores destacan que el proceso de extracción, transformación y carga (ETL), son las tareas más importantes y las que mayores recursos demanda en función al tiempo, esta afirmación es totalmente verdadera, puesto que se necesita tener la información previa al procesamiento con los algoritmos de aprendizaje automático, los cuales deberán estar totalmente validados.

2.2.2. Antecedentes Nacionales

A. Camborda Z, Gabriela M (2014). *Aplicación de árboles de decisión para la predicción del rendimiento académico de los estudiantes de los primeros ciclos de la carrera de Ingeniería Civil de la Universidad Continental*, Huancayo, Perú.

Es una tesis desarrollada para obtener el grado de Magister en Ingeniería de Sistemas, refiere a que en la carrera de Ingeniería Civil existe muy bajo rendimiento académico debido a varios factores demográficos, académicos, institucionales actitudinales los cuales afectan al estudiante en su desenvolvimiento académico en la universidad, el objetivo de la investigación fue predecir el rendimiento académico identificando las variables de los factores que más influyen en el estudiante en su rendimiento académico, utilizando para dicho propósito árboles de decisión, específicamente el algoritmo J48 de WEKA.

Las conclusiones descritas son:

- El desarrollo del trabajo permitió validar que la técnica de árboles de decisión aplicados a los factores demográficos, socioculturales, académicos, institucionales predicen el rendimiento académico de los estudiantes de los primeros ciclos de la carrera de Ingeniería civil de la universidad continental.
- Hemos podido aprender a utilizar una herramienta muy potente y con licencia OpenGPL para la minería de datos llamada WEKA. A través de esta potente aplicación, se han podido analizar y aplicar de manera práctica y constructiva uno de sus algoritmos implementados. Se ha analizado los

resultados y discernido entre atributos influyentes y no influyentes, se han filtrado instancias de los datos obtenidos y se han creado nuevos modelos de minería de datos para nuestro conjunto de datos de muestra.

- Las variables académicas fueron las que resultaron con más ganancia de información por lo que con estas variables la predicción tiene mayor exactitud y definen el rendimiento académico del estudiante.
- En términos de minería de datos, el árbol de decisión al estar basado en algoritmos de extracción en reglas de asociación, no solo que es eficiente, sino que también es escalable. Esta es una característica esencial en la resolución de problemas en Minería de datos ya que es de vital importancia el conocimiento de técnicas que permitan anticiparse y predecir los posibles resultados de las decisiones a tomar, "apuntando" siempre a tener mayores posibilidades de éxito y una adecuada gestión administrativa.
- Se encuentra mejores resultados sobre todo en el nivel de concordancia cuando la cantidad de intervalos en las salidas del árbol de decisión son menores. Se pudo obtener mejores resultados cuando se utilizó el aprobado y desaprobado a cambio de colocar varios intervalos de notas para su rendimiento académico

El autor recomienda también profundizar e investigar mucho más las bondades de WEKA que es una herramienta muy potente y con licencia OpenGPL para la minería de datos, además, el trabajo en mención describe y concluye que con los Árboles de decisión se predice el rendimiento académico con una exactitud mayor al 80% para el caso de estudio.

B. Acosta de la Cruz, P & Pizarro, P (2011), *Predicción del rendimiento académico en la educación superior usando minería de datos y su comparación con técnicas estadísticas,* Lima, Perú.

El trabajo y como lo describen los autores pretende elaborar modelos predictivos usando las técnicas de Redes Neuronales de retro propagación, la Regresión Logística y la Regresión múltiple, afirman también que la predicción del rendimiento académico consiste en predecir la nota que obtendrá en el curso o predecir si saldrá aprobado en el curso.

La tesis busca predecir el rendimiento académico que obtendrá un estudiante universitario (considerando si aprobará o no, así como la nota que sacará) en cada uno de los cursos en los que se inscribe, utilizando tres técnicas de predicción: redes neuronales de retro propagación, regresión logística y regresión múltiple para los estudiantes de la especialidad de Ingeniería Química de la Facultad de Ingeniería Química y Textil de la UNI (Universidad Nacional de Ingeniería).

Entre las conclusiones más importantes que describen los autores están:

- Que la aplicación de las técnicas de redes neuronales de retro propagación y de regresión logística para la predicción de la aprobación o no de un curso, arrojan promedios de porcentajes de aciertos similares, de 70.45 %y 70.39% para los modelos, y de 72.83% y 74.04% para los pronósticos, respectivamente.
- No se requiere de herramientas sofisticadas para la aplicación del modelo de redes neuronales de retro propagación y que es suficiente la utilización del Excel de Microsoft con su complemento Solver para la implementación de la red neuronal con diferentes números de capas y neuronas por capas.

Los datos que utilizan para el desarrollo de la tesis son la base histórica académica de los estudiantes de la especialidad de Ingeniería Química de la Universidad Nacional de Ingeniería desde 1993 hasta el 2010, es decir de 7 años. La información suministrada por estos modelos predictivos, según los autores permitirán al estudiante tomar en consideración para decidir su inscripción a las asignaturas y que le ayudarán a avanzar su carrera en forma satisfactoria.

La regresión Logística en el trabajo tiene un mejor pronóstico, respecto de las redes neuronales y también menciona que la utilización de una herramienta como Excel sería suficiente para la implementación de la red neuronal.

C. Gonzales, C & Rodríguez, C (2017) *Propuesta de un Modelo de Business Intelligence para identificar el perfil de deserción estudiantil en la Universidad Científica del Sur*, Lima Perú

Este trabajo fue desarrollado para optar el grado de Magister en Dirección de Sistemas y Tecnologías de la Información.

Entre las conclusiones del proyecto están:

- A través de las herramientas de Minería de Datos, se puede crear modelos predictivos que ayuden al descenso de la tasa de deserción, que podría incrementar la recaudación
- Se ha identificado los factores de riesgo como son: Carrera, Semestre Cursado, Materias cursadas, Modalidad de Ingreso, Materias perdidas, Horario o Programación, Promedio, Edad, Ciudad de procedencia,

Domicilio, Sexo, Estado Civil, Nivel de estudios del Padre, Nivel de estudios de la Madre, Ingresos, entre otros

- Sobre el modelo predictivo desarrollado en WEKA, se concluyó que existe la posibilidad de deserción de la carrera de Administración de Negocios Internacionales, cuando los cursos son de Turno Noche. En el caso de Artes Escénicas y Teatro, la probabilidad de deserción se da en los alumnos que se encuentran matriculados en 1 curso, y de turno mañana. En el caso de la carrera de Ingeniería de Sistemas, existe la probabilidad de deserción cuando se matriculan en 1 curso en las tardes. Existe la probabilidad de deserción en la carrera de Medicina Humana, cuando se encuentra matriculado en 3 o menos cursos, y se da en el Turno Tarde.

El autor hace referencia a la necesidad de utilizar la inteligencia de negocios para identificar los perfiles de deserción, de manera que se brinden recomendaciones que permitan ejecutar acciones enfocadas para lograr su disminución, mejorando la eficiencia y rentabilidad de la empresa. Así mismo afirma que la aplicación de técnicas de minería de datos, permite mejorar la eficiencia y rentabilidad de la institución y que, con la información generada, las diferentes áreas podrían diseñar desde nuevas estrategias de captación del postulante, campañas para atacar la morosidad de los alumnos, mejoras en los mecanismos de selección de los postulantes, hasta desarrollar políticas de retención del alumno

Los modelos predictivos de este trabajo fueron desarrollados en WEKA, así mismo el análisis de los modelos fueron desarrollados en RapidMiner Studio.

D. Menacho C & Higinio C (2017) *Predicción del rendimiento académico aplicando técnicas de minería de datos, Universidad Nacional Agraria La Molina,* Lima – Perú, los datos utilizados para el trabajo corresponden a los registros académicos de la Oficina de Estudios de la UNALM, para una muestra de 914 estudiantes matriculados en los ciclos 2013 II y 2014 I en el curso de Estadística General.

Describe las conclusiones de la siguiente manera:

- Las técnicas de minería de datos demuestran ser herramientas eficaces para obtener modelos que permitan predecir el resultado de los estudiantes matriculados en el curso de Estadística General.
- La técnica de la red Naive de Bayes resultó con una la mayor precisión, al obtener un 71,0% de correcta clasificación.
- Así mismo, se aprecia que, en las cuatro técnicas (regresión logística, redes neuronales, redes bayesianas y árboles de decisión) respecto a la precisión de cada clase, resultó con mayores porcentajes de correcta clasificación para la clase Aprobado y menores para la clase Desaprobado.
- Las variables que influyen en el resultado del curso de Estadística General, fueron el promedio ponderado, situación del curso, nota en diferencial y número de veces que llevó diferencial.
- Se recomienda aplicar las Técnicas de minería de datos con información socio económica, de los estudiantes a fin de mejorar el modelo predictivo.

Las variables consideradas en el estudio fueron: situación del curso, sexo, promedio ponderado, situación académica, nro. de veces que llevo matemáticas, créditos cursados, resultado final entre los más importantes,

utilizando para la predicción los métodos de árboles de decisión, regresión logística binaria, redes bayesianas, y para los resultados y discusión de las técnicas de minería de datos (TMD) utilizaron la herramienta WEKA.

Capítulo 3

HIPÓTESIS Y VARIABLES

3.1. Hipótesis

a.- Hipótesis General

Los algoritmos de aprendizaje automático determinan la predicción del rendimiento académico de los estudiantes de la UNSAAC del primer semestre, a partir de sus datos de ingreso con una eficiencia de 70%

b.- Hipótesis Específicas

H1: La predicción del rendimiento académico de los estudiantes de la UNSAAC a partir de sus datos de ingreso dependen de los factores sociodemográficos y socioeducativas.

H2: El algoritmo con mejor performance de predicción del rendimiento académico de los estudiantes de la UNSAAC en su primer semestre es Random Forest.

3.2. Identificación de Variables e Indicadores

VARIABLES DE ESTUDIO

- ✓ **Variables socio demográficas**
  - Edad
  - Genero
  - Lugar de procedencia
  - Trabaja
  - Forma de elección de la E.Profesional

- Tipo de preparación

✓ **Variables socio educativas**

- Nota de ingreso
- Modalidad de Ingreso
- Financiamiento de estudios universitarios
- Tipo de preparación para el ingreso
- Escuela Profesional

✓ **Rendimiento académico**, Obtenido con los algoritmos de aprendizaje automático expresado como: Bueno o Malo (Porcentaje de predicción acertada)

3.3. Operacionalización de Variables

Tabla 2. *Operacionalización de las variables de estudio*

Variables de Estudio	Definición Conceptual	Dimensiones	Indicadores
Rendimiento académico	Gonzales, R. (1989), El rendimiento académico de los estudiantes depende estrictamente de la Universidad donde estudia, de sus docentes y sobre todo de las capacidades de los alumnos. Es un concepto que se utiliza en los ámbitos educativos de todos los niveles para referirse a la evaluación del conocimiento adquirido por parte de los alumnos, expresado en los resultados de sus evaluaciones	Socio Demográficas	Edad
			Genero
			Lugar de Procedencia
			Trabaja
			Forma de Elección de E. Profesional
			Tipo de Preparación
		Socio Educativas	Nota de Ingreso
			Modalidad de Ingreso
			Financiamiento de estudios universitarios
			Tipo de preparación para el ingreso
			Cursos Matriculados
			Escuela Profesional
Algoritmos de aprendizaje automático	Kelleher, J. & Mac, B. & D'Arcy, A. (2017), El aprendizaje automático se define como un proceso automatizado que extrae patrones de los datos para la construcción de modelos de análisis de datos y para la predicción utilizamos algoritmos supervisados.	Algoritmos de Aprendizaje automático	Porcentaje de Predicción Acertada

Fuente: Elaboración Propia

Capítulo 4

METODOLOGÍA

4.1. Ámbito de Estudio

El ámbito de estudio de la presente investigación está enmarcada a la Universidad Nacional de San Antonio Abad del Cusco, utilizando la información de los alumnos ingresantes a partir del semestre 2014-I hasta el 2018-I. Y el modelo de predicción está orientado a los alumnos del primer semestre de todas las Escuelas Profesionales

4.2. Tipo y Nivel de Investigación

Es una investigación cuantitativa, del tipo correlacional y no experimental.

Los estudios correlacionales asocian variables mediante un patrón predecible para un grupo o población. Este tipo de estudios tiene como finalidad conocer la relación o grado de asociación que exista entre dos o más conceptos, categorías o variables en un contexto en particular, (Hernández, 2015) En estos estudios se ofrecen predicciones y explican las relaciones entre las variables, y es precisamente lo que desarrollaremos “predecir el rendimiento académico” estas investigaciones pretenden responder a preguntas de investigación como: ¿Los algoritmos de aprendizaje automático pueden predecir el rendimiento académico de los estudiantes de la UNSAAC a partir de sus datos de ingreso?.

4.3. UNIDAD DE ANÁLISIS

Los sujetos de estudio de los cuales se pretende determinar su rendimiento académico, son los estudiantes de la UNSAAC, considerándose para cada uno de estos sus datos de ingreso o de admisión a la Universidad, expresados en factores socio demográficos y socio educativas.

4.4. POBLACIÓN DE ESTUDIO

La población de estudio es el conjunto de los elementos a ser evaluados y para la presente investigación son los estudiantes ingresantes a la UNSAAC desde el semestre 2014-I hasta el 2018-I

4.5. TAMAÑO DE MUESTRA

Tabla 3. *Cantidad de ingresantes por Semestre*

Semestre	**Ingresantes**
2014-1	1462
2014-2	955
2015-1	1505
2015-2	885
2016-1	1751
2016-2	1140
2017-1	1849
2017-2	1310
2018-1	1841
TOTAL	**12698**

Fuente: elaboración propia

La muestra es de 12,698 alumnos ingresantes a la UNSAAC por las diferentes modalidades, y estará compuesta por la totalidad de la población.

4.6. Técnicas de Selección de Muestra

Para realizar la selección de la muestra se utilizó la información proporcionada por la Unidad de Centro de Cómputo de la Universidad conteniendo:

1. Encuestas a los postulantes. - Esta información es obtenida a través de la información que ingresan los postulantes vía web, el formulario en mención se encuentra en el ANEXO Nro. 01.

2. Información de los estudiantes ingresantes con sus promedios ponderados en el primer semestre. Esta información se encuentra en archivos de Excel descritos de la siguiente manera:

Tabla 4. *Información proporcionada por el Centro de Computo, de estudiantes ingresantes*

Denominación	**Ejemplo del dato**
Semestre de Ingreso	2014-1, 2014-2, ….2018-1
Escuela a la que ingreso	Agronomía, Ing. Civil, Enfermería, y las 42 Escuelas profesionales.
Código de Alumno	En formato encriptado por temas éticos.
Genero	1: Masculino; 0: Femenino
Ubigeo del Colegio	7-12-1 : (Dpto., Provincia, Distrito)
Modalidad de Ingreso	Admisión ordinario, primera oportunidad, CEPRU.
Posición de Ingreso	Puesto en el que ingreso
Nota	Vigesimal (0 a 20)

Fecha de Nacimiento	1991-01-04, a partir del cual se obtendrá la edad.
Créditos Matriculados	Numérico
Créditos Aprobados	Numérico
Promedio Ponderado	Nota vigesimal, es el promedio ponderado de las asignaturas desarrolladas en el primer semestre

Fuente: Elaboración propia

La información proporcionada por la Unidad del Centro de Computo de la UNSAAC, han sido utilizados en esta investigación con ética y responsabilidad, cuidando la privacidad respecto de la información personal, considerando que el uso de la información que permita identificar individualmente a las personas requiere del consentimiento y la autorización previa para su uso, en tal sentido se eliminaron atributos que pudieran identificarlos, asignándose valores numéricos correlativos, de esta manera se evita exponer la información personal.

4.7. Técnicas de Recolección de Datos e Información

La información fue extraída de las bases de datos del Centro de Computo de la UNSAAC, considerando en esta información, las respuestas a las encuestas que realizan a los postulantes, los datos de los ingresantes y su rendimiento académico en el primer semestre, los cuales fueron posteriormente normalizados, considerando la teoría descrita en las bases teóricas para su posterior utilización, es decir se hizo la preparación y depuración de la data.

4.8. Técnicas de Análisis e Interpretación de la Información

Para el análisis e interpretación de la información extraída del Centro de Computo respecto de los alumnos ingresantes, en primer lugar se hizo la revisión de la información generada, en cada una de las 06 preguntas, adicionando o eliminando aquellos registros que mostraban inconsistencias, valores nulos o incompletos, en segundo lugar se procedió a la transformación de los datos para ser procesados, adicionando columnas para el promedio ponderado y la condición(“Bueno” o “Malo”), en tercer lugar, se hizo el análisis, comprensión y la explotación de los resultados, para finalmente ser interpretados con gráficas y matrices, se utilizaron herramientas como Microsoft SQL Server 2014, Power Pivot, Power Query, de Microsoft Excel, así como con la herramienta de WEKA 3.8.

Capítulo 5

RESULTADOS Y DISCUSIÓN

5.1. Procesamiento, Análisis, Interpretación y Discusión de Resultados

Para lograr el procesamiento, análisis e interpretación de la data proporcionada por el Centro de Cómputo, utilizaremos la metodología CRISP-DM (*Cross Industry Standard Process for Data Mining*), descrito en el marco teórico, en sus diferentes fases, que se describen a continuación:

5.1.1. FASE I: Comprensión del Negocio

Que a su vez comprende las siguientes etapas:

5.1.1.1. Comprensión del contexto y determinación de objetivos

Esta fase se enfoca en la comprensión del contexto y de los objetivos del proyecto, para luego convertir este conocimiento de los datos en la definición de un problema y en un plan preliminar para alcanzar los objetivos, en este sentido describiremos lo siguiente:

Contexto de la Universidad y del proyecto

Las universidades como instituciones educativas tienen como propósito fundamental garantizar la calidad de la educación de sus estudiantes en sus diferentes programas, los cuales generalmente están expresados en las notas que estos obtienen, pero muchas veces en el intento de lograr lo anterior no se logra, por falta de herramientas que permitan monitorear o apoyar al rendimiento académico de los principales actores de las Universidades (los estudiantes).

Objetivos del proyecto

Con el presente trabajo buscaremos predecir el rendimiento académico de los estudiantes del primer semestre a partir de sus datos de ingreso o admisión a la UNSAAC, para que las autoridades correspondientes puedan tomar las acciones correspondientes para mejorar el éxito de los estudiantes en su rendimiento académico y evitar o eliminar el fracaso de estos.

Para determinar un modelo que permita predecir el rendimiento académico de los estudiantes del primer semestre, se obtuvo la información de los estudiantes desde el semestre 2014-I hasta el semestre 2018-I, proporcionado por la Unidad de Centro de Computo de la UNSAAC, esta información en su mayoría ha sido procesada a partir de la encuesta que realizan los postulantes a la UNSAAC vía web, como se describe en el ANEXO 01 del presente documento.

5.1.1.2. Evaluación de la situación

Desde una perspectiva institucional

Considerando que la UNSAAC, es una entidad de formación de profesional de alto nivel académico y los procesos de licenciamiento y acreditación de las Escuelas profesionales, sería importante contar con una herramienta que les permita a las autoridades competentes, tomar decisiones respecto de los estudiantes que podrían fracasar en su rendimiento académico en los primeros semestres de estancia en la institución.

Desde una perspectiva del aprendizaje automático y minería de datos

Como se describió en el marco teórico, en el aprendizaje automático se clasificaron las técnicas de algoritmos supervisados y no supervisados, y dentro de los supervisados a su vez las técnicas de regresión y de clasificación, es precisamente en las técnicas de clasificación donde utilizamos los algoritmos para determinar el de mejor desempeño, la estadística descriptiva también nos ayuda a la valoración de los resultados encontrados en la etapa inicial.

Y respecto de la metodología utilizada para los resultados del presente se utilizó CRISP-DM, por la flexibilidad que ofrece nos ayudara a trabajar de manera ordenada.

5.1.1.3. Determinación de los Objetivos de Minería de Datos aplicados al proyecto

Los objetivos de la minería de datos aplicados al proyecto de aprendizaje automático de la presente tesis, se resumen en los siguientes pasos:

- Realizar la limpieza de los datos proporcionada por el centro de computo
- Determinar los factores más importantes que afectan al rendimiento académico
- Analizar y determinar el mejor algoritmo de aprendizaje automático para desarrollar la predicción
- Predecir el rendimiento académico con el algoritmo más óptimo de predicción.

El criterio de la predicción se basa primero en el análisis estadístico de los datos para determinar los factores que influyen en el rendimiento académico y segundo la utilización de la herramienta WEKA para determinar la performance de los algoritmos y la selección del mejor predictor para realizar la predicción.

5.1.1.4. Desarrollo del plan del proyecto

El tiempo estimado para la realización del presente proyecto en su etapa de generación y evaluación de resultados de la predicción utilizando la metodología CRISP-DM se muestran en la siguiente tabla:

Tabla 5. *Plan del proyecto utilizando la metodología CRISP-DM*

Fase	**Tiempo**	**Recursos**	**Riesgos**
I.-Comprensión del negocio	15 días	Entrevistas con encargados de Centro de Computo	Ninguno
II.-Comprensión de los datos	15 días	Microsoft SQL Server Rapid Mining Microsoft EXCEL	Datos incoherentes e incompletos
III.-Preparación de los datos	15 días	R-Studio, Python Excel, WEKA 3.8	Problemas de atípicos en los datos
IV.-Modelado	1 mes	WEKA 3.8	No encontrar un buen modelo
V.-Evaluación	15 días	WEKA 3.8	No llegar al porcentaje de predicción esperada
VI.-Despliegue	15 días	Java	Ninguno

Fuente: Elaboración propia

5.1.2. FASE II: Comprensión de los Datos

La fase de entendimiento de datos comienza con la colección de datos inicial y continúa con actividades que permiten familiarizarse con los datos, identificar los problemas de calidad, descubrir conocimiento preliminar sobre los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.

En la labor de investigación de antecedentes al presente, se pudo evidenciar que la mayoría de autores coinciden que existen 3 factores que afectan al rendimiento académico de los estudiantes universitarios y en concreto son los siguientes:

- El estudiante
- La institución
- Los docentes

Y respecto a los estudiantes es un factor aún más complejo por que intervienen aspectos psicológicos y emocionales así como también la familia y otros aspectos inherentes a la persona; estos no serán estudiados en este trabajo de investigación, solo se considerarán factores de sus datos de ingreso a la UNSAAC, y sobre las que se realizaron actividades de conocimiento preliminar de la data, para mejorar la consistencia de los mismos (inserción, modificación y eliminación) de datos, utilizando herramientas como SQL – SERVER 2018 y Microsoft Excel 2017.

5.1.2.1. Recopilación de datos iniciales

La información procesada sobre los datos iniciales en bruto, contiene la siguiente información:

- Semestre
- Nombre de la escuela Profesional

- Fecha de Nacimiento
- Sexo
- Procedencia
- Tipo de Colegio
- Modalidad de ingreso
- Nota de Ingreso
- Posición de Ingreso
- Colegio de Procedencia
- Tipo de Preparación
- Forma de elección de la Escuela Profesional
- Trabaja
- Conocimientos de computación
- Financiamiento de los estudios
- Créditos matriculados
- Créditos aprobados
- Cursos matriculados
- Cursos desaprobados
- Promedio ponderado

Y la predicción estará en relación a la:

- **Condición** (Bueno o Malo)

La principal fuente de extracción de la información en bruto se obtuvo de las siguientes tablas:

Tabla 6. *Encuesta a postulantes: ¿Trabajas?*

Pregunta	**Opción**	**Texto**
¿Trabajas?	1	No
	2	Sí, en forma permanente
	3	Sí, en forma eventual

Fuente: Elaboración propia

Tabla 7. *Encuesta a postulantes: ¿Quién financia tus estudios?*

Pregunta	**Opción**	**Texto**
¿Quién financia tus estudios?	1	Padre
	2	Madre
	3	Ambos
	4	Yo mismo
	5	Hermanos
	6	Otros parientes
	7	Otras personas o Instituciones

Fuente: Elaboración propia

Tabla 8.- *Encuesta a postulantes: ¿Cómo elegiste tu carrera?*

Pregunta	**Opción**	**Texto**
¿Cómo elegiste tu carrera profesional?	1	Por orientación vocacional
	2	Por las posibilidades de trabajo
	3	Por influencia familiar
	4	Pensando en mis aptitudes
	5	Por el costo de la profesión
	6	Otro

Fuente: Elaboración propia

Tabla 9. *Encuesta a postulantes: ¿Que conocimientos de computación tienes?*

Pregunta	**Opción**	**Texto**
¿A través de qué medio de difusión se enteró del Examen de Admisión?.	1	TV.
	2	Radio.
	3	Página "Web UNSAAC"
	4	Periódico.
	5	Otro.

Fuente: Elaboración propia

Tabla 10. *Encuesta a postulantes: ¿Que conocimientos de computación tienes?*

Pregunta	**Opción**	**Texto**
¿Qué conocimientos en computación tienes?	1	Word.
	2	Excel
	3	Power Point
	4	Todos
	5	Ninguno

Fuente: Elaboración propia

Tabla 11. *Encuesta a postulantes: ¿Qué te incentiva a postular a la UNSAAC?*

Pregunta	**Opción**	**Texto**
¿Qué te incentiva postular a la UNSAAC?	1	Por la gratuidad de la enseñanza
	2	Por su prestigio
	3	Por diversidad de carreras que ofrece
	4	Por presión familiar
	5	Otros

Fuente: Elaboración propia

La cual a su vez fue extraída a partir de la página Web de admisión de la UNSAAC, y se muestra en el ANEXO 02.

5.1.2.2. Descripción de los datos

- **Semestre**. - Expresado a partir del semestre 2014-I hasta el 2018-I
- **Nombre de la escuela Profesional**. - La tesis está en función de los ingresantes a las 42 Escuelas profesionales de la UNSAAC (Ejemplo: Contabilidad, Administración, Ing. Informática y de sistemas, Medicina)
- **Fecha de nacimiento**. - Este atributo nos permite obtener la edad de los ingresantes a la Universidad (01/05/1999)
- **Sexo**. - Expresado como Masculino (1) o Femenino (0)
- **Procedencia**. - Es el distrito, provincia o departamento del cual proviene el ingresante a la universidad (Ejemplo: Wanchaq-Cusco-Cusco)
- **Tipo de Colegio**. - Expresado como particular o nacional
- **Modalidad de ingreso**. - LA forma de como ingreso, es decir: admisión ordinario, CEPRU, primera oportunidad u otro.
- **Nota de Ingreso**. - Nota vigesimal de ingreso a la universidad (Ejemplo: 14.56)
- **Posición de Ingreso**. - Es el lugar o el puesto en el que ingreso.
- **Colegio de Procedencia**. - Nombre del colegio del cual proviene el ingresante
- **Tipo de Preparación**. - Nombre de la academia, CEPRU de la UNSAAC, profesor particular, solo u otro
- **Forma de elección de la Escuela Profesional**. - Ingresante a la opción deseada o ingresante a la segunda opción elegida por el postulante
- **Trabaja**. - Si, No, Eventualmente, Permanentemente
- **Conocimientos de computación**. - Conocimientos de las plataformas de ofimática

- **Financiamiento de los estudios**. - Quien financia los estudios del ingresante (Ejemplo: papá, mamá, ambos)
- **Créditos matriculados**. - Cantidad de créditos matriculados en el rango de 3 hasta 46 créditos
- **Créditos aprobados**. - Que puede ser desde 3 hasta 46 Créditos
- **Cursos matriculados**. - Cantidad de cursos matriculados que pueden estar entre 1 y 12 cursos
- **Cursos desaprobados**. - Que puede ser desde 1 hasta 12 Créditos
- **Promedio ponderado**. - Es el promedio ponderado de los cursos matriculados y la cantidad de créditos que tiene la asignatura

5.1.2.3. Exploración de los datos

Para la exploración de los datos utilizaremos estadísticos en base al cuestionario de ingreso a la UNSAAC, que cuenta con 7 preguntas de las cuales 5 preguntas se describen a continuación, las otras 2 preguntas, se consideran no relevantes para el presente trabajo.

Pregunta 01. *¿Trabajas?*

Tabla 12. *Estadísticas: Pregunta: ¿Trabajas?*

Opción	Texto	Cantidad	Porcentaje
0	No respondió la encuesta	1400	11,03%
1	No	8927	70,30%
2	Sí, en forma permanente	303	2,39%
3	Sí, en forma eventual	2068	16,29%

Fuente: Elaboración propia

- Como se puede apreciar un 70% de los estudiantes ingresantes no trabajan, y un 16% lo realiza de manera eventual.

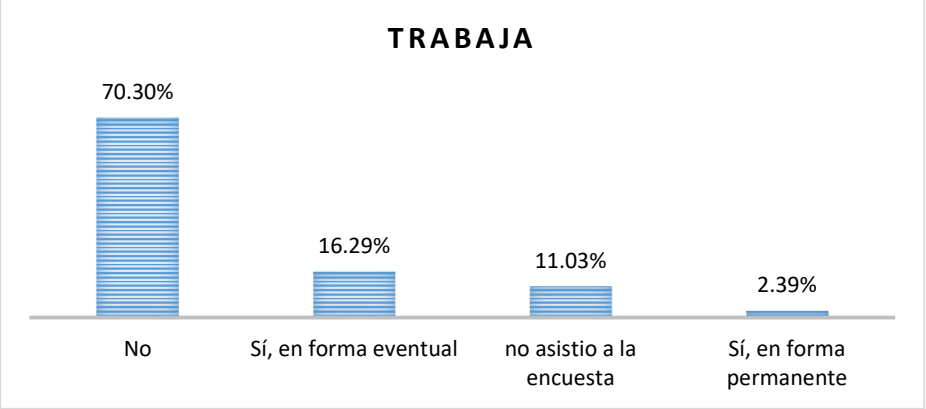

Gráfico 1.- Pregunta: ¿Trabajas?

Fuente: Elaboración propia

Pregunta 02. *¿Quién financia tus estudios?*

Tabla 13.- Estadísticas: ¿Quién financia tus estudios?

Opción	Texto	Cantidad	Porcentaje
0	No respondió la encuesta	1400	11,03%
1	Padre	1376	10.84%
2	Madre	1380	10,87%
3	Ambos	7448	58,65%
4	Yo mismo	765	6,02%
5	Hermanos	214	1,69%
6	Otros parientes	93	0,73%
7	Otras personas o Instituciones	22	0,17%

Fuente: Elaboración propia

- En 58% de los ingresantes, sus estudios son financiado por los dos padres.

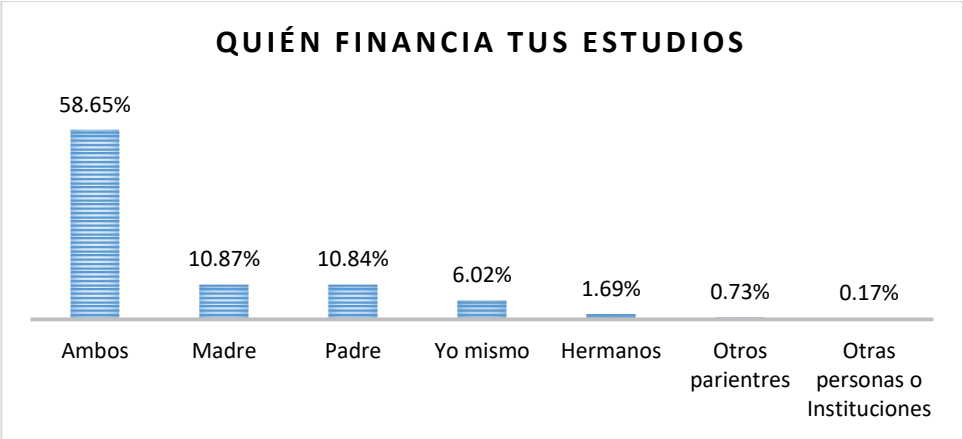

Gráfico 2.- Pregunta: ¿Quién financia tus estudios?

Fuente: Elaboración propia

Pregunta 03.- ¿Qué tipo de preparación recibiste para ingresar a la Universidad?

Tabla 14. *Estadísticas: ¿Qué tipo de Preparación recibiste para ingresar a la Universidad?*

Opción	**Texto**	**Cantidad**	**Porcentaje**
0	No respondió la encuesta	1400	11,03%
1	Por mi cuenta	1819	14,33%
2	CEPRU-UNSAAC	2640	20,79%
3	Profesor particular	51	0,40%
4	Sólo academias	5157	40,61%
5	CEPRU y Academia	1631	12,84%

Fuente: Elaboración propia

Para el ingreso a la universidad el 40 % recibió capacitación en las academias de nuestro medio, un 20% lo hicieron a través del Centro Pre Universitario (CEPRU), hay un porcentaje importante de 14% que su preparación fue por su propia cuenta, consideramos por falta de aspectos económicos.

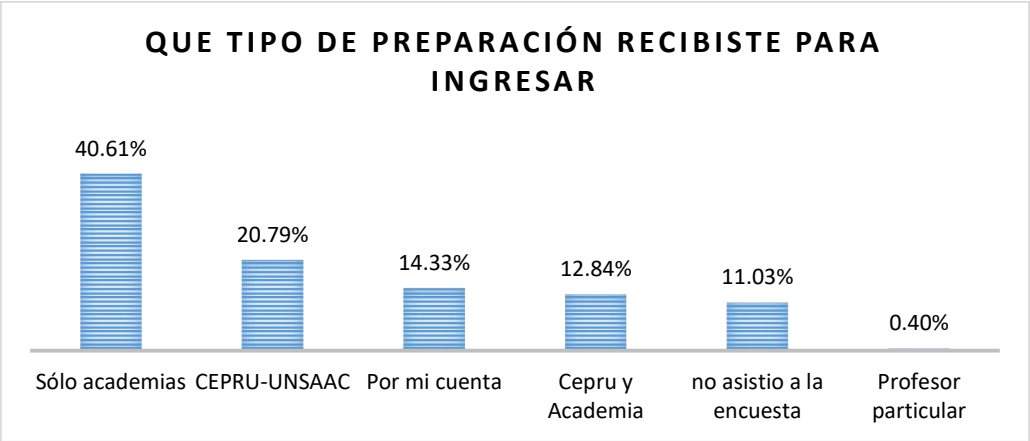

Gráfico 3.- Pregunta: ¿Qué tipo de Preparación recibiste para ingresar a la Universidad?

Fuente: Elaboración propia

Pregunta 04.- ¿Cómo elegiste tu Escuela profesional?

Tabla 15. *Estadísticas: ¿Cómo elegiste tu escuela profesional?*

Opción	**Texto**	**Cantidad**	**Porcentaje**
0	No respondió la encuesta	1400	11,03%
1	Por orientación vocacional	5732	45,14%
2	Por las posibilidades de trabajo	662	5,21%
3	Por influencia familiar	362	2,85%
4	Pensando en mis aptitudes	3765	29,65%
5	Por el costo de la profesión	46	0,36%
6	Otro	731	5,76%

Fuente: Elaboración propia

El 45% de los encuestados escogió su escuela profesional por orientación vocacional y un 29% pensado en sus aptitudes.

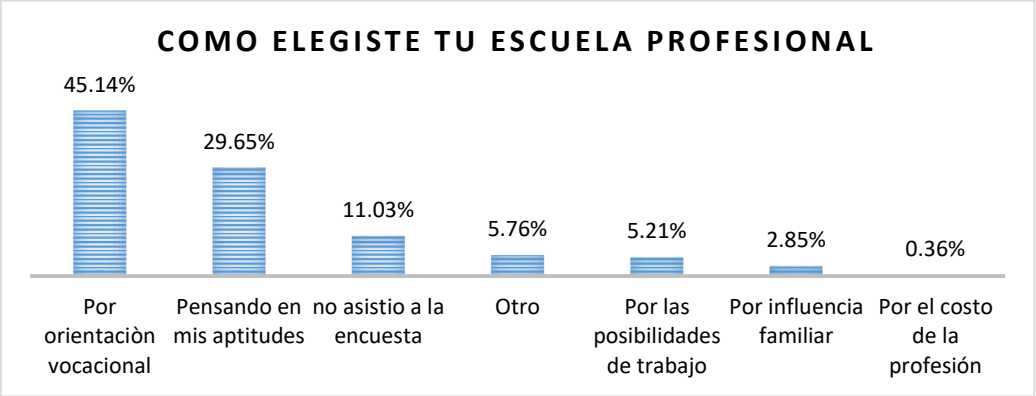

Gráfico 4.- Pregunta: ¿Cómo elegiste tu carrera profesional?

Fuente: Elaboración propia

***Pregunta 05.- ¿A través de qué medio de difusión se enteró del Examen de Admisión*? (No relevante para la presente investigación)**

Pregunta 06.- ¿Qué conocimientos de computación tienes?

Tabla 16. *Estadísticas: ¿Qué conocimientos de computación tienes?*

Opción	Texto	Cantidad	Porcentaje
0	No respondió la encuesta	1400	11,03%
1	Word.	3044	23,98%
2	Excel	265	2,09%
3	Power Point	442	3,48%
4	Todos	6517	51,35%
5	Ninguno	1024	8,07%

Fuente: Elaboración propia

- El 51% de los ingresantes a la UNSAAC tiene los conocimientos de computación básicos.

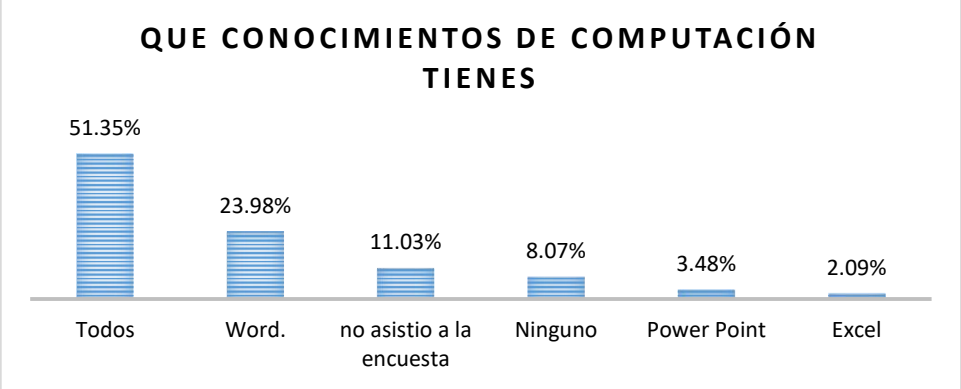

Gráfico 5.- Pregunta: ¿Qué conocimientos de computación tienes?

Fuente: Elaboración propia

Pregunta Nro. 07.- ¿Que te incentiva postular a la UNSAAC?

(No relevante para el presente trabajo de tesis)

Es importante también indicar que un 11% de los ingresantes no completo la encuesta realizada.

Otras estadísticas importantes para la exploración de los datos

- **Ingresantes por genero desde el 2014-I al 2018-I**

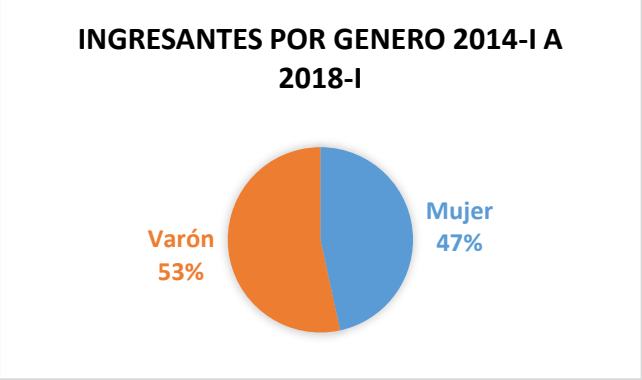

Gráfico 6.- Ingresantes por genero del semestre 2014-1 al 2018-1

Fuente: Elaboración propia

Como se aprecia en el Gráfico 6 existe una equivalencia entre el porcentaje de ingresantes varones y mujeres a la UNSAAC, con una ligera ventaja de los varones.

- **Ingresantes por tipo de colegio desde el 2014-I al 2018-I**

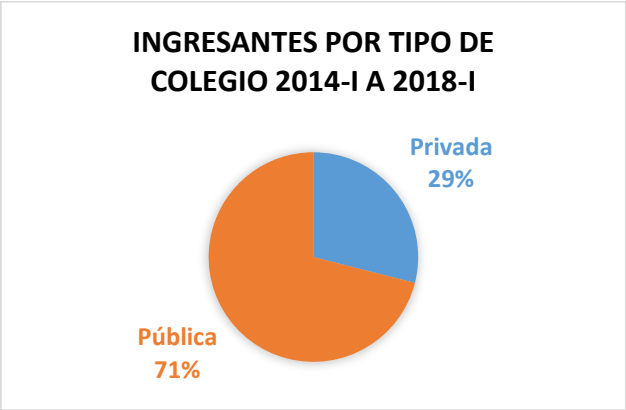

Gráfico 7.-.- Ingresantes por tipo de colegio del semestre 2014-1 al 2018-1

Fuente: Elaboración propia

El mayor porcentaje de estudiantes de la UNSAAC son de colegios públicos, es decir de cada 10 estudiantes de la universidad 7 son de colegios públicos y 3 de colegios privados.

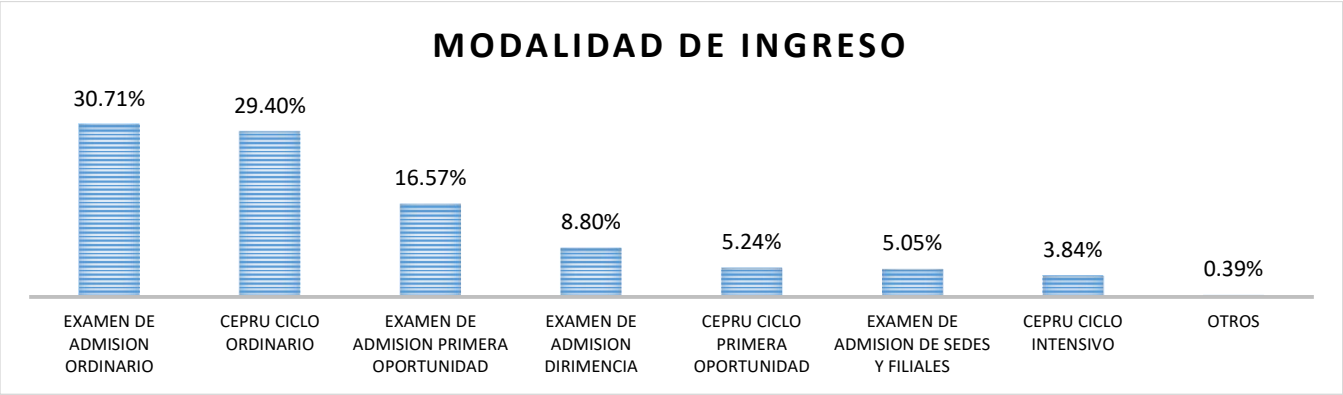

Gráfico 8.- Ingresantes por Modalidad de Ingreso

Las modalidades de examen de admisión ordinario y CEPRU, son las que tienen los mayores porcentajes de admisión a la UNSAAC.

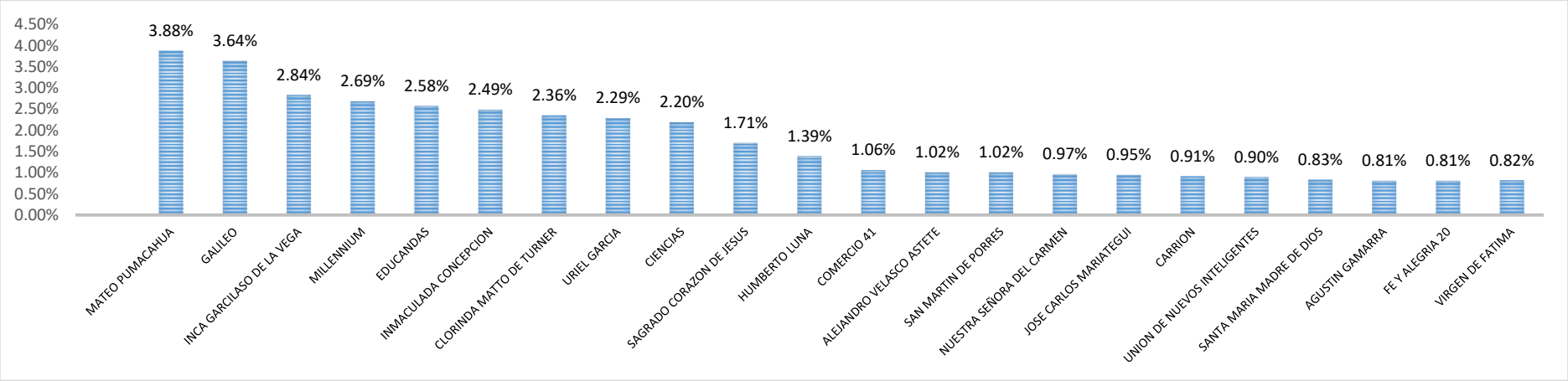

Gráfico 9.- Ingresantes por Colegio de Procedencia

Importante.-El 61.8% de ingresantes a la UNSAAC, corresponden al grupo de Otros Colegios.

5.1.2.4. Verificación de la calidad de los datos

En esta fase se hace un examen de la calidad de los datos, es decir respondernos a las preguntas:

- ✓ **¿Están completos?** La Información proporcionada por la Unidad de Centro de Computo, lamentablemente no cuenta con datos completos, en varios registros de las encuestas en promedio 11% del total, sin embargo, estos datos que están incompletos podría considerarse como una nueva clase, es decir, todos aquellos estudiantes que no ingresaron o no completaron las preguntas de las encuestas podrían pertenecer a un grupo dentro del rendimiento académico.

  Y respecto de los valores incompletos que, en otras instancias con porcentajes mínimos, por ejemplo, en la edad, cantidad de créditos matriculados o nota de ingreso, fueron completados con el promedio.
- ✓ **¿Cubren todos los casos requeridos?** La Información del *dataset*, si cubren todos los casos requeridos, considerando que están completos, en cada uno de las instancias para la presente tesis.
- ✓ **¿Son correctos o contienen errores? Si hay errores, ¿cómo son de frecuentes?** Los errores más frecuentes fueron los encontrados en la cantidad de créditos matriculados, puesto que existían estudiantes matriculados hasta en 46 créditos en diferentes escuelas profesionales.
- ✓ **¿Hay valores omitidos? Si es así, ¿cómo se representan, ¿dónde ocurre esto, y cómo son de frecuentes?** Si hay valores omitidos, pero en porcentajes mínimos, los cuales fueron completados como se indicó anteriormente.

5.1.3. FASE III.- Preparación de los Datos

En esta fase de Preparación, desarrollaremos todas las actividades necesarias para construir el conjunto final de datos (los datos que se utilizarán en las herramientas de modelado) a partir de los datos en bruto iniciales.

Las tareas incluirán la selección de atributos y registros, así como la transformación y la limpieza de datos para las herramientas que permitan generar los modelos predictivos.

5.1.3.1. Selección de los datos más relevantes

Para determinar los datos más o factores más relevantes que influyen en el rendimiento académico de los estudiantes desarrollaremos el análisis estadístico del *dataset*, desarrollando una limpieza preliminar de los mismos, para lo cual utilizaremos herramientas estadísticas. Con los baremos definidos se plantea la comparación con el test de chi cuadrado para tablas de contingencia y coeficiente de correlación de Pearson. La tabla 17 resume la significancia de cada variable implicada en su asociación con el rendimiento académico.

Descriptivos de las variables de estudio:

Tabla 17.*Tabla de contingencia para determinar factores importantes que inciden en el rendimiento académico*

		Chi cuadrado de PEARSON			Coeficiente contingencia	Coeficiente contingencia
	Variable de estudio	**Valor**	**GL**	**P-Valor**	**CRAMER**	**PEARSON**
Condición	**Nota de ingreso**	808.89	2	0.0001	0.18	0.25
	Escuela Profesional	553.75	28	0.0001	0.15	0.21
	Semestre	381.98	8	0.0001	0.12	0.17
	Sexo	195.69	1	0.0001	0.12	0.12
	Modalidad de ingreso	179.63	7	0.0001	0.08	0.12
	Cursos matriculados	84.34	2	0.0001	0.06	0.08
	Forma de ingreso	71.99	1	0.0001	0.05	0.08
	Procedencia	63.89	21	0.0001	0.05	0.07
	Edad	33.31	2	0.0001	0.04	0.05
	Trabaja	21.65	3	0.0001	0.03	0.04
	Tipo de preparación	11.89	3	0.0078	0.02	0.03
	Forma Elección EP	8.67	3	0.0340	0.02	0.03
	Financiamiento de estudios	14.39	3	0.0024	0.02	0.03
	Tipo de colegio	9.92	1	0.0016	0.02	0.03

Fuente: Elaboración propia

De la tabla anterior se puede inferir que si el **P-Valor** es pequeño, indica que existe asociación entre las 2 variables (VARIABLES DE ESTUDIO Y RENDIMIENTO ACADÉMICO), por consiguiente, podemos afirmar que:

- Todas las variables descritas presentan significancia con un **P-Valor** inferior a 0.05 debido a la gran cantidad de datos analizados; sin embargo, la interpretación del P-Valor no es del todo acertada por lo cual se sugiere la interpretación del coeficiente de correlación de Pearson el cual tiene una variación de 0 a 1 donde 0 es ninguna relación y 1 relación perfecta máxima.
- El coeficiente de correlación de Pearson nos sugiere trabajar la predicción con las variables que consiguieron un coeficiente de correlación superior a 0.10 siendo estas: nota de ingreso, escuela profesional, semestre, sexo, modalidad de ingreso.

Descriptivo para la variable dependiente. - Esta variable, se determinó en función a la siguiente tabla, considerando que los datos refieren a 2 agrupaciones por el semestre y respecto de las notas mínimas de aprobación en la UNSAAC

Tabla 18. *Descriptivo para la variable dependiente*

SEMESTRE	**OBSERVACIÓN**	
	Promedio Ponderado	**Condición**
2014-I hasta 2015-II	De 00.00 a 10.99	MALO
	De 11.00 a 20.00	BUENO
2016-I hasta 2018-I	De 00.00 a 13.99	MALO
	De 14.00 a 20.00	BUENO

Fuente: Elaboración propia

Es importante mencionar que, en la UNSAAC hasta el semestre 2015-II, la nota mínima aprobatoria era de 10.5; y a partir del semestre 2016-I, es de 13.5.

5.1.3.2. Limpieza de datos

En esta fase, se describen los factores más importantes que nos permitirán predecir el rendimiento académico, y para poder lograr este objetivo en primer lugar se realizó la eliminación de 89 Registros de la data inicial, porque estos tenían inconsistencias en su información por ejemplo créditos matriculados de hasta 46, valores incoherentes en el género, en los nombres de las escuelas profesionales, en las fechas de nacimiento, en la procedencia y algunos otros en las encuestas, utilizándose las herramientas de WEKA y Microsoft Excel 2016.

Para la limpieza de los datos realizaremos un análisis de cada uno de los atributos de nuestra información y lo clasificaremos considerando 3 apreciaciones de importancia para desarrollar el modelo (Alta, Media y Baja), considerando el ítem anterior (Selección de los datos más relevantes) y que se describe a continuación:

- **Semestre** (**dato ordinal/categórico**)

Importancia : Media

Observación : Es Media considerando que no existen cambios muy drásticos en el performance del rendimiento de los alumnos por semestre.

- **NomEscuela (dato categórico)**

Importancia : Alta

Observación : Es Alta, considerando que el promedio ponderado de cada escuela es diferente una de otra, así como la cantidad de créditos y cursos aprobados o desaprobados, así como el nivel de deserción.

- **Edad (Nominal)**

Importancia : Baja

Observación : Baja, considerando que nuestro rango de edades está entre 16 y 61 años, la edad será normalizada en sus valores a 1, 2 y 3 de manera proporcional, y no incide bastante en el rendimiento académico.

- **Sexo (categórico)**

Importancia : Alta

Observación : Podría ser determinante a para determinar el rendimiento académico, el sexo o genero se normalizará sus valores e irán entre 0 y 1

- **Procedencia (categórico)**

Importancia : Media

Observación : Se considerará la procedencia a Cusco y Provincias

- **Tipo Colegio (categórico)**

Importancia : Media

Observación : Pública o Privada, podría determinar el rendimiento académico de los estudiantes.

- **Modalidad Ingreso (categórico)**

Importancia : Alta

Observación : Como escogieron estudiar la escuela, es decir ingresaron a la que realmente querían ingresar o fue su segunda opción, será determinante en el rendimiento académico.

- **Nota Ingreso (Nominal)**

Importancia : Alta

Observación : Dato importante, se normalizará entre 0 y 1 considerando máx. = 20 y min = 0

- **Posición Ingreso (Nominal)**

Importancia : Alta

Observación : Dato importante, que también será normalizado, considerando que el puesto en el que ingreso el estudiante podría determinar el alto o bajo rendimiento.

- **Colegio Procedencia (dato categórico)**

Importancia : Baja

Observación : Dato podría ser relevante, lo complejo de este dato es que tiene muchas posibles valores.

- **Tipo Preparación (dato categórico)**

Importancia : Media

Observación : Podría ser relevante, el principal problema es que hay 1457 datos con valor NULL, es decir que no completaron la encuesta.

- **FormaDeEleccionEP (dato categórico)**

Importancia : Media

Observación : Se reducirá a las categorías: primera oportunidad, ordinario y CEPRU.

- **Trabaja (dato categórico)**

Importancia : Media

Observación : Podría ser relevante, el principal problema es que hay 1476 datos con valor NULL, que no respondieron a la encuesta

- **Conocimiento Computación (dato categórico)**

Importancia : Baja

Observación : Podría ser relevante, el principal problema es que hay 1564 datos con valor NULL, que no respondieron a la encuesta

- **Financiamiento Estudios (dato categórico)**

Importancia : Baja

Observación : Podría ser relevante, el principal problema es que hay 1452 datos con valor NULL, que no respondieron a la encuesta.

5.1.3.3. Construcción de nuevos datos

Para la construcción de los nuevos datos se realizó la normalización y discretización de los mismos, es decir los datos se convirtieron a intervalos numéricos a continuación se muestra un fragmento de la construcción de los nuevos datos, la construcción total de los nuevos datos se muestra en el Anexo Nro. 04:

- Id : Numeric
- Nota de Ingreso : {0,1,2}

  Donde:

  0: Si : 0<=Nota<11.5

  1: Si : 11.5<=Nota<15

  2: Si : 15<=Nota
- Edad : {0,1,2}

  Donde:

  0: Si Edad<19

  1: Si 19<=Edad<20

  2: Si 20<=Edad>=20
- Sexo : {0,1}

  Donde:

  0: Mujer

  1: Varón

- CursosMatriculados : {0,1,2}

  Donde:

  0: Si: 1<=CursosMatriculados <6

  1: Si: 6<=CursosMatriculados <7

  2: Si: 7<=CursosMatriculados

- TipoPreparacion : {0,1,2,3}

  Donde:

  0: Si TipoPreparacion = No respondió la encuesta

  1: Si TipoPreparacion = Admisión Ordinario

  2: Si TipoPreparacion = Cepru, todas las modalidades

  3: Si TipoPreparacion = Admisión primera oportunidad, otros.

- FormaDeEleccionEP : {0,1,2,3}

  Donde:

  0: FormaDeEleccionEP = No respondió la encuesta

  1: FormaDeEleccionEP = Por orientación Vocacional

  2: FormaDeEleccionEP = Pensando en mis actitudes

  3: FormaDeEleccionEP = Por posibilidades de trabajo, costo de la profesión, otros.

- Trabaja : {0,1,2,3}

  Donde:

  0: Trabaja = No respondió la encuesta

  1: Trabaja = No

  2: Trabaja = Si, en forma permanente

  3: Trabaja = Si, en forma eventual

- FinanciamientoEstudios : {0,1,2,3}

  Donde:

  0: FinanciamientoEstudios = No respondió la encuesta

  1: FinanciamientoEstudios = Ambos

  2: FinanciamientoEstudios = Padre o Madre

  3: FinanciamientoEstudios = Yo mismo, hermanos, otros parientes, instituciones

- Semestre_2014-1 : {0,1}

  Donde:

  0: No es el semestre 2014-1

  1: Si es el semestre 2014-1

- Semestre_2014-2 : {0,1}

  Donde:

  0: No es el semestre 2014-2

  1: Si es el semestre 2014-2

- Semestre_2017-2 : {0,1}

  Donde:

  0: No es el semestre 2017-2

  1: Si es el semestre 2017-2

- Semestre_2018-1 : {0,1}

  Donde:

  0: No es el semestre 2018-1

  1: Si es el semestre 2018-1

- NomEscuela_ADMINISTRACION : {0,1}

  Donde:

  0: No es ADMINISTRACION

  1: Si es ADMINISTRACION

- NomEscuela_AGRONOMIA : {0,1}

  Donde:

  0: No es AGRONOMIA

  1: Si es AGRONOMIA

- NomEscuela_ZOOTECNIA : {0,1}

  Donde:

  0: No es ZOOTECNIA

  1: Si es ZOOTECNIA

- Procedencia_ACOMAYO : {0,1}

  Donde:

  0: No es ACOMAYO

  1: Si es ACOMAYO

- Procedencia_ANTA : {0,1}

  Donde:

  0: No es ANTA

  1: Si es ANTA

- Procedencia_URUBAMBA : {0,1}

  Donde:

  0: No es URUBAMBA

  1: Si es URUBAMBA

- TipoColegio_PRIVADA : {0,1}

  Donde:

  0: No es Colegio_PRIVADA

  1: Si es Colegio_PRIVADA

- TipoColegio_PUBLICA : {0,1}

  Donde:

  0: No es Colegio_PUBLICA

  1: Si es Colegio_PUBLICA

- FormaIngreso_INGRESANTE : {0,1}

  Donde:

  0: No es INGRESANTE

  1: Si es INGRESANTE

- FormaIngreso_INGRESANTE_SO : {0,1}

  Donde:

  0: No es INGRESANTE_SO

  1: Si es INGRESANTE_SO

- ModalidadIngreso_ADMISION_DIRIMENCIA : {0,1}

  Donde:

  0: No es ADMISION_DIRIMENCIA

  1: Si es ADMISION_DIRIMENCIA

- ModalidadIngreso_ADMISION_ORDINARIO : {0,1}

  Donde:

  0: No es ADMISION_ORDINARIO

  1: Si es ADMISION_ORDINARIO

- CONDICION : {0,1}

  Donde:

  0: RENDIMIENTO MALO

  1: RENDIMIENTO BUENO

5.1.3.4. Integración de datos

Luego de la construcción de los nuevos datos, estos fueron integrados y en total se cuenta con 79 instancias o atributos y 12609 registros, los cuales serán utilizados para la predicción del rendimiento académico.

5.1.4. FASE IV: Modelado

En esta fase, se seleccionan y aplican las técnicas de modelado que sean pertinentes al problema (cuantas más mejor), y se calibran sus parámetros a valores óptimos.

5.1.4.1. Selección de técnicas de modelado.

Para el modelado se utilizará la herramienta WEKA 3.8, en donde se realizará el modelado de los datos utilizando 5 de los algoritmos más importantes para este tipo de casos de estudio, los resultados generados a partir de los datos tienen el formato siguiente:

```
=== Summary ===

Correctly Classified Instances       12451               98.7469 %
Incorrectly Classified Instances       158                1.2531 %
Kappa statistic                          0.9749
Mean absolute error                      0.0127
Root mean squared error                  0.0991
Relative absolute error                  2.5433 %
Root relative squared error             19.8315 %
Total Number of Instances            12609

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,987    0,012    0,987      0,987    0,987      0,975    0,995     0,993     0
                 0,988    0,013    0,988      0,988    0,988      0,975    0,995     0,993     1
Weighted Avg.    0,987    0,013    0,987      0,987    0,987      0,975    0,995     0,993

=== Confusion Matrix ===

    a    b   <-- classified as
 6039   80 |    a = 0
   78 6412 |    b = 1
```

Figura 23.- Modelo de resultados generados por WEKA 3.8 de los algoritmos de predicción

Fuente: Elaboración propia

1. ALGORITMO ARBOLES DE DECISIÓN J48

Los resultados generados por la herramienta WEKA 3.8 para el algoritmo "Arboles de decisión J-48", los resumimos en la tabla 19

Tabla 19. *Clasificación de instancias Algoritmo: Arboles de decisión J-48*

Algoritmo : ARBOLES DE DECISIÓN J-48		
Instancias correctamente clasificadas	8482	67,27%
Instancias incorrectamente clasificadas	4127	32,73%
TOTAL:	**12609**	**100,00%**

Fuente: Elaboración propia

Tabla 20. *Matriz de consistencia Algoritmo: Arboles de decisión J-48*

Matriz de Consistencia			
	Desaprobado	***Aprobado***	***Total***
Desaprobado	3898	2221	6119
Aprobado	2158	4332	6490

Fuente: Elaboración propia

Interpretación de resultados con árboles de decisión J-48

Como se puede apreciar clasifico de manera correcta hasta un 67.3%, y en la matriz de confusión se aprecia que:

De 6119 registros 3898 fueron clasificados correctamente, para la condición de "MALO"

De 6490 registros 4332 fueron clasificados correctamente, para la condición de "BUENO"

2.- ALGORITMO ARBOLES DE DECISIÓN RANDOM FOREST

Los resultados generados por la herramienta WEKA 3.8 para el algoritmo "Random Forest", los resumimos en la Tabla 21.

Tabla 21. *Clasificación de instancias Algoritmo: Random Forest*

Algoritmo : RANDOM FOREST		
Instancias correctamente clasificadas	8745	69,36%
Instancias incorrectamente clasificadas	3864	30,64%
TOTAL:	**12609**	**100,00%**

Fuente: Elaboración propia

Tabla 22. *Matriz de consistencia Algoritmo: Random Forest*

Matriz de Confusión			
	Desaprobado	***Aprobado***	***Total***
Desaprobado	4094	2025	6119
Aprobado	2091	4399	6490

Fuente: Elaboración propia

Interpretación de resultados con Árboles de Decisión Random Forest

Como se puede apreciar clasifico de manera correcta hasta un 69.4 %, y en la matriz de confusión se aprecia que:

De 6119 registros 4094 fueron clasificados correctamente, para la condición de "MALO"

De 6490 registros 4399 fueron clasificados correctamente, para la condición de "BUENO"

3. ALGORITMO DE VECINOS MÁS CERCANOS

Los resultados generados por la herramienta WEKA 3.8 para el algoritmo "Vecinos más cercanos", los resumimos en lo siguiente:

Tabla 23. *Clasificación de instancias Algoritmo: Vecinos más cercanos*

Algoritmo : VECINOS MÁS CERCANOS		
Instancias correctamente clasificadas	8051	63,85%
Instancias incorrectamente clasificadas	4558	36,15%
TOTAL:	**12609**	**100,00%**

Fuente: Elaboración propia

Tabla 24. *Matriz de consistencia Algoritmo: Vecinos más cercanos*

Matriz de Consistencia			
	Desaprobado	***Aprobado***	***Total***
Desaprobado	3866	2253	6119
Aprobado	2557	3933	6490

Fuente: Elaboración propia

Interpretación de resultados del algoritmo Vecinos Más Cercanos

Como se puede apreciar clasifico de manera correcta hasta un 63.8 %, y en la matriz de confusión se aprecia que:

De 6119 registros 3866, fueron clasificados correctamente, para la condición de "MALO"

De 6490 registros 3933, fueron clasificados correctamente, para la condición de "BUENO".

4.- ALGORITMO DE FUNCIÓN LOGÍSTICA

Los resultados generados por la herramienta WEKA 3.8 para el algoritmo "Función Logística", los resumimos en la tabla 25.

Tabla 25. *Clasificación de instancias Algoritmo: Función Logística*

Algoritmo : FUNCIÓN LOGISTICA		
Instancias correctamente clasificadas	8620	68,37%
Instancias incorrectamente clasificadas	3989	31,63%
TOTAL:	**12609**	**100,00%**

Fuente*: Elaboración propia*

Tabla 26. *Matriz de consistencia Algoritmo: Función logística*

Matriz de Consistencia			
	Desaprobado	***Aprobado***	***Total***
Desaprobado	4009	2110	6119
Aprobado	2131	4359	6490

Fuente: Elaboración propia

Interpretación de resultados algoritmo Función Logística

Como se puede apreciar clasifico de manera correcta hasta un 68 %, y en la matriz de confusión se aprecia que:

De 6119 registros 4009 fueron clasificados correctamente, para la condición de "MALO"

De 6490 registros 4359 fueron clasificados correctamente, para la condición de "BUENO"

5. - ALGORITMO PERCEPTRÓN MULTICAPA

Los resultados generados por la herramienta WEKA 3.8 para el algoritmo "Perceptrón multicapa", los resumimos en lo siguiente:

Tabla 27. *Clasificación de instancias Algoritmo: Perceptrón Multicapa*

Algoritmo : PERCEPTRÓN MULTICAPA		
Instancias correctamente clasificadas	8493	64,80%
Instancias incorrectamente clasificadas	4116	35,20%
TOTAL:	**12609**	**100,00%**

Fuente: Elaboración propia

Tabla 28. *Matriz de consistencia Algoritmo: Función logística*

Matriz de Consistencia			
	Desaprobado	***Aprobado***	***Total***
Desaprobado	3657	2462	6119
Aprobado	2228	4262	6490

Fuente: Elaboración propia

Interpretación de resultados Algoritmo Perceptrón Multicapa

Como se puede apreciar clasifico de manera correcta hasta un 68 %, y en la matriz de confusión se aprecia que:

De 6119 registros 3657 fueron clasificados correctamente, para la condición de "MALO"

De 6490 registros 4262 fueron clasificados correctamente, para la condición de "BUENO".

5.1.4.2. Generación de un diseño de comprobación.

El diseño de comprobación será generado a partir de los resultados mostrados en la Tabla 29.

Tabla 29.-Resumen de algoritmos con porcentajes de predicción y aciertos

Nro.	ALGORITMO	Porcentaje de predicción acertada	Acertados para "MALO" de 6119 Registros	Acertados para "BUENO" de 6490 Registros
1	Árboles de decisión J-48	67,27%	3898	4323
2	Árboles de decisión Random Forest	69,35%	4094	4399
3	Algoritmo de Vecino más Cercano	63,85%	3866	3933
4	Algoritmo de Función Logística	68,33%	4009	4359
5	Algoritmo de Perceptrón Multicapa	64,80%	3657	4262

Fuente: Elaboración propia

Como se puede apreciar el Algoritmo de árboles de decisión Random Forest, es el que tiene el mejor desempeño, por consiguiente, es con este algoritmo que se realizará la predicción del rendimiento académico de los estudiantes de la UNSAAC.

5.1.4.3. Generación del modelo.

Para la generación de los modelos utilizaremos los mismos algoritmos descritos anteriormente, y para esto dividiremos nuestra información en 2 grupos:

- **Primer grupo (Entrenamiento):**

Formado por todos las instancias o registros desde el semestre 2014-I hasta el semestre 2017-II; en total 10790 Registros

- **Segundo Grupo (Test):**

Formado únicamente por las instancias o registros del semestre 2018-I, en total 1819 Registros.

Para el segundo grupo (Test), el atributo **Observación** (1,0): "MALO" o "BUENO", será desconocido y lo representaremos con un signo de interrogación "**?**", el algoritmo será el que determine en la interrogación si el rendimiento del estudiante será "BUENO" o "MALO"

5.1.4.4. Evaluación y comprobación del modelo.

Para la evaluación y comprobación del modelo se implementó cada uno de los algoritmos en WEKA 3.8, generándose los siguientes resultados:

El software nos genera 2 resultados en función a los datos utilizados para el Entrenamiento y para el Test, similares a las ventanas siguientes:

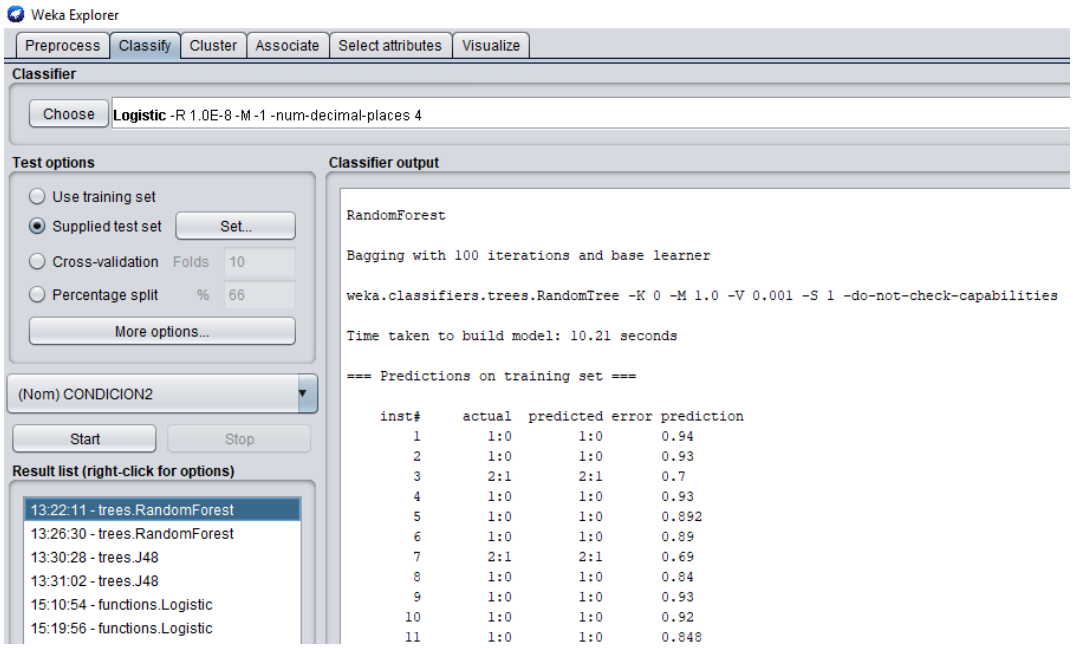

Figura 24.- Pantalla con los datos de Entrenamiento

Fuente: Elaboración propia

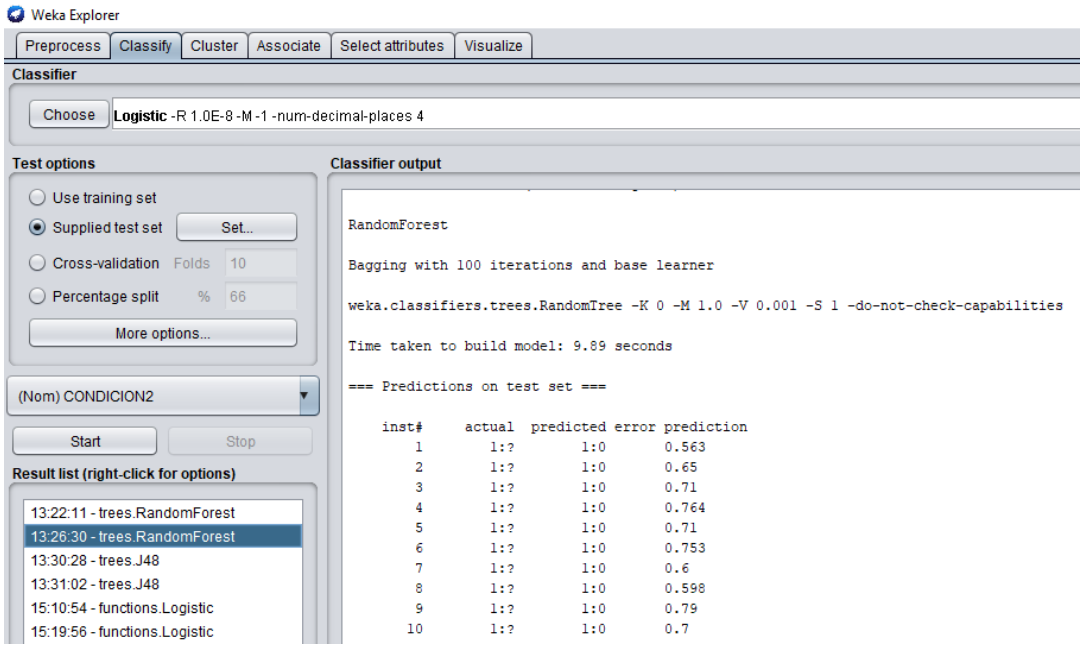

Figura 25.- Pantalla con los datos del Test

Fuente: Elaboración propia

5.1.5. FASE V: Evaluación

En esta etapa en el proyecto, se han construido uno o varios modelos que parecen alcanzar calidad suficiente desde la una perspectiva de análisis de datos.

5.1.5.1. Evaluación de resultados.

Los resultados son obtenidos a partir del test generado por WEKA 3.8, como se muestra un fragmento en la pantalla siguiente:

Tabla 30. *Algoritmo Random Forest para predecir rendimiento académico del semestre 2018-1*

```
RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V

Time taken to build model: 9.39 seconds

=== Predictions on test set ===

    inst#     actual  predicted error prediction
        1        1:?        1:0       0.563
        2        1:?        1:0       0.65
        3        1:?        1:0       0.71
        4        1:?        1:0       0.764
        5        1:?        1:0       0.71
        6        1:?        1:0       0.753
        7        1:?        1:0       0.6
        8        1:?        1:0       0.598
        9        1:?        1:0       0.79
       10        1:?        1:0       0.7
       11        1:?        1:0       0.8
       12        1:?        1:0       0.533
       13        1:?        2:1       0.628
       14        1:?        2:1       0.508
       15        1:?        1:0       0.698
       16        1:?        1:0       0.58
       17        1:?        1:0       0.635
       18        1:?        1:0       0.598
       19        1:?        1:0       0.6
       20        1:?        1:0       0.81
       21        1:?        1:0       0.82
       22        1:?        2:1       0.52
       23        1:?        1:0       0.637
       24        1:?        1:0       0.695
       25        1:?        1:0       0.633
       26        1:?        1:0       0.857
       27        1:?        1:0       0.677
       28        1:?        1:0       0.6
       29        1:?        1:0       0.67
       30        1:?        1:0       0.85
       31        1:?        1:0       0.683
       32        1:?        1:0       0.8
       33        1:?        1:0       0.656
       34        1:?        1:0       0.678
```

Fuente: Elaboración propia

Para la interpretación de la pantalla anterior utilizamos una hoja electrónica utilizando la función SI, para determinar si “ACERTÓ” o “NO ACERTÓ”, a continuación, un fragmento de la tabla

Tabla 31. *Aciertos y desaciertos del algoritmo Random Forest*

Instancia	**Valor real Semestre 2018-1**	**Predicción con Random Forest**	**Observación**
1	0	0	ACERTO
2	0	0	ACERTO
3	0	0	ACERTO
4	1	0	NO ACERTO
5	0	0	ACERTO
6	0	0	ACERTO
7	0	0	ACERTO
8	0	0	ACERTO
9	1	0	NO ACERTO
10	0	0	ACERTO
11	0	0	ACERTO
12	0	0	ACERTO
13	0	1	NO ACERTO
14	1	1	ACERTO
15	1	0	NO ACERTO
16	0	0	ACERTO
17	0	0	ACERTO
18	1	0	NO ACERTO
19	0	0	ACERTO
20	0	0	ACERTO
21	1	0	NO ACERTO
22	0	1	NO ACERTO
23	0	0	ACERTO
24	0	0	ACERTO
25	1	0	NO ACERTO
26	0	0	ACERTO
27	1	0	NO ACERTO
28	0	0	ACERTO
29	0	0	ACERTO
30	0	0	ACERTO
31	1	0	NO ACERTO
32	0	0	ACERTO
33	0	0	ACERTO

Fuente: Elaboración propia

5.1.5.2. Proceso de revisión.

Para el proceso de revisión de los resultados se generó un prototipo desarrollado con el lenguaje de programación JAVA y NETBEANS IDE 8.2, por ser herramientas compatibles con WEKA 3.8 descrito en las siguientes pantallas:

Figura 26.- Formulario de Inicio para generar el rendimiento académico en la UNSAAC

Fuente: Elaboración propia

Figura 27.- Rendimiento académico de los estudiantes de la UNSAAC, por semestre

Fuente: Elaboración propia

El paso 1 es el de abrir el archivo de entrenamiento, el paso 2 es abrir el archivo de test ambos de extensión ".arff", para luego hacer clic en el botón "Cargar" y nos mostrará la predicción del rendimiento académico del semestre seleccionado, en este caso del semestre 2018-1

Figura 28.-Rendimiento académico de los estudiantes de la UNSAAC por Escuelas Profesionales

Fuente: Elaboración propia

Con el formulario anterior el prototipo nos da la opción de mostrar el rendimiento académico de los estudiantes de la UNSAAC, realizando filtros por Escuela Profesional, y del mismo modo poder generar reportes que permitan visualizar la información como se muestra en la Figura 29

UNIVERSIDAD NACIONAL DE SAN ANTONIO ABAD DEL CUSCO

PREDICCIÓN DEL RENDIMIENTO ACADÉMICO

Tabla de Alumnos por Carreras

Nro	Escuela	Alumno	Condición
1	GEOLOGICA	Alumno1	Bueno
2	GEOLOGICA	Alumno2	Malo
3	GEOLOGICA	Alumno3	Bueno
4	GEOLOGICA	Alumno4	Malo
5	GEOLOGICA	Alumno5	Bueno
6	GEOLOGICA	Alumno6	Malo
7	GEOLOGICA	Alumno7	Malo
8	GEOLOGICA	Alumno8	Malo
9	GEOLOGICA	Alumno9	Malo
10	GEOLOGICA	Alumno10	Bueno
11	GEOLOGICA	Alumno11	Bueno
12	GEOLOGICA	Alumno12	Bueno
13	GEOLOGICA	Alumno13	Bueno
14	GEOLOGICA	Alumno14	Malo
15	GEOLOGICA	Alumno15	Bueno
16	GEOLOGICA	Alumno16	Bueno
17	GEOLOGICA	Alumno17	Malo
18	GEOLOGICA	Alumno18	Malo
19	GEOLOGICA	Alumno19	Bueno
20	GEOLOGICA	Alumno20	Malo
21	GEOLOGICA	Alumno21	Bueno
22	GEOLOGICA	Alumno22	Malo
23	GEOLOGICA	Alumno23	Bueno
24	GEOLOGICA	Alumno24	Malo
25	GEOLOGICA	Alumno25	Bueno
26	GEOLOGICA	Alumno26	Malo
27	GEOLOGICA	Alumno27	Bueno
28	GEOLOGICA	Alumno28	Malo
29	GEOLOGICA	Alumno29	Malo
30	GEOLOGICA	Alumno30	Malo
31	GEOLOGICA	Alumno31	Malo
32	GEOLOGICA	Alumno32	Malo
33	GEOLOGICA	Alumno33	Bueno
34	GEOLOGICA	Alumno34	Malo
35	GEOLOGICA	Alumno35	Bueno
36	GEOLOGICA	Alumno36	Malo
37	GEOLOGICA	Alumno37	Bueno

Figura 29.- Modelo de reporte de rendimiento académico (Ing. Geológica).

Fuente: Elaboración propia

5.1.5.3. Determinar los pasos siguientes a base de los resultados

Los pasos siguientes a estos resultados seria desarrollar la predicción del rendimiento académico de los estudiantes ingresantes en los semestres posteriores, y buscar que el archivo de entrenamiento sea cada vez mayor y pueda generar predicciones con mayor porcentaje.

5.1.6. FASE VI: DESPLIEGUE

Generalmente, la creación del modelo no es el final del proyecto. Incluso si el objetivo del modelo es de aumentar el conocimiento de los datos, el conocimiento obtenido tendrá que organizarse y presentarse para que el cliente pueda usarlo.

5.1.6.1. Planificación de distribución.

La planificación de la distribución de la presente tesis, estará enfocada a la entrega del prototipo a las instancias correspondientes de la Universidad Nacional de San Antonio Abad del Cusco, específicamente al Vice rectorado académico, que es el órgano que vela por la parte académica de la Institución, así mismo a los coordinadores de cada escuela Profesional para que tomen las consideraciones correspondientes.

5.1.6.2. Creación del informe final.

El informe final está descrito en la redacción del presente documento.

5.1.6.3. Revisión final del proyecto.

El presente proyecto podrá ser revisado y evaluado por los señores dictaminantes del presente proyecto así mismo por los encargados de velar por la parte académica de la UNSAAC.

DISCUSIÓN

- Este trabajo de tesis tuvo como propósito realizar la predicción del rendimiento académico de los estudiantes de la UNSAAC en su primer semestre, a partir de sus datos de ingreso y la encuesta web que los postulantes completan para poder inscribirse al examen de admisión a la institución, para lograr este objetivo se utilizaron los algoritmos de aprendizaje automático, también se pudo determinar los factores más importantes que inciden en el rendimiento académico, utilizando técnicas estadísticas.
- Los algoritmos con la mejor performance es decir con la mejor predicción en porcentajes fueron el algoritmo de Random Forest con un 69.35 % de acierto en la predicción, seguido del algoritmo de Función Logística con un 68.33%.
- Gonzales, C & Rodríguez, C (2017), en su trabajo Propuesta de un Modelo de Business Intelligence para identificar el perfil de deserción estudiantil en la Universidad Científica del Sur, identifica como factores de riesgo a: la Carrera, semestre cursado, cantidad de materias cursadas, modalidad de Ingreso, edad, ciudad de procedencia, Genero y el nivel de estudios de los padres como las más importantes, en la presente tesis se encontraron como factores importantes que afectan al rendimiento académico los siguientes: nota de ingreso a la Universidad, escuela profesional que estudia, semestre, genero, modalidad de ingreso, cantidad de cursos matriculados, forma de ingreso, procedencia, edad, si trabaja y el tipo de preparación.
- Acosta de la Cruz, P. & Pizarro, P. (2011), utilizando redes neuronales de retro propagación y Función Logística, obtuvo 70.45% y 70.39% respectivamente para los modelos de predicción de aprobación o no de una asignatura, y nuestros

resultados se aproximan a estos valores, cabe destacar que la predicción de este trabajo no es de solamente una asignatura sino del rendimiento académico de todo un semestre.

- Menacho C & Higinio C (2017) *Predicción del rendimiento académico aplicando técnicas de minería de datos, Universidad Nacional Agraria La Molina,* utiliza la técnica de la red Naive de Bayes alcanzando una precisión del 71,0% de correcta clasificación para solamente un curso con datos de solo 2 semestres (2013-II y 2014-I), considero que la información es muy escasa, puesto que a mayor cantidad de datos la predicción es mejor.

CONCLUSIONES

- El rendimiento académico de los estudiantes es un tema bastante complejo y que depende de muchas variables no solo socio demográficas y socio educativas, sino que dependen también de otros factores como el aspecto emocional de los estudiantes y también la familia, como se verificó en la revisión bibliográfica, sin embargo, el rendimiento académico se puede predecir a través de los datos de ingreso o de admisión a la UNSAAC, utilizando los algoritmos de aprendizaje automático hasta en un 69 % de efectividad.
- Los factores claves de los datos de ingreso que determinan el rendimiento académico de los estudiantes a partir de los datos de ingreso son diversos, los principales son: la Nota de ingreso, la escuela profesional que se estudia, el semestre, el género, y la modalidad de ingreso, esta información fue generada por el análisis estadístico utilizando chi cuadrado y el coeficiente de correlación de Pearson, cumpliéndose con la hipótesis planteada.
- Según los resultados encontrados, el algoritmo de árboles de decisión “Random Forest”, fue el algoritmo que tuvo el mejor performance para la predicción del rendimiento académico de los ingresantes en los primeros semestres a la UNSAAC con un 69% de predicción, el segundo algoritmo con mejor performance fue algoritmo de Regresión Logística con un 68% para el presente caso de estudio.

RECOMENDACIONES

- A las instancias correspondientes y en el caso de la UNSAAC a la unidad de Admisión generar políticas para mejorar la obtención de la información de los postulantes a la Universidad lo cual permitiría tener información más relevante y precisa y de esta manera mejorar sustancialmente el porcentaje de predicción.
- Utilizar los algoritmos de aprendizaje automático y las técnicas de clasificación para predecir el rendimiento académico de los estudiantes en función a los datos históricos con los que cuenta la institución educativa, podría ser una herramienta que apoye a la toma de decisiones y a mejorar las políticas para un buen rendimiento académico de los estudiantes universitarios.
- Utilizar la metodología CRISP-DM, por ser una metodología que guía a través de sus 6 fases bien definidas todo el proceso de la generación de modelos predictivos, considerando además que esta metodología es bastante sencilla de entender y de poder aplicarla a cualquier situación de minería de datos y de aprendizaje automático, independientemente de la herramienta de predicción a utilizar.
- Utilizar para este tipo de investigaciones el algoritmo *Random Forest* así como la herramienta WEKA para el procesamiento de la información y la predicción por ser un algoritmo con mejor performance para este tipo de situaciones y una herramienta bastante visual y versátil respectivamente.
- Apoyarse en las herramientas estadísticas para mejorar la comprensión de la información y de los datos.
- Como trabajos futuros se recomienda investigar la predicción del posible rendimiento académico del docente universitario, del mismo modo sería importante investigar los algoritmos no supervisados y su aplicación para casos similares al desarrollado en la presente tesis.

BIBLIOGRAFÍA

Acosta de la Cruz, P. & Pizarro, P. (2011). *Predicción del rendimiento académico en la educación superior usando -minería de datos y su comparación con técnicas estadísticas*, Lima, Perú.

A visual guide to CRISP-DM methodology. (2009). Recuperado 27 de enero de 2019, de https://exde.wordpress.com/2009/03/13/a-visual-guide-to-crisp-dm-methodology/

Blanco, L. (1989). *Rendimiento académico en la universidad de Cantabria: abandono y retraso en los estudios*. Cantabria, España.

Camborda, M. (2014). *Aplicación de árboles de decisión para la predicción del rendimiento académico de los estudiantes de los primeros ciclos de la carrera de Ingeniería Civil de la Universidad Continental*, Huancayo, Perú.

El Comercio (2016). *Gasto público anual en el Perú por alumno*, recuperado de: https://elcomercio.pe/peru/gasto-publico-anual-peru-alumno-us-1-100-172528

Escorza, Y. & Camacho, D. (2014). *Factores que afectan el desempeño académico*. Monterrey, México.

EcuRed. (2014). *Matrices de confusión*, Recuperado 5 de marzo de 2019, de: https://www.ecured.cu/Matrices_de_confusi%C3%B3n

Estatuto Asamblea Universitaria de la Universidad Nacional de San Antonio Abad del Cusco (2015), Recuperado 27 de enero de 2019 de: http://transparencia.unsaac.edu.pe/links/datosgenerales/documentos/ESTATUTO%20UNSAAC%20-%202015.pdf

Galán, V. (2015) *Aplicación de la metodología CRISP-DM a un proyecto de minería de datos en el entorno universitario*, Universidad Carolos III de Madrid, España.

García, D (2015), *Construcción de un modelo para determinar el rendimiento académico de los estudiantes basado en Learning AnalyTIC's (Análisis del aprendizaje), mediante el uso de técnicas multivariantes*, Universidad de Sevilla. Sevilla, España.

Gonzales, C. & Rodríguez, C. (2017). *Propuesta de un Modelo de Business Intelligence para Identificar el perfil de deserción estudiantil en la Universidad Científica del Sur*, Lima, Perú.

González, L (2018), *Aprende todo sobre Inteligencia Artificial*, Venezuela, recuperado de:

http://ligdigonzalez.com/aprendizaje-supervisado-random-forest-classification/

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

Gonzalo, A (2018), *Machine learning, data science y analítica avanzada*, recuperado de: http://machinelearningparatodos.com/tipos-de-aprendizaje-automatico/

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

Kaufmann, M. & Han, J. Kamber, M & Pei, J. (2011) *Minería de datos: conceptos y técnicas,* Tercera edición. Illinois, Estados Unidos.

Kelleher, J. & Mac, B. & D'Arcy, A. (s.f.) *Fundamnetals of Machine learning for predictive data analyTIC's algorithms, worked, and case of studies*, London, Inglaterra.

Marqués, C (2015), *Predicción del fracaso y el abandono escolar mediante técnicas de minería de datos*, Córdoba, España

Management solutions (2018), *Machine Learning, una pieza clave en la transformación de los modelos del negocio*, recuperado de: https://www.managementsolutions.com/sites/default/files/publicaciones/esp/machine-learning.pdf

Marqués, M (2014), *Minería de datos a través de ejemplos*, RC LIBROS (SC LIBRO)

Mayorga, H. (2016). *Minería de procesos: Fundamentos y metodología de aplicación.* Editorial Pontificia Universidad Javeriana, Bogotá, Colombia.

Menacho, C. & Higinio, C. (2017) *Predicción del rendimiento académico aplicando técnicas de minería de datos*, Universidad Nacional Agraria La Molina, Lima, Perú.

Mitchell, Th. (1997). *Aprendizaje automático.* McGRAW-Hill, 1ra. Edición.

Moreno, A. & Armengol, E. & Béjar, J. & Belanche, L. & Cortés, U. & Gavalda, R. & Gimeno, J. & López, B. & Martin, M. & Sánchez, M. (1994), *Aprendizaje automático*, Barcelona, España. Edicions de la Universidad Politécnica de Catalunya.

Pérez, C. & Gonzales, S. (2007). *Minería de datos técnicas y herramientas*, Madrid España.

Quintana, V & Yagual, S (2017). *Propuesta de aplicación predictiva de aprobación de una asignatura con flujo previo a través de algoritmos basados en software WEKA para estudiantes del último semestre de la Carrera de Ingeniería en Sistemas Computacionales de la Universidad de Guayaquil*, Guayaquil, Ecuador.

Russell, S & Norvig, P, (2004). *Inteligencia Artificial un enfoque moderno*, Madrid, España.

Recuero, P. (2019). *Los 2 tipos de aprendizaje en Machine Learning: supervisado y no supervisado.* Recuperado 26 de enero de 2019, de

https://data-speaks.luca-d3.com/2017/11/que-algoritmo-elegir-en-ml-aprendizaje.html

Sierra, B. (2006). Aprendizaje Automático: *conceptos básicos y avanzados, Aspectos prácticos utilizando el software WEKA*, Madrid, España, Editorial PEARSON PRENTICE HALL.

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

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

WEKA 3, *Minería de datos con software de aprendizaje de código abierto en Java*. (2014). Recuperado 21 de febrero de 2019, de https://www.cs.waikato.ac.nz/ml/weka/

Zambrano, C. & Rojas D. & Carvajal, K. & Acuña**,** G. (2011). *Análisis de rendimiento académico estudiantil usando data warehouse y redes neuronales*, Universidad de Atacama, Arica, Chile.

Zambrano, J. (2018). *¿Aprendizaje supervisado o no supervisado?* Recuperado 26 de enero de 2019, de https://medium.com/@juanzambrano/aprendizaje-supervisado-o-no-supervisado-39ccf1fd6e7b

ANEXOS

Anexo 01.- Matriz de consistencia

PROBLEMA	OBJETIVOS	HIPÓTESIS	VARIABLES	INDICADORES	INSTRUM
Problema general: ¿Es posible predecir el rendimiento académico de los estudiantes de la UNSAAC en el primer semestre a partir de sus datos de ingreso utilizando algoritmos de aprendizaje automático? **Problemas específicos** **1.**- ¿Cuáles son los factores claves de los datos de ingreso que determinan el rendimiento académico de los estudiantes de la UNSAAC en su primer semestre? **2**. ¿Cuál es el algoritmo de aprendizaje automático más eficiente que predice el rendimiento académico de los estudiantes de la UNSAAC a partir de sus datos de ingreso en el primer semestre?	**Objetivo general** Predecir el rendimiento académico de los estudiantes de la UNSAAC del primer semestre a partir de sus datos de ingreso utilizando algoritmos de aprendizaje automático. **Objetivos específicos** **1**.-Determinar los factores claves de los datos de ingreso a la UNSAAC, que permiten la predicción del rendimiento académico de los estudiantes en el primer semestre. **2**.- Determinar el algoritmo de aprendizaje automático más eficiente, para predecir el rendimiento académico de los estudiantes de la UNSAAC del primer semestre a partir de sus datos de ingreso a la Universidad.	**Hipótesis General** Los algoritmos de aprendizaje automático, determinan la predicción del rendimiento académico de los estudiantes de la UNSAAC del primer semestre, a partir de sus datos de ingreso con una eficiencia de 70% **Hipótesis especificas** 1: La predicción del rendimiento académico de los estudiantes de la UNSAAC a partir de sus datos de ingreso dependen de los factores sociodemográficos y socioeducativas. 2: El algoritmo con mejor performance de predicción del rendimiento académico de los estudiantes de la UNSAAC en su primer semestre es Random Forest.	**Variables independientes** Algoritmos de aprendizaje automático Factores socio demográficos y socio educativos **Variables dependientes** Rendimiento académico	• Performance de los Algoritmos de aprendizaje automático • Edad • Genero • Procedencia • Tipo de colegio • Forma de Ingreso • Modalidad de Ingreso • Nota de ingreso • Colegio de procedencia • Financiamiento de estudios • Tipo de preparación para el ingreso Condición de: **Nivel de Bueno** **Nivel de Malo**	Encuesta web de la página de admisión de la UNSAAC e información de datos de ingreso

Anexo 02.- Encuesta a postulantes a la UNSAAC

ENCUESTA A POSTULANTES

Lee con atención las preguntas que se presentan a continuación y marca la opción que corresponde a la alternativa que mejor coincida con tu realidad o tus opiniones según los casos; responde con sinceridad, pues los datos que proporciones ayudarán a conocer mejor a quienes postulan a nuestra universidad y a brindarles el apoyo que corresponde a nuestra institución.

1: ¿Trabajas?

- No
- Sí, en forma permanente
- Sí, en forma eventual

2: ¿Quién financia tus estudios?

- Padre
- Madre
- Ambos
- Yo mismo
- Hermanos
- Otros parientes
- Otras personas o Instituciones

3: ¿Qué tipo de preparación recibiste para el ingreso a la Universidad?

- Por mi cuenta
- CEPRU-UNSAAC
- Profesor particular
- Sólo academias
- CEPRU y Academia

4: ¿Cómo elegiste tu carrera profesional?

- Por orientación vocacional
- Por las posibilidades de trabajo
- Por influencia familiar

- Pensando en mis aptitudes
- Por el costo de la profesión
- Otro

5: ¿A través de qué medio de difusión se enteró del Examen de Admisión?

- TV.
- Radio.
- Página "Web UNSAAC"
- Periódico.
- Otro.

6: ¿Qué conocimientos en computación tienes?

- Word.
- Excel
- Power Point
- Todos
- Ninguno

7: ¿Que te incentiva postular a la UNSAAC?

- Por la gratuidad de la enseñanza
- Por su prestigio
- Por la diversidad de carreras que ofrece
- Por presión familiar
- Otros

Anexo 03.- Documento de entrega de información de la Unidad de Centro de Computo

UNIVERSIDAD NACIONAL DE SAN ANTONIO ABAD DEL CUSCO

UNIDAD DEL CENTRO DE CÓMPUTO

Cusco, 17 de octubre de 2018

SEÑOR:
DR. LAURO ENCISO RODAS
ASESOR DE PROYECTO DE TESIS DE POSTGRADO EN CIENCIAS MENCION INFORMÁTICA

ASUNTO : ALCANZA INFORMACIÓN SOLICITADA PARA PROYECTO DE TESIS

Es grato dirigirme a usted, con la finalidad de hacer llegar la información solicitada para el desarrollo de proyecto de tesis de postgrado intitulado "PREDICCIÓN DEL RENDIMIENTO ACADÉMICO DE LOS ESTUDIANTES DE LA UNSAAC A PARTIR DE SUS DATOS DE INGRESO, UTILIZANDO ALGORITMOS DE APRENDIZAJE AUTOMATICO".

Debo aclarar que la información alcanzada corresponde a encuestas que se aplican a postulantes, y rendimiento académico de ingresantes a nuestra Universidad desde el semestre 2014-1 al semestre 2018-1, información que solicitamos se utilizada con ética y responsabilidad.

Sin otro particular, uso de la ocasión para expresarle mis consideraciones más distinguidas.

Atentamente.

UNIVERSIDAD NACIONAL DE SAN ANTONIO ABAD DEL CUSCO
DIRECCIÓN DE LA UNIDAD DE CENTRO DE COMPUTO

Universidad Nacional de San Antonio Abad del Cusco
DIRECCION DE LA UNIDAD DE CENTRO DE COMPUTO

Ing. Luis Beltran Palma Ttito
DIRECTOR

CC/LPT
c.c.
Archivo

Anexo 04.- Fragmento del código preparado para el procesamiento con WEKA 3.8

```
@relation DATOS_TESIS_PREDICCION
%Predicción del rendimiento académico de los estudiantes de la UNSAAC utilizando Machine Learning (Algoritmos de
%clasificación)
%Diciembre de2018
%ATRIBUTOS
@attribute NotaIngresoCP {0,1,2}
@attribute Edad2 {0,1,2}
@attribute Genero {0,1}
@attribute CursosMatriculados {0,1,2,3}
@attribute TipoPreparacion {0,1,2,3}
@attribute FormaDeEleccionEP {0,1,2,3}
@attribute Trabaja {0,1,2,3}
@attribute FinanciamientoEstudios {0,1,2,3}

@attribute Semestre_2014-1 {0,1}
@attribute Semestre_2014-2 {0,1}
@attribute Semestre_2015-1 {0,1}
@attribute Semestre_2015-2 {0,1}
@attribute Semestre_2016-1 {0,1}
@attribute Semestre_2016-2 {0,1}
@attribute Semestre_2017-1 {0,1}
@attribute Semestre_2017-2 {0,1}
@attribute Semestre_2018-1 {0,1}

@attribute NomEscuela_ADMINISTRACION numeric
@attribute NomEscuela_AGRONOMIA numeric
@attribute NomEscuela_ANTROPOLOGIA numeric
@attribute NomEscuela_ARQUEOLOGIA numeric
@attribute NomEscuela_ARQUITECTURA numeric
@attribute NomEscuela_BIOLOGIA numeric
@attribute NomEscuela_COMUNICACION numeric
@attribute NomEscuela_CONTABILIDAD numeric
@attribute NomEscuela_DERECHO numeric
@attribute NomEscuela_ECONOMIA numeric
@attribute NomEscuela_EDUCACION numeric
@attribute NomEscuela_ENFERMERIA numeric
@attribute NomEscuela_FARMACIA numeric
@attribute NomEscuela_FISICA numeric
@attribute NomEscuela_GEOLOGICA numeric
@attribute NomEscuela_HISTORIA numeric
@attribute NomEscuela_INFORMATICA numeric
@attribute NomEscuela_INGENIERIA_QUIMICA numeric
@attribute NomEscuela_MATEMATICAS numeric
```

@attribute NomEscuela_MECANICA numeric
@attribute NomEscuela_MEDICINA numeric
@attribute NomEscuela_METALURGICA numeric
@attribute NomEscuela_ODONTOLGIA numeric
@attribute NomEscuela_OTROS numeric
@attribute NomEscuela_PSICOLOGIA numeric
@attribute NomEscuela_QUIMICA numeric
@attribute NomEscuela_TURISMO numeric
@attribute NomEscuela_VETERINARIA numeric
@attribute NomEscuela_ZOOTECNIA numeric

@attribute Procedencia_ACOMAYO numeric
@attribute Procedencia_ANTA numeric
@attribute Procedencia_APURIMAC numeric
@attribute Procedencia_AREQUIPA numeric
@attribute Procedencia_CALCA numeric
@attribute Procedencia_CANAS numeric
@attribute Procedencia_CANCHIS numeric
@attribute Procedencia_CHUMBIVILCAS numeric
@attribute Procedencia_CUSCO_CUSCO numeric
@attribute Procedencia_CUSCO_SANTIAGO numeric
@attribute Procedencia_CUSCO_SAN_JERONIMO numeric
@attribute Procedencia_CUSCO_SAN_SEBASTIAN numeric

@attribute Procedencia_CUSCO_WANCHAQ numeric
@attribute Procedencia_ESPINAR numeric
@attribute Procedencia_LIMA numeric
@attribute Procedencia_MADRE_DIOS numeric
@attribute Procedencia_OTROS numeric
@attribute Procedencia_PARURO numeric
@attribute Procedencia_PAUCARTAMBO numeric
@attribute Procedencia_PUNO numeric
@attribute Procedencia_QUILLABAMBA numeric
@attribute Procedencia_QUISPICANCHI numeric
@attribute Procedencia_URUBAMBA numeric

@attribute TipoColegio_PRIVADA {0,1}
@attribute TipoColegio_PUBLICA {0,1}

@attribute FormaIngreso_INGRESANTE {0,1}
@attribute FormaIngreso_INGRESANTE_SO {0,1}

@attribute ModalidadIngreso_ADMISION_DIRIMENCIA numeric
@attribute ModalidadIngreso_ADMISION_ORDINARIO numeric
@attribute ModalidadIngreso_ADMISION_PRIMERA_OPORTUNIDAD numeric

@attribute ModalidadIngreso_ADMISION_SEDES_FILIALES numeric
@attribute ModalidadIngreso_ADMISION_CEPRU_INTENSIVO numeric
@attribute ModalidadIngreso_ADMISION_CEPRU_ORDINARIO numeric
@attribute ModalidadIngreso_ADMISION_CEPRU_PRIMERA_OPORTUNIDAD numeric
@attribute ModalidadIngreso_OTROS numeric

@attribute CONDICION {0,1} % 0: Malo; 1: Bueno

%Datos de Excel formato csv
@data

1,1,1,3,3,3,2,1,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,
0,1,1,0,0,1,0,0,0,0,0,0,10.2,1,0
1,1,1,3,3,1,3,1,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,
0,1,1,0,0,0,1,0,0,0,0,0,11.4,2,0
1,1,1,3,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,
0,1,1,0,0,0,0,0,0,1,0,12.6,2,1
1,3,1,3,3,3,2,3,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,1,1,0,0,0,0,0,0,1,0,9.6,1,0
1,3,1,3,2,2,3,1,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,1,0,1,0,0,0,0,0,0,1,11.2,2,0
1,1,1,3,3,3,1,1,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,
0,1,1,0,0,0,1,0,0,0,0,0,8.5,1,0
1,1,2,3,2,2,1,1,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,
0,1,1,0,0,0,1,0,0,0,0,0,12.5,2,1
1,1,1,3,3,3,1,1,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,
0,1,1,0,0,0,1,0,0,0,0,0,10.0,1,0
…

…

…