



UNIVERSIDAD CÉSAR VALLEJO

FACULTAD DE INGENIERÍA Y ARQUITECTURA  
ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

**Machine Learning para predecir el rendimiento académico de los  
estudiantes universitarios**

TESIS PARA OBTENER EL TÍTULO PROFESIONAL DE:

Ingeniero de Sistemas

**AUTOR:**

García Dionisio, Jeancarlos Donato (ORCID: 0000-0001-6739-169X)

**ASESORA:**

Mg. Menéndez Mueras, Rosa (ORCID: 0000-0003-2403-7679)

**LÍNEA DE INVESTIGACIÓN:**

Sistemas de Información y comunicaciones

**LIMA-PERÚ**

**2021**

## Dedicatoria

A mis padres Donato Garcia y Susana Dionisio, por darme la oportunidad de ser profesional, por ser el apoyo y la fortaleza en momentos de dificultad y debilidad. A mi hermana Diana Garcia por sus consejos y por enseñarme a persistir.

## Agradecimiento

A mis padres y hermana porque fueron los que me impulsaron a dar este gran paso, siempre me motivan para ser una mejor persona y un gran profesional. Finalmente, a la Universidad César Vallejo y a toda su plana docente.

## Índice de contenidos

Dedicatoria .....	ii
Agradecimiento.....	iii
Índice de contenidos.....	iv
Índice de tablas .....	v
Índice de gráficos y figuras .....	vi
Resumen .....	vii
Abstract .....	viii
<b>I. INTRODUCCIÓN.....</b>	<b>1</b>
<b>II. MARCO TEÓRICO.....</b>	<b>6</b>
<b>III. METODOLOGÍA.....</b>	<b>20</b>
3.1 Tipo de investigación .....	21
3.2 Variable y operacionalización.....	21
3.3 Población, muestra y muestreo.....	22
3.4 Técnicas e instrumentos de recolección de datos .....	22
3.5 Procedimiento .....	23
3.6 Método de análisis de datos .....	23
3.7 Aspectos éticos.....	24
<b>IV. RESULTADOS .....</b>	<b>25</b>
<b>V. DISCUSIÓN .....</b>	<b>36</b>
<b>VI. CONCLUSIONES .....</b>	<b>40</b>
<b>VII. RECOMENDACIONES .....</b>	<b>42</b>
<b>REFERENCIAS .....</b>	<b>44</b>
<b>ANEXOS .....</b>	<b>52</b>

## Índice de tablas

Tabla 1: Matriz de confusión .....	17
Tabla 2: Tabla de calificaciones.....	26
Tabla 3: Matriz de confusión – árbol de decisión .....	26
Tabla 4: Matriz de observación – árbol de decisión .....	26
Tabla 5: Matriz de confusión - SVM .....	27
Tabla 6: Matriz de observación - SVM .....	27
Tabla 7: Matriz de confusión - KNN .....	27
Tabla 8: Matriz de observación - KNN .....	27
Tabla 9: Tabla cruzada – cálculo de precisión con algoritmo árbol de decisión .....	28
Tabla 10: Tabla cruzada – cálculo de precisión con algoritmo SVM .....	299
Tabla 11: Tabla cruzada – cálculo de precisión con algoritmo KNN.....	29
Tabla 12: Cuadro comparativo de resultados según el indicador precisión .....	29
Tabla 13: Tabla cruzada – cálculo de sensibilidad con árbol de decisión.....	30
Tabla 14: Tabla cruzada – cálculo de sensibilidad con algoritmo SVM .....	30
Tabla 15: Tabla cruzada – cálculo de sensibilidad con algoritmo KNN .....	31
Tabla 16: Cuadro comparativo de resultados según el indicador sensibilidad.....	31
Tabla 17: Tabla cruzada – cálculo de especificidad con árbol de decisión.....	322
Tabla 18: Tabla cruzada – cálculo de especificidad con algoritmo SVM .....	32
Tabla 19: Tabla cruzada – cálculo de especificidad con algoritmo KNN .....	32
Tabla 20: Cuadro comparativo de resultados según el indicador especificidad.....	33
Tabla 21: Resumen de cuadro comparativo de algoritmos .....	33
Tabla 22: Medida de Kappa de Cohen – Árbol de decisión.....	34
Tabla 23: Medida de Kappa de Cohen – SVM .....	34
Tabla 24: Medida de Kappa de Cohen – K-NN .....	34

## Índice de gráficos y figuras

Figura 1: Ejemplo de clasificación con SVM .....	13
Figura 2: Kernel Lineal, Kernel polinómico y Gaussian kernel respectivamente.....	14
Figura 3: Estructura de árbol de decisión.....	14
Figura 4: Modelo de Neurona Artificial .....	15
Figura 5: Etapas de la metodología KDD .....	16
Figura 6: Fórmula para calcular la sensibilidad .....	17
Figura 7: Fórmula para calcular la especificidad .....	18
Figura 8: Fórmula para calcular la especificidad .....	18
Figura 9: Diseño pre-experimental con un solo grupo.....	21
Figura 10: Nivel de Kappa de Cohen .....	28
Fuente: Manterola (2018) .....	28
Figura 11: Cuestionario utilizando Google Forms .....	65
Figura 12: Definición de variables – SPSS Statistics .....	66
Figura 13: Datos importados de Google Forms a Excel.....	69
Figura 14: Datos importados de Excel a SPSS Statistics.....	69
Figura 15: Cuantificación de ítems no contestados parte 1 - SPSS Statistics .....	70
Figura 16: Cuantificación de ítems no contestados parte 2 - SPSS Statistics .....	70
Figura 17: Cuantificación de ítems no contestados parte 3 - SPSS Statistics .....	71
Figura 18: Carga de datos con nodo origen - SPSS Modeler.....	71
Figura 19: Transformación de datos con nodo derivar - SPSS Modeler.....	72
Figura 20: Ejemplo de la transformación de datos - SPSS Modeler.....	72
Figura 21: Aplicación del nodo tipo - SPSS Modeler.....	73
Figura 22: Ejemplo del nodo tipo - SPSS Modeler .....	73
Figura 23: Aplicación del nodo filtro - SPSS Modeler.....	73
Figura 24: Algoritmos de aprendizaje automático - SPSS Modeler .....	74
Figura 25: Proyecto de rendimiento académico - SPSS Modeler.....	74
Figura 26: Precisión de modelo utilizando árbol de decisión - SPSS Modeler.....	75
Figura 27: Variables con relevancia utilizando árbol de decisión - SPSS Modeler .....	75
Figura 28: Árbol de decisión – SPSS Modeler .....	76
Figura 29: Precisión de modelo utilizando SVM - SPSS Modeler.....	77
Figura 30: Precisión de modelo utilizando K vecinos - SPSS Modeler .....	77

## Resumen

En el presente trabajo de investigación se elaboró un modelo Machine Learning para predecir el rendimiento académico de los estudiantes universitarios, se utilizó la metodología KDD, así mismo herramientas como SPSS statistic y SPSS Modeler para la creación del modelo predictivo.

El objetivo de esta investigación es determinar en qué porcentaje Machine Learning permite predecir el rendimiento académico con precisión, sensibilidad y especificidad, con el fin de poder identificar a los alumnos con probabilidad de éxito o fracaso.

En esta investigación se utilizó una población de 87 alumnos, así mismo se usó la totalidad de la población como muestra. Por otro lado, el estudio es de tipo aplicada, con un diseño de investigación experimental de tipo pre-experimental de un solo grupo, ya que luego de aplicar Machine Learning se podrá observar los resultados y realizar la medición.

Como resultado en relación a la precisión, sensibilidad y especificidad para los algoritmos de árbol de decisión, Máquina de vectores y K-NN, se valida que Machine Learning para predecir el rendimiento académico de los estudiantes universitarios, así mismo el algoritmo con mejores resultados para esta casuística fue Máquina de vectores (SVM) con un valor de 100%.

Palabras clave: Machine Learning, rendimiento académico, métricas de precisión.

## Abstract

In the present research work, a Machine Learning model was developed to predict the academic performance of university students, the KDD methodology was used, as well as tools such as SPSS statistic and SPSS Modeler for the creation of the predictive model.

The objective of this research is to determine in what percentage Machine Learning allows to predict academic performance with precision, sensitivity and specificity, in order to be able to identify students with probability of success or failure.

In this research a population of 87 students was used, likewise the entire population was used as a sample. On the other hand, the study is of an applied type, with a pre-experimental experimental research design of a single group, since after applying Machine Learning, the results can be observed and the measurement carried out.

As a result in relation to the precision, sensitivity and specificity for the decision tree algorithms, Vector Machine and K-NN, it is validated that Machine Learning to predict the academic performance of university students, as well as the algorithm with the best results for This casuistry was Vector Machine (SVM) with a value of 100%.

Keywords: Machine Learning, academic performance, precision metrics.



## **I. INTRODUCCIÓN**

Desde hace mucho tiempo atrás, uno de los problemas controversiales en la educación es el rendimiento académico, con ello nos referimos principalmente al éxito o fracaso del estudiante. Teniendo en cuenta que los alumnos son parte fundamental de las instituciones educativas y sobre todo del país, nace la necesidad de poder identificar con anticipación aquellos alumnos que tendrán un periodo educativo bueno o malo con la finalidad de que la universidad pueda brindar alternativas para mejorar.

Partiendo del concepto de rendimiento académico muchos autores entre los cuales tenemos a Jiménez (como se citó en Navarro, 2014, párr.11) lo define como el “nivel de conocimiento demostrado por el alumno en un área o materia comparado con la norma de edad y nivel académico”, así mismo Sánchez (como se citó en Garbanzo, 2012, párr. 16) indica que el rendimiento académico es un conjunto de diversos factores involucrados con el alumno y se mide bajo las calificaciones obtenidas, ello demuestra la aprobación de cursos, la deserción y el grado de éxito.

En el ámbito internacional, según un estudio realizado por la fundación BBVA (2019) en España, cerca del 33 % de alumnos matriculados en un determinado periodo no logran culminar el grado, un 21 % abandona la universidad sin terminar sus estudios y cerca del 12 % prefiere cambiarse de carrera universitaria. También indica que puede deberse a varios factores y que se da principalmente en los primeros años del grado académico y también en los cursos posteriores. Según el estudio estas cifras son alarmantes por las pérdidas económicas de 400 millones de euros. Bajo lo mencionado, se puede concluir que muchas veces las universidades no están comprometidas con el alumno al grado de orientarlos u ofrecer programas que permitan culminar la universidad de manera satisfactoria como también existen alumnos que, teniendo la oportunidad de ser mejores, no existe interés. En Chile, el Servicio de Información de Educación Superior (2016), en su informe “Avance curricular en educación superior” menciona que cerca del 50 % de los estudiantes matriculados en una carrera profesional no logran culminarlo y casi se pierde la mitad de los alumnos en 4 años esto debido al bajo rendimiento académico dado por factores personales, familiares e institucionales.

En el ámbito nacional, Chilca (2017) en su investigación “Autoestima, hábitos de estudio y rendimiento académico en estudiantes universitarios” analizó a los estudiantes de Ingeniería de la Universidad Tecnológica del Perú, donde se pudo comprobar en base a registros académicos que un 47.8 % de los alumnos durante el año 2016 - II obtuvieron notas desaprobatorias, verificando que el promedio general de notas fue 10.68. Por otro lado, Orihuela (2019) en su estudio, presenta estudios sobre la Universidad Continental indicando que el 33.1% de los alumnos logran obtener un promedio de nota menor o igual a 10.5 de manera presencial mientras que bajo la modalidad virtual cerca al 55.7% obtienen notas menores o iguales a 10.5. En conclusión, estas cifras porcentuales son de preocupación puesto que representan casi la mitad de alumnos matriculados con bajo rendimiento académico y de alguna manera representan pérdidas económicas y de prestigio para la Universidad, por lo cual es de importancia poder identificar y brindar opciones que permitan ayudar al alumno con sus metas y objetivos académicos.

La presente investigación se realizó en la Universidad Nacional, creada en el año 1984 en el distrito de Nuevo Chimbote en el departamento de Ancash. En la actualidad cuenta con la facultad de ingeniería, ciencia y educación y humanidades. Teniendo en cuenta que las instituciones de nivel superior son entidades de enseñanza, de mejora cognitiva y además de ayuda en la inserción laboral, se da el caso de que la universidad no monitorea el rendimiento académico de sus estudiantes, por tanto, se desconoce a aquellas personas con riesgo de fracaso o éxito. Para el estudio se realizó un cuestionario a un total de 87 personas que pertenecen a la carrera de Ingeniería de Sistemas e Informática, verificando que el 51.7% presenta calificaciones de hasta 13, el 43.7% notas de hasta 16 y 4.6% notas hasta 18. De los resultados obtenidos se comprueba el predominio de notas bajas y por ende la dificultad que tienen los alumnos para el aprendizaje. Ante esta situación nace la interrogante ¿Qué ocurrirá si esta problemática continúa?, la respuesta es bastante evidente ya que si no se identifica a los estudiantes con dificultades puede traer consecuencias como la deserción estudiantil y pérdidas económicas en la universidad. Poder identificar con alta precisión y anticipación el desempeño del estudiante ayudará a la

institución a implementar acciones de mejora que ayuden a persistir y lograr los objetivos académicos.

Debido al panorama presentado se considera la siguiente problemática general: ¿En qué medida Machine Learning permitirá predecir el rendimiento académico de los estudiantes universitarios? Como problemas específicos, en primer lugar: ¿En qué medida Machine Learning permite predecir con precisión el rendimiento académico de los estudiantes universitarios? En segundo lugar: ¿En qué medida Machine Learning permite predecir con sensibilidad el rendimiento académico de los estudiantes universitarios? Y, en tercer lugar: ¿En qué medida Machine Learning permite predecir con especificidad el rendimiento académico de los estudiantes universitarios?

Este estudio contó con justificación teórica, puesto que se comparan algoritmos de aprendizaje automático para predecir el rendimiento académico con alta precisión, de tal manera que ayude a identificar a los alumnos en riesgo de fracaso, así mismo se contrasta resultados con otras investigaciones. También se justifica de manera social, puesto que se está contribuyendo con la población estudiantil, al identificar el rendimiento académico, la universidad puede brindar distintas opciones para culminar el desarrollo profesional y en un futuro estos puedan tener una mejor calidad de vida. Además, se justifica de manera tecnológica puesto que se está creando un modelo predictivo con el cual se puede construir un software de apoyo institucional para identificar el rendimiento académico.

Por lo explicado se plantea el siguiente objetivo general, Aplicar Machine Learning para predecir el rendimiento académico de los estudiantes universitarios. Como objetivos específicos; en primer lugar: Determinar en qué porcentaje Machine Learning permite predecir con precisión el rendimiento académico de los estudiantes universitarios. En segundo lugar: Determinar en qué porcentaje Machine Learning permite predecir con sensibilidad el rendimiento académico de los estudiantes universitarios y, por último, Determinar en qué porcentaje Machine Learning permite predecir con especificidad el rendimiento académico de los estudiantes universitarios. Estos objetivos permiten plasmar la siguiente hipótesis general: El Machine Learning permite predecir el rendimiento académico de los

estudiantes universitarios. Las hipótesis específicas planteadas son: El Machine Learning permite predecir con precisión el rendimiento académico de los estudiantes universitarios, también El Machine Learning permite predecir con sensibilidad el rendimiento académico de los estudiantes universitarios y finalmente, El Machine Learning permite predecir con especificidad el rendimiento académico de los estudiantes universitarios.

## **II. MARCO TEÓRICO**

Para el desarrollo del estudio se tuvo en cuenta trabajos previos internacionales y nacionales, así como también teorías y enfoques conceptuales los cuales se procede a detallar.

Según Chahuan (2019), titulada "Prediction of Student's Performance Using Machine Learning". El objetivo principal de este trabajo de investigación fue crear una herramienta de aprendizaje automático que permita predecir el GPA (Promedio de calificaciones) del estudiante en base a datos pasados (2015-2019) en el curso de Ciencias de la Computación. Las variables que tomaron en cuenta fueron las calificaciones de teoría, calificaciones de prácticas; luego se aplicó técnicas de regresión como KNN, árbol de decisión, SVM, random forest y regresión lineal. Para finalizar se concluyó que la técnica de regresión lineal múltiple es más óptima con un error cuadrático igual a 0.040, error cuadrático medio de 0.2, error absoluto de 0.149 y R-cuadrado igual 0.940. La relevancia de esta investigación es la comparación que se realizó entre técnicas para obtener un modelo con mejor precisión, en este caso regresión lineal múltiple.

Así mismo, Canagareddy (2019), "A Machine Learning Model to Predict the Performance of University Students". La problemática identificada fue que el número de estudiantes que se matriculaban no coincidía con el número de graduados, esto debido a que con mayor frecuencia los estudiantes repiten el año o varios módulos. Como objetivo principal se creó un modelo predictivo que permita pronosticar el rendimiento de los estudiantes con la finalidad de que se puedan tomar acciones correctivas. En busca del mejor modelo, se utilizó algoritmos de clasificación como Naive Bayes, Logistic Classifier y J48 Classifier; también algoritmos de predicción como SVM, Random forest y Logic Regression. El análisis se hizo en base a 2000 registros estudiantiles que es equivalente al tamaño de una facultad en la Universidad de Mauricio, concluyendo que el mejor algoritmo de clasificación es J48, ya que su precisión de 100 % no varía para 50, 100, 200 y 400 registros y un porcentaje de error en promedio menor a 0.05 %. Por otro lado, el algoritmo de predicción más eficiente es Random Forest. De esta investigación se rescata la utilización de la herramienta Weka para aplicar algoritmos de clasificación y precisión.

Para Alsaman (2019), "Using Decision Tree and Artificial Neural Network to Predict Students Academic Performance", el cual tuvo como objetivo crear un modelo de clasificación para predecir el rendimiento académico, en esta investigación se utilizó la metodología CRISP-DM y se realizó una comparación de mejores resultados o precisión en base a dos algoritmos de clasificación: árbol de decisión J48 y redes neuronales MLP (perceptrón multicapa) proporcionados por la herramienta WEKA. Para la recolección de datos se utilizó un cuestionario de elaboración propia los cuales fueron aplicados a 524 estudiantes de universidades privadas y públicas de Jordania. Concluyendo que la red neuronal es mucho más preciso con un 97 % aplicando cross-validation, mientras que un árbol de decisión 66%, así mismo se pudo comprobar que varios atributos como el tiempo de reprobación de los cursos y la dependencia de internet tienen impacto en la predicción, el estado laboral y civil no influyen, mientras que "beca" y "sostén de la familia" tuvieron un grado de eficacia en la predicción. De esta investigación se tomará como referencia la utilización de la técnica de Cross-validation para mejorar la precisión del modelo.

También Burman (2019) con su investigación titulada "Predicting Students Academic Performance Using Support Vector Machine", su objetivo fue ayudar a los estudiantes a mejorar su rendimiento académico con el uso de aplicaciones basadas en minería de datos, para lo cual creó un modelo predictivo basado en SVM (Máquina de Vectores de Soporte) para clasificar a los alumnos en 3 categorías: alto, medio y bajo. La metodología propuesta por el autor para el desarrollo consta de 6 pasos: input data set, uso de clasificación SVM, etapa de entrenamiento usando Linear Kernel y Radial Basis Kernel, etapa de testeo al modelo y etapa de estudio comparativo. La recolección de datos se realizó a través de un cuestionario basado en parámetros psicológicos, motivación, psicosocial, estrategias de aprendizaje, enfoque de aprendizaje y situación socioeconómica, obteniendo como resultado un total de 1000 registros para el análisis, de los cuales el 70% se utilizó para entrenamiento y el 30% para validación o testeo. Para validar el modelo se evaluó la sensibilidad, especificidad y precisión, obteniendo mejores resultados la función de base radial con un 90% sobre el núcleo lineal con un 64%. De esta investigación se utilizará como marco



de referencia la aplicación de SVM y los indicadores para evaluar el modelo predictivo.

Por otra parte, el autor Candia (2019) en su investigación titulada “Predicción del rendimiento académico de los estudiantes de la UNSAAC a partir de sus datos de ingreso utilizando algoritmos de aprendizaje automático”. Para el desarrollo del modelo predictivo se utilizó la metodología CRISP-DM, la herramienta weka y se consideró factores demográficos y educativos. Los datos para el análisis son de los estudiantes a partir del año 2014-I hasta 2018-I de la Universidad Nacional de San Antonio Abad del Cusco siendo un total de 12698 alumnos. La investigación es cuantitativa de tipo correlacional y no experimental. Así mismo, los algoritmos que se utilizaron fueron: árbol de decisión J48 se tuvo una precisión de 67.3%, para el algoritmo random forest 69.4%, para el algoritmo vecinos más cercanos 63.8%, para función logística 68% y por ultimo para el perceptron multicapas un 68%. De esta investigación se tendrá en cuenta ciertos conceptos definidos y correctamente estructurados para la fundamentación de factores asociados al rendimiento académico.

Para Vega (2019) en su trabajo de investigación “Modelo de pronóstico de rendimiento académico de alumnos en los cursos del programa de estudios básicos de la Universidad Ricardo Palma usando algoritmos de Machine Learning”. Tiene como finalidad pronosticar la cantidad de alumnos aprobados y desaprobados mediante el uso de técnicas de Machine Learning, esta investigación tiene un enfoque cuantitativo y es de tipo aplicado, la población es representada por los estudiantes desde el ciclo 2015 – I hasta el ciclo 2019 – 0, siendo un total de 9118 alumnos y 574,283 calificaciones. El método que se utilizó fue la metodología CRISP-DM y se realizaron comparaciones entre tres algoritmos; Redes neuronales, GBM y XGBoosting; dando como resultado una mayor precisión para el algoritmo XGBoosting con un porcentaje 0.89%. De esta investigación se tomará en cuenta el algoritmo con mayor precisión para realizar una comparación.

También, Hamoud (2018) con la investigación “Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis”. El objetivo principal fue la implementación de un modelo predictivo para buscar el factor que afecta al

éxito o fracaso del estudiante, esta investigación utilizó técnicas de minería de datos específicamente árbol de decisión. Para la recolección de datos se realizó una encuesta con 60 preguntas relacionadas a la salud, actividad social, la relaciones y el rendimiento académico, las cuales fueron tomadas a 161 alumnos. Para finalizar luego de comparar resultados entre algoritmos como J48, Random Tree y RepTree, se pudo comprobar que J48 tuvo mayor precisión de 0.629 y una TP RATE (tasa de verdadero positivo) igual a 0.634 mientras que RECALL (casos positivos) con una tasa de 0.634 y una tasa FP (caso positivo pero predicho negativo) igual a 0.409. De esta investigación se tendrá en cuenta el método para recolectar datos de los estudiantes (cuestionario), así mismo se tendrá en cuenta la influencia de los factores como la edad, trabajo, genero, etapa y estado influyen en la precisión de manera negativa y finalmente para mejorar la precisión, la aplicación de la técnica cross-validation.

De modo similar, Segura (2018) en su investigación "Using Decision Trees For Predicting Academic Performance Based On Socio-Economic Factors" tuvo como objetivo determinar de qué manera los factores socioeconómicos afectan a la educación, para ellos se consideró datos socioeconómicos y académicos, luego se aplicó algoritmos de clasificación y técnicas de aprendizaje automático. En total se recolectaron 1115.445 registros académicos y 19069 registros socioeconómicos, los resultados del algoritmo de árbol de decisión mostró una precisión de 59.85% y de árboles con degradado una precisión de 67.41%, se observó que ambos tienen indicadores en común que ayudan en la predicción de rendimiento académico como: Beca académica, edad actual y edad del estudiante cuando comenzaron el proyecto, un punto importante de mencionar es que se utilizó el indicador precisión para validar el modelo predictivo. Del estudio, se concluyó que los factores socioeconómicos no guardan ninguna relación ni influyen en el rendimiento académico de los estudiantes. La presente investigación servirá para considerar los indicadores relevantes y las recomendaciones de utilizar factores como psicológicos y emocionales en la definición del cuestionario para la recolección de data.

De igual manera Chiheb (2017) en su investigación "Predicting students' performance using decision trees: Case of an Algerian University", tuvo como

objetivo la creación de un modelo de clasificación y predictivo basado en técnicas de minería de datos para identificar a los estudiantes de buen desempeño, así como también ayudar a los alumnos graduados a elegir su maestría según los resultados que obtuvo sobre la disciplina que se ajusta. Para el estudio se consideró la metodología CRISP-DM y la recolección de datos de dos departamentos (Departamento de matemáticas e Informática, Ciencias y Departamento de Ciencias de la Computación) de la Universidad de Jijel, estudiantes entre los años 2009-2010 y 2014-2015, luego de pasar por la etapa de preparación de datos se empleó la técnica de árbol de decisión específicamente el algoritmo J48. Para finalizar los resultados obtenidos indican que se logró una tasa de clasificación correcta de 79.55% para la variable “éxito” para el nivel Licenciatura, 55% para Master SIAD y 100% para Master R&S; para la variable “apreciación” el 44.15% nivel Licenciatura, 58.3% Master SIAD y 100% Master R&S; finalmente con respecto a la variable “diploma” se logró un 58.81% para el nivel Licenciatura, 83.33% para Master SIAD y 100% para Master R&S. De esta investigación se rescata la utilización de la metodología CRISP-DM y la utilización del algoritmo J48, los cuales servirán para realizar una comparación de metodología y algoritmo de clasificación.

En el trabajo de investigación de Soto (2015), que tiene como título “Uso de técnicas de Machine Learning para predecir el rendimiento académico de los estudiantes de la carrera de Ingeniería Civil en Informática de la Universidad del Bío Bío, Chillan”. Tuvo como objetivo principal la creación de una aplicación que permita predecir el rendimiento académico. Para la obtención de datos se realizó una encuesta con 33 preguntas a estudiantes que ingresaron desde el año 2013 hasta 2015, se consideraron factores validados por otras investigaciones como auto concepto, motivación, enfoque de aprendizaje, factores socio-culturales, educación de los padres, inteligencia emocional, autoestima, factores económicos, factores emocionales, interés en la carrera, escuela de origen, determinantes personales, puntaje de acceso a la universidad, motivos para estudiar y expectativas laborales. Con respecto al algoritmo de aprendizaje se utilizó k Nearest Neighbors, dando una efectividad máxima de 41% con un índice de error de 1.24 en la predicción del rendimiento académico para el curso de introducción a la programación y un 60% de efectividad con un índice de error de

0.4 para el curso de programación orientada a objetos. De esta investigación de tomará en cuenta la técnica de recolección de datos y los factores asociados al rendimiento académico.

Para un adecuado respaldo de la investigación, se han tomado referencias teóricas, entre las cuales se explicará sobre el Machine Learning o aprendizaje automático, algoritmos supervisado como por ejemplo SVM, arboles de decisión, K – vecinos y redes neuronales; la metodología KDD, métricas de precisión para evaluar el modelo predictivo y finalmente sobre el rendimiento académico.

Machine Learning, fue definido por primera vez por Arthur Samuel (1952) como "Campo de estudio que brinda a las computadoras la capacidad de aprender sin estar programadas explícitamente", mientras que para Tom Mitchel (1997) "es el estudio de algoritmos informáticos que mejoran automáticamente a través de la experiencia". Se puede concluir que el aprendizaje automático es parte de la inteligencia artificial y tiene como objetivo la autonomía en el aprendizaje.

Según Hinestroza (2018, p. 4), el aprendizaje automático se da de tres tipos, a través de algoritmos supervisados; en el cual se utilizan etiquetas que permitan clasificar a los datos con la finalidad de identificar patrones y luego poder utilizarlo en otros grupos de entrada. Así mismo el autor menciona que existen algoritmos de aprendizaje no supervisado, en el cual no se utiliza etiquetas, sino, se obtiene conocimiento a partir de la búsqueda de patrones de los datos de entrada; por tanto, se podría decir que el conocimiento se da sin la intervención humana a través de un proceso de clusterización con los datos de entrada. Finalmente, el aprendizaje de refuerzo el autor menciona: "es cuando los datos no están etiquetados, pero después realiza varias acciones y de cierto periodo, el sistema será retroalimentado mediante actualizaciones"; por tanto, se podría decir que el sistema aprende a través de prueba y error, de la retroalimentación del análisis de datos. Adicionalmente existe un enfoque a considerar, es decir Deep Learning e IBM lo define como "el método que utiliza redes neuronales por capas para con los datos no estructurados", en otras palabras, trata de imitar el comportamiento del cerebro humano.

Para esta investigación se utilizó algoritmos de tipo supervisado, puesto que se empleó etiquetas para definir el rendimiento académico en categorías. En base a los trabajos previos descritos, se observó el predominio de árbol de decisión obteniendo buenos resultados, así mismo de acuerdo al autor Uskov (2019, p. 1371) en su estudio “Machine Learning – based Predictive Analytics of Student Academic Performance in STEM Education” quien realizó un análisis de los algoritmos a utilizar para el análisis académico bajo los siguientes aspectos: el enfoque, si es paramétrico, la velocidad de entrenamiento, velocidad de predicción y características automáticas, ver Anexo 4. De acuerdo a ello se decidió utilizar el algoritmo SVM (Maquinas de vectores de soporte) principalmente y se realizará una comparación con el algoritmo árbol de decisión, y K – NN.

Según Carmona (2016, p. 1). Las SVM son algoritmos inicialmente pensados para problemas de clasificación; sin embargo, en la actualidad son utilizadas para problemas de regresión, agrupamiento y multi-clasificación. El autor menciona “Las SVM pertenecen a los clasificadores lineales, puesto que inducen separadores lineales, también llamados hiper-planos”.

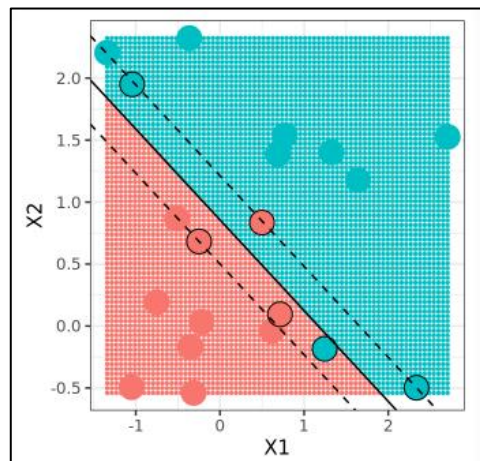


Figura 1: Ejemplo de clasificación con SVM

Fuente: Joaquín (2017)

Joaquín (2017) menciona que se obtienen buenos resultados cuando las clases tienen una separación lineal, de lo contrario es necesaria la utilización de kernel, que es principalmente crear una dimensión con la finalidad de encontrar un

hiperplano. El autor menciona “Kernel es una función que devuelve el resultado del dot product entre dos vectores realizado en un nuevo espacio dimensional distinto al espacio original”. Alguno de los kernel más utilizados son: Kernel lineal, kernel polinómico y gaussian kernel (RBF).

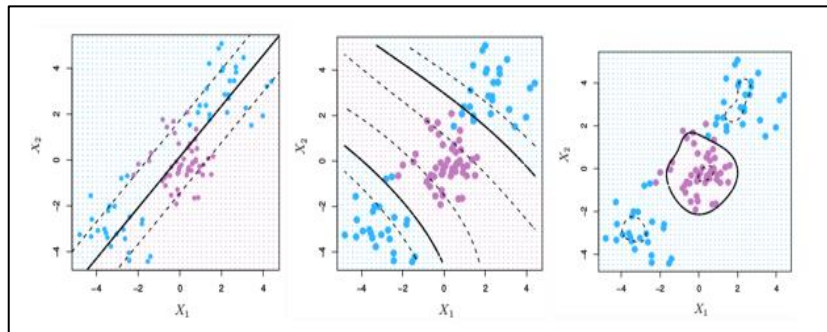


Figura 2: Kernel Lineal, Kernel polinómico y Gaussian kernel respectivamente  
Fuente: Joaquín (2017)

Por otro lado, el algoritmo de árbol de decisión según Charris (2018, p. 1), “Su objetivo principal es el aprendizaje inductivo a partir de observaciones y construcciones lógicas”. En otras palabras, tiene una estructura similar a un diagrama de flujo, también se podría decir que tiene una similitud con un conjunto de reglas estructuradas. Un árbol está representado por un nodo raíz el cual hace referencia al atributo principal; este desencadena en nodos hijos, los cual representan reglas de decisión y por ultimo terminan en nodos hojas que son principalmente la decisión o resultado.

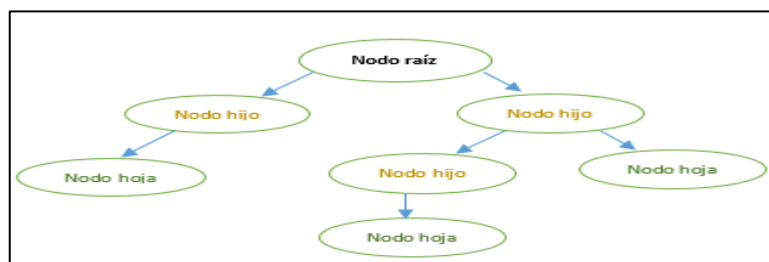


Figura 3: Estructura de árbol de decisión  
Fuente: Charris (2018)

Según el autor Rokash (2015), al ser un algoritmo de clasificación la característica de los datos recolectados para entrenamiento es de importancia, ya que servirán para aprender a tal punto de llegar a inferir las etiquetas de clase. Así mismo

Barrientos (2009, p. 22) menciona que los algoritmos de árbol de decisión más utilizados debido a la simplicidad, precisión y bajo costo de computo en su ejecución son los siguiente: ID3, emplea una búsqueda descendente con el objetivo de encontrar al mejor atributo, para lo cual es necesario gran cantidad de información (Garcia, 2005, p. 31). También se encuentra algoritmo J48, su particularidad es la creación de un árbol de manera iterativa, además puede utilizar atributos numéricos y vacíos para crear nuevas reglas (Leon, 2018, p.18). Finalmente, según Barrientos (2009, p 22), el algoritmo Naive Bayes, que es una técnica basada en estadística, que se caracteriza por asumir que los atributos son independientes, es decir que la característica en un conjunto de datos no guarda relación con otra característica presente.

Según Zapata (2014, p.221), el clasificador K vecinos más cercanos utiliza las características de un dato de entrada con la finalidad de validar si tiene relación con las características de su vecino más cercano, de ser verdadero acabará siendo de la misma clase. Para Quezada (2018, p.36), este algoritmo “es un método retardado y supervisado, puesto que la fase de entrenamiento y prueba se realizan en momentos diferentes”, así mismo menciona que el principal problema de este algoritmo es determinar el valor de k. Ante ello se han planteado reglas, la principal y la más básica es 1-NN.

Según García (2006, p.3) “Se basa en la suposición de que la clase del patrón a etiquetar, es decir X, es la del prototipo más cercano en R al que notaremos por  $X_{nn}$ ”. A continuación, un ejemplo de clasificación entre dos clases utilizando 1-NN. Según el autor la figura 6 muestra 4 prototipos de clase los cuales están representadas por cruces rojas y 5 prototipos de clase representadas por asteriscos de color azul, de ello se puede observar que el prototipo más cercano es de clase 2 por tanto X estará asociada a dicha clase.

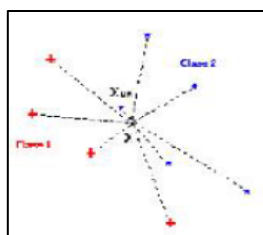


Figura 4: Modelo de Neurona Artificial

Fuente: Matich (2001)

Con relación a la metodología, se utilizará KDD (Knowledge Discovery in Database), Según el autor Fayyad (1996), “es el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia entendibles en los datos”, básicamente trata de explicar que hay una iteración entre etapas, una validación de los patrones a través de nuevos datos, novedoso en el sentido que se busca nuevo conocimiento y útil para poder tomar decisiones. Según Moine (2013, p. 11), “KDD es un proceso iterativo e interactivo”. Con iterativo se refiere a que se puede repetir en alguna etapa con la finalidad de obtener conocimiento de calidad y con interactivo porque el experto en el dominio tiene el fin de ayudar en la preparación de datos y validación de conocimiento. KDD consta de 5 etapas:

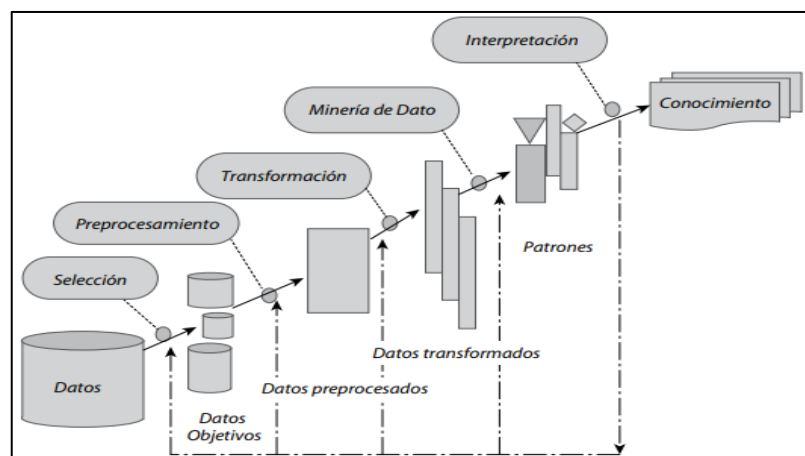


Figura 5: Etapas de la metodología KDD

Fuente: Reyes (2005)

La primera etapa es selección, luego de un entendimiento del problema y definido las metas del proceso, se crea un conjunto de datos sobre el cual se buscará conocimiento nuevo (Reyes, 2005, p. 65).

La segunda etapa es el pre-procesamiento / limpieza, básicamente se trata de una etapa de análisis de calidad de la data, donde se eliminan datos ruidosos, se utilizan estrategias para homogenizar datos desconocidos, datos nulos, duplicados, etc (Moine, 2013, p. 11).



La tercera etapa es transformación / reducción, tiene como objetivo eliminar variables no influyentes según la meta del proceso, para lo cual se utilizan técnicas de reducción para disminuir el número de variables (Fayyad 1996).

La cuarta etapa es minería de datos, de la vista minable generada en la etapa anterior se aplica técnicas con el fin de descubrir patrones o reglas (Reyes, 2005, p. 66). Finalmente, la etapa de interpretación, donde se evalúa el conocimiento descubierto para integrarlo con algún otro sistema probablemente (Reyes, 2005, p. 67).

Otro de los aspectos importantes en este trabajo es la validación del modelo predictivo, para lo cual se utilizará las métricas de precisión, sensibilidad y especificidad. Para AGGARWAL (2019, p. 497), la evaluación se da a través de una matriz de confusión, considerando los siguientes resultados:

Tabla 1: Matriz de confusión

		Predicted Class	
		Positive	Negative
Actual Class	Positive	(TP) True Positive	(FN) False Negative
	Negative	(FP) False Positive	(TN) True Negative

Fuente: Segura (2018)

- True Positives (TP): son las predicciones positivas que realmente son positivos para la clase.
- False Positives (FP): son las predicciones positivas que realmente son negativos para la clase.
- True Negatives (TN): son las predicciones negativas que realmente son negativas para la clase.
- False Negatives (FN): son las predicciones negativas que realmente son positivas

Según Burman (2019, p. 758), la sensibilidad “es una métrica estadística de logro que mide los valores positivos”

$$SENSIBILIDAD = \frac{TP}{TP + FN} \times 100$$

Figura 6: Fórmula para calcular la sensibilidad

Fuente: Burman (2019)

Según López (2015, p. 5), la especificidad “es La fracción de las instancias de la clase negativa que se clasifican correctamente.”

$$ESPECIFICIDAD = \frac{TN}{TN + FP} \times 100$$

Figura 7: Fórmula para calcular la especificidad

Fuente: Burman (2019)

Según Burman (2019, p. 758) accuracy “Es el sesgo estadístico lo que mide la veracidad, es decir, diferencia entre el valor observado y el valor real”.

$$PRECISIÓN = \frac{TP + TN}{TP + TN + FP + FN} \times 100$$

Figura 8: Fórmula para calcular la especificidad

Fuente: Burman (2019)

Otro de los aspectos importantes dentro de la presente investigación es el rendimiento académico y los factores que intervienen en la predicción. Para muchos autores el objetivo principal de cualquier universidad es formar estudiantes con un gran rendimiento académico, sin embargo, muchos son los factores que están inmersos en el proceso de formación, logrando el éxito o fracaso del estudiante (Puarungroj, 2018, p.1). Sobre lo mencionado anteriormente, Ramaphosa (2018, p.1), considera que el objetivo de las instituciones educativas es dar una enseñanza de calidad y para lograr ello es necesario analizar diversos factores.

Bajo dicha premisa autores como Hernandez (2013, p.20), conceptualiza al rendimiento académico como el resultado de factores que se relacionan y están asociados al estudiante. Así mismo, Navarro (2014, p. 3), define que el rendimiento académico guarda mucha relación con el proceso de evaluación, sin embargo, también es necesario tomar en cuenta factores externos que rodean al estudiante como por ejemplo el aula, plana docente, centro de estudio, entre otros.

Por otro lado, el autor Bohorquez (2013, p. 18) mencionan que el rendimiento académico es una métrica relacionada a la capacidad del estudiante y que

expresa los logros a lo largo del periodo formativo. Mientras que para el autor Navarro (2014, p.2) la manera de evaluar el rendimiento y contribuir con mejorarlo se realiza haciendo un estudio sobre los factores influyentes entre los cuales se encuentran los socioeconómicos, amplitud de los programas de estudio, metodologías de enseñanza, dificultad de enseñanza personalizada, conceptos previos y nivel de pensamiento de los alumnos. Por el contrario, Jimenez (2000), considera que un alumno puede tener las aptitudes y actitudes sin embargo no estar obteniendo un rendimiento adecuado.

También Navarro (2014, p.4), en su artículo expone que los factores con mayor relación hacia el rendimiento académico son la motivación escolar, el autocontrol y las habilidades sociales. Así mismo Segura (2018) en su trabajo de investigación considera que los factores socioeconómicos permiten evaluar o medir el rendimiento académico, estos a su vez esta divididos en factores demográficos, capital humano y estabilidad residencial. Para Shakil (2017) además de los factores socioeconómicos, también considera a los factores psicológicos y factores académicos dentro del estudio para determinar el rendimiento estudiantil.

Por tanto, de lo expuesto anteriormente por diversos autores, se puede mencionar que el rendimiento académico es multifactorial y existe gran cantidad de líneas de investigación para determinar cómo influyen, teniendo en cuenta la complejidad e importancia dentro del ámbito académico.

### **III. METODOLOGÍA**

### 3.1 Tipo de investigación

La presente investigación es de tipo aplicada, debido a que se utilizó la máquina de aprendizaje para resolver la predicción de rendimiento académico de los alumnos, la cual permite identificar al estudiante con un alto grado de precisión. Según Schubert (2017, p.5), “nace desde la práctica social y genera resultados que se pueden aplicar, pero estos no necesariamente acaban en producción, usualmente debido a los costos.” Para el autor Vargas (2009, p. 159), “se caracteriza porque busca la aplicación o utilización de los conocimientos adquiridos, a la vez que se adquieren otros, después de implementar y sistematizar la práctica basada en investigación”

Así mismo, el diseño de investigación es experimental de tipo pre-experimental, debido a que se tomó un grupo de alumnos representados por sus datos, los cuales fueron obtenidos a través de un cuestionario, para luego aplicar Machine Learning y posteriormente obtener una observación para realizar una medición. Según Hernández (2014, p 141) “los pre-experimentos consisten en administrar un estímulo o tratamiento a un grupo y después aplicar una medición de una o más variables para observar cual es el nivel del grupo en estas.” Para Bernal (2010, p. 153) “un diseño pre-experimental con un solo grupo, es un diseño sin grupo control y donde solo se efectúa una medición posterior”.

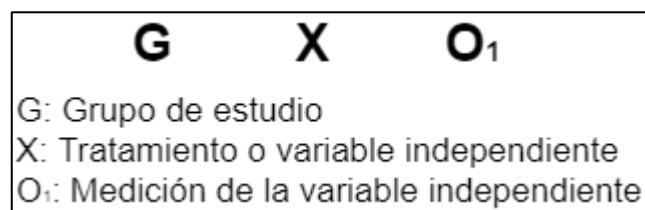


Figura 9: Diseño pre-experimental con un solo grupo  
Fuente: Bernal (2010)

**G:** Datos de los estudiantes de ingeniería de sistemas

**X:** Machine Learning

**O:** Métricas de precisión

### 3.2 Variable y operacionalización

La variable con las que cuenta la presente investigación son: Machine Learning, como variable independiente y como variable dependiente “predecir el rendimiento académico”, la variable dependiente consta de tres indicadores que son la exactitud, sensibilidad y especificidad. La operacionalización a detalle de estas variables se encuentran en el ANEXO N° 3.

### 3.3 Población, muestra y muestreo

Según Boddy (2016, p. 4321), el término población es definido como la agrupación de objetos o personas para recabar ciertos datos que sirvan en una investigación, esta puede ser de diversas formas y tipos desde; registros hasta animales. Para Tamayo (2004), “La población se define como la totalidad del fenómeno a estudiar donde la unidad de población posee una característica común la cual se estudia y da origen a los datos de la investigación”.

Por tanto, para esta investigación la población estuvo constituida por una base de datos que cuenta con un aproximado de 87 registros de estudiantes obtenidas a través de un cuestionario. Así mismo, para la muestra se tomó la totalidad de la población, debido a que los datos recolectados representan la cantidad mínima para realizar el experimento. El muestreo que se realizó es de tipo no probabilístico debido a que se tomó todos los datos.

### 3.4 Técnicas e instrumentos de recolección de datos

Según García (2003, p.2), “El cuestionario es un procedimiento considerado clásico en la ciencia social para la obtención y registro de datos. Su versatilidad permite utilizarlo como instrumento de investigación y como instrumento de evaluación de personas, procesos y programas de formación”. Así mismo para el autor Aigner (2009, p.6), “El cuestionario es un formulario con una lista de preguntas estandarizadas y estructuradas que se han de formular de idéntica manera a todos los encuestados”.

Por tal motivo, para esta investigación se estableció usar como técnica de recolección de datos, la realización de un cuestionario, el cual consta de 28 preguntas las cuales fueron desarrolladas con la herramienta google formulario y estuvieron relacionadas con determinantes personales (6 preguntas), auto-concepto (3 preguntas), la motivación (3 preguntas), factores socio-culturales (6

preguntas), educación de los padres (2 preguntas), inteligencia emocional (3 preguntas), factores económicos (2 preguntas), escuela origen (2 preguntas) y variable objetiva (1 pregunta), estas preguntas fueron diseñadas a partir de artículos relacionados al objeto de estudio, entre las cuales tenemos autores como Segura (2018), Hamoud (2018), Alsalman (2019). Así mismo se utilizó la ficha de observación para registrar la precisión del algoritmo el cual se encuentra en el ANEXO 5, 6, 7. Cabe mencionar que esta ficha de observación fue validada por expertos, esto se puede observar en el ANEXO 8, 9, 10, 11.

### 3.5 Procedimiento

Esta investigación tiene como fin utilizar Machine Learning para predecir el rendimiento académico de los estudiantes universitarios, siendo “predecir el rendimiento académico” la variable dependiente, para lo cual se realizó la búsqueda de investigaciones similares a nivel nacional e internacional con el objetivo de analizar la solución de ese momento. Por eso se estudió a detalle tanto la variable dependiente como independiente recolectando información de tesis y artículos científicos de tal manera que se obtenga antecedentes y bases teóricas de las cuales se obtienen dimensiones e indicadores sustentables.

Para este proyecto de investigación se recolectó datos de los estudiantes de Ingeniería de sistemas de una Universidad Nacional a través de un cuestionario creado en google formulario, en el ANEXO 12 se visualiza las preguntas consideradas para este estudio, luego se aplicó la metodología KDD que nos permitió crear una vista minable en donde se aplicó técnicas de aprendizaje automático como arboles decisiones, K-vecinos y máquinas de vectores (SVM).

Finalmente, para la validación de hipótesis en esta investigación, se utilizó la matriz de confusión que proporciona la herramienta SPSS Modeler para visualizar el desempeño del algoritmo, además de coeficiente de Kappa de Cohen para validar el nivel de concordancia y el grado de significancia.

### 3.6 Método de análisis de datos

Para el presente trabajo de investigación la data se obtuvo a través de un cuestionario realizado en un formulario virtual online (ANEXO 12) aplicados a los alumnos de Ingeniería de Sistemas de una Universidad Nacional, luego se realizó

la eliminación de los datos inconsistentes, también se hizo la transformación de la variable de rendimiento de valor cuantitativo a valor cualitativo haciendo uso de la herramienta SPSS statistic, así mismo se realizó un análisis predictivo haciendo uso de la herramienta SPS Modeler, el cual nos mostró información relacionado con la precisión, sensibilidad y especificidad, cabe mencionar que se está utilizando estadística predictiva.

### 3.7 Aspectos éticos

Para esta investigación se respetó la autoría de las fuentes obtenidas de revistas indexadas y repositorios de universidad, así mismo se respetó los aspectos relevantes del código de ética de investigación de la universidad. El artículo 15, 16 corresponde a las políticas anti plagio y los derechos de autor. Así mismo los artículos 37, 42, 44 del código de ética del colegio de ingenieros del Perú relacionado a la divulgación de información o la omisión de autor o coautor que intervienen en la investigación las cuales se están cumpliendo.

Por último, el estudiante se compromete a respetar la confiabilidad de la información recolectada sin realizar ninguna modificación que permita alterar la demostración de la hipótesis planteada, de tal manera el proyecto mostró la calidad y puede usar para proyectos futuros como referencia.



## **IV. RESULTADOS**

En este capítulo se detalla los resultados obtenido de la investigación, basándose en los indicadores de precisión, sensibilidad y especificidad, los cuales fueron comparados entre 3 algoritmos de aprendizaje para definir el mejor, la validación de la hipótesis se realizó utilizando el índice de Kappa de Cohen el cual muestra el nivel de concordancia como también el nivel de significancia. Por otro lado, el Ministerio de Educación (MINEDU), indica utilizar la siguiente escala para medir la capacidad del alumno durante el periodo educativo. Así mismo la Universidad de Granada en la elaboración de tabla de conversión de calificaciones con respecto a países de intercambio clasifica de la siguiente manera la evaluación en el Perú:

Tabla 2: Tabla de calificaciones

Excelente	19 – 20	5
Distinguido	17 – 18.9	4
Bueno	14 – 16.9	3
Regular	11 – 13.9	2
Reprobado	0 – 10.9	1

Fuente: Elaboración propia

A continuación, se presenta los resultados de la matriz de confusión obtenida a partir de la herramienta SPSS Modeler para los siguientes algoritmos:

- Árbol de decisión

Tabla 3: Matriz de confusión – árbol de decisión

Rendimiento	Predicción		
	2	3	4
2	40	5	0
3	4	34	0
4	0	4	0

Fuente: Elaboración propia

Tabla 4: Matriz de observación – árbol de decisión

Clase	Medidas			
	TP	TN	FP	FN
2	40	34	4	5
3	34	40	9	4
4	0	74	0	4

Fuente: Elaboración propia

Para un total de 87 registros, se identificó a 40 estudiantes con notas regulares de manera correcta y 5 de manera incorrecta, 34 estudiantes con notas buenas de manera correcta y 4 incorrectos, finalmente 4 predicciones incorrectas para estudiantes distinguidos y 0 predicciones para alumnos excelentes.

- Máquina de Vectores de soporte (SVM)

Tabla 5: Matriz de confusión - SVM

Rendimiento	Predicción		
	2	3	4
2	45	0	0
3	0	38	0
4	0	0	4

Fuente: Elaboración propia

Tabla 6: Matriz de observación - SVM

Clase	Medidas			
	TP	TN	FP	FN
2	45	42	0	0
3	38	49	0	0
4	4	83	0	0

Fuente: Elaboración propia

Para un total de 87 registros en la data de entrenamiento, se identificó a 45 estudiantes con notas regulares de manera correcta, 38 estudiantes con notas buenas de manera correcta, 4 estudiantes con notas distinguidas, finalmente 0 predicciones para estudiantes con notas excelentes

- K-vecinos (K-NN)

Tabla 7: Matriz de confusión - KNN

Rendimiento	Predicción		
	2	3	4
2	43	2	0
3	20	18	0
4	0	2	2

Fuente: Elaboración propia

Tabla 8: Matriz de observación - KNN

Clase	Medidas			
	TP	TN	FP	FN
2	43	20	20	2
3	18	45	4	20
4	2	61	0	2

Fuente: Elaboración propia

Para un total de 87 registros en la data de entrenamiento, se identificó a 43 estudiantes con notas regulares de manera correcta y 2 predicciones incorrectas, 18 estudiantes con notas buenas de manera correcta y 20 de manera incorrecta, 2 estudiantes con notas distinguidas correctamente y 2 incorrectos, finalmente 0 predicciones para estudiantes con notas excelentes

A continuación, se presentan los resultados obtenidos para las hipótesis específicas planteadas en esta investigación, a partir de la herramienta SPSS Statistics se aplicó el índice de Kappa de Cohen el cual nos muestra el nivel de concordancia a partir de dos observaciones. Según el autor Manterola (2018, p.261) “el índice de Kappa de Cohen es una forma de correlación y como todo

ellos, puede variar de -1 a +1, donde 1 representa la concordancia perfecta entre dos observadores”, así mismo el autor Salas (2019, p.2) lo define de la siguiente manera “es un valor que busca establecer el acuerdo que existe entre distintos evaluadores”

Valores	Interpretación
< 0,01	No acuerdo
0,01 - 0,20	Ninguna a escaso
0,21 - 0,40	Regular o razonable
0,41 - 0,60	Moderado
0,61 - 0,80	Substancial
0,81 - 1,00	Casi perfecto

Figura 10: Nivel de Kappa de Cohen

Fuente: Manterola (2018)

**HE1:** El Machine Learning permite predecir con precisión el rendimiento académico de los estudiantes universitarios.

- Árbol de decisión

Tabla 9: Tabla cruzada – cálculo de precisión con algoritmo árbol de decisión

		Predicción		Total	
		2	3		
Rendimiento	2	Recuento	40	5	45
		% del total	46,0%	5,7%	51,7%
	3	Recuento	4	34	38
		% del total	4,6%	39,1%	43,7%
	4	Recuento	0	4	4
		% del total	0,0%	4,6%	4,6%
Total		Recuento	44	43	87
		% del total	50,6%	49,4%	100,0%

Fuente: SPSS Statistics

$\text{PRECISIÓN} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100$	$= (74+148) / (74+174) \times 100$ $= (222/248) \times 100 = 89.51$
---	---

Interpretación: Machine Learning permite predecir el rendimiento académico de los estudiantes universitarios con una precisión igual a 89.51% utilizando el algoritmo de árbol de decisión.

- Máquina de vectores de soporte (SVM)

Tabla 10: Tabla cruzada – cálculo de precisión con algoritmo SVM

			Predicción			Total
			2	3	4	
Rendimiento	2	Recuento	45	0	0	45
		% del total	51,7%	0,0%	0,0%	51,7%
	3	Recuento	0	38	0	38
		% del total	0,0%	43,7%	0,0%	43,7%
	4	Recuento	0	0	4	4
		% del total	0,0%	0,0%	4,6%	4,6%
Total		Recuento	45	38	4	87
		% del total	51,7%	43,7%	4,6%	100,0%

Fuente: SPSS Statistics

$\text{PRECISIÓN} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100$	$= (87+174) / (87+174) \times 100$ $= (261 / 261) \times 100 = 100$
---	---

Interpretación: Machine Learning permite predecir el rendimiento académico de los estudiantes universitarios con una precisión igual a 100% utilizando el algoritmo de clasificación de Máquina de vectores (SVM).

- K-Vecinos (K-NN)

Tabla 11: Tabla cruzada – cálculo de precisión con algoritmo KNN

Tabla cruzada						
			Predicción			Total
			2	3	4	
Rendimiento	2	Recuento	43	2	0	45
		% del total	49,4%	2,3%	0,0%	51,7%
	3	Recuento	20	18	0	38
		% del total	23,0%	20,7%	0,0%	43,7%
	4	Recuento	0	2	2	4
		% del total	0,0%	2,3%	2,3%	4,6%
Total		Recuento	63	22	2	87
		% del total	72,4%	25,3%	2,3%	100,0%

Fuente: SPSS Statistics

$\text{PRECISIÓN} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100$	$= (63+126) / (63+174) \times 100$ $= (189 / 237) \times 100 = 79.75$
---	---

Interpretación: Machine Learning permite predecir el rendimiento académico de los estudiantes universitarios con una precisión igual a 79.75% utilizando el algoritmo K-vecinos (K-NN).

Tabla 12: Cuadro comparativo de resultados según el indicador precisión

ALGORITMO	RESULTADO (%)
Árbol de decisión	89.51 %
SVM	100 %
K-NN	79.75 %

Fuente: Elaboración propia

Interpretación: En la tabla 12 se observa que el algoritmo con mejor resultado en cuanto al indicador precisión que evalúa el porcentaje de elementos clasificados correctamente es la máquina de vectores (SVM) con un 100%, seguido del algoritmo árbol de decisión con un resultado igual a 89.5% y K-vecinos (K-NN) con un 79.75%

**HE2:** El Machine Learning permite predecir con sensibilidad el rendimiento académico de los estudiantes universitarios.

- Árbol de decisión

Tabla 13: Tabla cruzada – cálculo de sensibilidad con árbol de decisión

			Predicción		Total
			2	3	
Rendimiento	2	Recuento	40	5	45
		% del total	46,0%	5,7%	51,7%
	3	Recuento	4	34	38
		% del total	4,6%	39,1%	43,7%
	4	Recuento	0	4	4
		% del total	0,0%	4,6%	4,6%
Total		Recuento	44	43	87
		% del total	50,6%	49,4%	100,0%

Fuente: SPSS Statistics

$\text{SENSIBILIDAD} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100$	$= \frac{74}{74+13} \times 100$ $= 85.05$
--	---

Interpretación: Machine Learning permite predecir el rendimiento académico de los estudiantes universitarios con una sensibilidad igual a 85.05% utilizando el algoritmo árbol de decisión.

- Máquina de vectores de soporte (SVM)

Tabla 14: Tabla cruzada – cálculo de sensibilidad con algoritmo SVM

			Predicción			Total
			2	3	4	
Rendimiento	2	Recuento	45	0	0	45
		% del total	51,7%	0,0%	0,0%	51,7%
	3	Recuento	0	38	0	38
		% del total	0,0%	43,7%	0,0%	43,7%
	4	Recuento	0	0	4	4
		% del total	0,0%	0,0%	4,6%	4,6%
Total		Recuento	45	38	4	87
		% del total	51,7%	43,7%	4,6%	100,0%

Fuente: SPSS Statistics

$\text{SENSIBILIDAD} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100$	$= \frac{87}{87} \times 100$ $= 100$
--	--------------------------------------

Interpretación: Machine Learning permite predecir el rendimiento académico de los estudiantes universitarios con una sensibilidad igual a 100% utilizando el algoritmo máquina de vectores de soporte (SVM).

- K-Vecinos (K-NN)

Tabla 15: Tabla cruzada – cálculo de sensibilidad con algoritmo KNN

Tabla cruzada			Predicción			Total
			2	3	4	
Rendimiento	2	Recuento	43	2	0	45
		% del total	49,4%	2,3%	0,0%	51,7%
	3	Recuento	20	18	0	38
		% del total	23,0%	20,7%	0,0%	43,7%
	4	Recuento	0	2	2	4
		% del total	0,0%	2,3%	2,3%	4,6%
Total		Recuento	63	22	2	87
		% del total	72,4%	25,3%	2,3%	100,0%

Fuente: SPSS Statistics

$\text{SENSIBILIDAD} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100$	$= \frac{63}{63+24} \times 100$ $= 72.41$
--	---

Interpretación: Machine Learning permite predecir el rendimiento académico de los estudiantes universitarios con una sensibilidad igual a 72.41% utilizando el algoritmo K-vecinos (K-NN).

Tabla 16: Cuadro comparativo de resultados según el indicador sensibilidad

ALGORITMO	RESULTADO (%)
Árbol de decision	85.05 %
SVM	100 %
K-NN	72.41 %

Fuente: Elaboración propia

Interpretación: En la tabla 16 se observa que el algoritmo con mejor resultado en cuanto al indicador sensibilidad, es decir casos verdaderos positivos es la máquina de vectores (SVM) con un 100%, seguido del algoritmo árbol de decision con un resultado igual a 85.05% y K-vecinos (KNN) con un 72.41%

**HE3:** El Machine Learning permite predecir con especificidad el rendimiento académico de los estudiantes universitarios.

- Árbol de decisión

Tabla 17: Tabla cruzada – cálculo de especificidad con árbol de decisión

			Predicción		Total
			2	3	
Rendimiento	2	Recuento	40	5	45
		% del total	46,0%	5,7%	51,7%
	3	Recuento	4	34	38
		% del total	4,6%	39,1%	43,7%
	4	Recuento	0	4	4
		% del total	0,0%	4,6%	4,6%
Total		Recuento	44	43	87
		% del total	50,6%	49,4%	100,0%

Fuente: SPSS Statistics

$\text{ESPECIFICIDAD} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100$	$= (148/148+13) \times 100$ $= 91.92$
---	---------------------------------------

Interpretación: Machine Learning permite predecir el rendimiento académico de los estudiantes universitarios con una especificidad igual a 91.92% utilizando el algoritmo árbol de decisión.

- Máquina de vectores de soporte (SVM)

Tabla 18: Tabla cruzada – cálculo de especificidad con algoritmo SVM

			Predicción			Total
			2	3	4	
Rendimiento	2	Recuento	45	0	0	45
		% del total	51,7%	0,0%	0,0%	51,7%
	3	Recuento	0	38	0	38
		% del total	0,0%	43,7%	0,0%	43,7%
	4	Recuento	0	0	4	4
		% del total	0,0%	0,0%	4,6%	4,6%
Total		Recuento	45	38	4	87
		% del total	51,7%	43,7%	4,6%	100,0%

Fuente: SPSS Statistics

$\text{ESPECIFICIDAD} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100$	$= (174/174) \times 100$ $= 100$
---	----------------------------------

Interpretación: Machine Learning permite predecir el rendimiento académico de los estudiantes universitarios con una especificidad igual a 100% utilizando el algoritmo de máquina de vectores (SVM).

- K-Vecinos (K-NN)

Tabla 19: Tabla cruzada – cálculo de especificidad con algoritmo KNN

Tabla cruzada						
			Predicción			Total
			2	3	4	
Rendimiento	2	Recuento	43	2	0	45
		% del total	49,4%	2,3%	0,0%	51,7%
	3	Recuento	20	18	0	38
		% del total	23,0%	21,1%	0,0%	27,1%



		% del total	23,0%	20,7%	0,0%	43,7%
	4	Recuento	0	2	2	4
		% del total	0,0%	2,3%	2,3%	4,6%
Total		Recuento	63	22	2	87
		% del total	72,4%	25,3%	2,3%	100,0%

Fuente: SPSS Statistics

$\text{ESPECIFICIDAD} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100$	$= (126/126+24) \times 100$ $= 84$
---	------------------------------------

Interpretación: Machine Learning permite predecir el rendimiento académico de los estudiantes universitarios con una especificidad igual a 84% utilizando el algoritmo K-vecinos (K-NN).

Tabla 20: Cuadro comparativo de resultados según el indicador especificidad

ALGORITMO	RESULTADO (%)
Árbol de decision	91.92 %
SVM	100 %
K-NN	84 %

Fuente: Elaboración propia

Interpretación: En la tabla 20 se observa que el algoritmo con mejor resultado en cuanto al indicador especificidad, es decir casos verdaderos negativos es la máquina de vectores (SVM) con un 100%, seguido del algoritmo árbol de decision con un resultado igual a 91.92 % y K-vecinos (K-NN) con un 84%

### Hipótesis General:

**Ho:** El Machine Learning no permite predecir el rendimiento académico de los estudiantes universitarios.

**Ha:** El Machine Learning permite predecir el rendimiento académico de los estudiantes universitarios.

Tabla 21: Resumen de cuadro comparativo de algoritmos

ALGORITMO	INDICADORES		
	PRECISIÓN	SENSIBILIDAD	ESPECIFICIDAD
Árbol de decisión	89.51 %	85.05 %	91.92 %
SVM	100 %	100 %	100 %
K-NN	79.75 %	72.41 %	84 %

Fuente: Elaboración propia

Frente a los resultados obtenidos con relación a la precisión, sensibilidad y especificidad para los algoritmos árbol de decisión, SVM y K-NN se valida que el algoritmo con mejores resultados para predecir el rendimiento académico es la máquina de vectores con un valor de 100 % en todos los indicadores.

Así mismo se utiliza el índice de Kappa de Cohen para evaluar la concordancia existente entre dos observaciones y el nivel de significancia. A continuación los resultados obtenidos por cada algoritmo.

Tabla 22: Medida de Kappa de Cohen – Árbol de decisión

Medidas simétricas					
		Valor	Error estándar asintótico	T aproximada	Significación aproximada
Medida de acuerdo	Kappa	,714	,070	7,151	,000
N de casos válidos		87			

Fuente: Elaboración propia

Interpretación: El valor de Kappa nos muestra un valor igual a 0.71, por tanto existe buena concordancia entre los valores observados en un primer momento y los valores predichos por el algoritmo árbol de decisión, además nos muestra un valor de significancia menor a 5%, por tanto se acepta la hipótesis alterna al utilizar el algoritmo de árbol de decisión.

Tabla 23: Medida de Kappa de Cohen – SVM

Medidas simétricas					
		Valor	Error estándar asintótico	T aproximada	Significación aproximada
Medida de acuerdo	Kappa	1,000	,000	10,521	,000
N de casos válidos		87			

Fuente: Elaboración propia

El valor de Kappa nos muestra un valor igual a 1, por tanto existe muy buena concordancia entre los valores observados en un primer momento y los valores predichos por el algoritmo SVM, además nos muestra un valor de significancia menor a 5%, por tanto se acepta la hipótesis alterna al utilizar el algoritmo SVM.

Tabla 24: Medida de Kappa de Cohen – K-NN

Medidas simétricas					
		Valor	Error estándar asintótico	T aproximada	Significación aproximada
Medida de acuerdo	Kappa	,463	,085	5,218	,000
N de casos válidos		87			

Fuente: Elaboración propia

El valor de Kappa nos muestra un valor igual a 0.46, por tanto existe una moderada concordancia entre los valores observados en un primer momento y los valores predichos por el algoritmo K-NN, además nos muestra un valor de significancia menor a 5%, por tanto se acepta la hipótesis alterna al utilizar el algoritmo K-NN.

En conclusión, el algoritmo con mejor nivel de concordancia luego de evaluar árbol de decisión, SVM y K-NN es la máquina de vectores con un 100% según la medida de acuerdo a Kappa de Cohen y un grado de significancia menor a 5%, por tanto se acepta la hipótesis alterna.

## **V. DISCUSIÓN**

A continuación, se muestran las discusiones realizadas en base a los resultados que se dieron durante la investigación

Para este estudio se utilizó la metodología KDD, un método que te indica que hacer mas no el cómo hacerlo, por sus características es mucho más ágil por ende toma menor tiempo la implementación. Autores como Chahuan (2019), Canagareddy (2019), Hamoud (2018), Segura (2018) y Soto (2015) utilizaron como referencia KDD para la extracción de conocimiento. Mientras que el autor Burman (2019) utilizó una metodología propia que consta de 6 pasos: input data set, uso de algoritmo, entrenamiento usando Linear Kernel y Radial Basis Kernel, testeo y estudio comparativo y finalmente autores como Alsalman (2019), Candia (2019), Vega (2019) y Chiheb (2017) quisieron estandarizar, es decir profundizar en las tareas y actividades en cada etapa por ello el usaron la metodología CRISP-DM.

Para desarrollar el modelo predictivo se utilizó como herramienta de modelado el SPSS Modeler por su interfaz amigable e interactivo, además por la documentación y comunidades de ayuda, esto permite realizar el desarrollo con mayor facilidad y rapidez, mientras que autores como Canagareddy (2019), Alsalman (2019), Candia (2019), Vega (2019), Hamoud (2018), Chiheb (2017) y Soto (2015) hicieron uso de la herramienta Weka y finalmente Segura (2018) prefirió utilizar RapidMiner.

Luego de realizar una comparación entre los algoritmos: árbol de decisión, máquina de vectores y K-NN se observó que el mejor modelo se creó utilizando máquina de vectores, ya que la sensibilidad, especificidad y precisión es de 100%, por tanto el más adecuado para la predicción del rendimiento académico de los alumnos. Otros autores por el contrario como Hamoud (2018), utilizaron otras métricas como recall, sensibilidad, FP Rate y precisión obteniendo como resultado 63%, 63%, 40% y 62% respectivamente utilizando arboles de decision j48. Otras investigaciones de los autores Chahuan (2019), Canagareddy (2019), Alsalman (2019), Candia (2019), Segura (2018), Chiheb (2017) y Soto (2015) solo consideraron como métrica de medición, la precisión. Otra de los puntos a tener en cuenta es la recolección de datos, pues en esta investigación se utilizó un cuestionario de 28 preguntas, considerando factores como personales, el auto-

concepto, motivación, socio-culturales, educación de los padres, inteligencia emocional, factores económicos, escuela origen y promedio. Estudios similares como el de los autores Alsalman (2019) consideró 19 preguntas para un total de 524 estudiantes, Burman (2019) utilizó factores psicológicos, motivación, psicosocial, estrategia de aprendizaje, enfoque de aprendizaje y socioeconómico, obteniendo un total de 1000 registros, Hamoud (2018) utilizó 60 preguntas relacionadas a la salud, actividad social, la relaciones y el promedio de notas obteniendo un total de 161 registros y Soto (2015) realizó un cuestionario de 33 preguntas a estudiantes de 2013 hasta el 2015. Por otro lado, autores como Chahuan (2019), Canagareddy (2019), Candia (2019), Vega (2019), Segura (2018), Chiheb (2017) decidieron tomar datos de los registros históricos de los alumnos de determinados años, esto permite tener data homogenizada y cantidad por ende un mejor resultado en el entrenamiento del modelo.

Con respecto a los indicadores de esta investigación se obtuvo resultados favorables.

En nuestra investigación se pudo observar una mejor precisión al utilizar Máquina de vectores (SVM) con un resultado de 100% para predecir el rendimiento académico, sin embargo, Burman (2019) utilizando el mismo algoritmo (SVM) obtuvo una precisión de 90%. Así mismo, para otros autores como Candia (2019) en su estudio comparativo alcanzó una precisión de 69.4% utilizando random forest, Hamoud (2018) observó que el algoritmo j48 alcanza una precisión de 62%, así mismo Segura (2018) obtuvo 67.41% utilizando arboles con degradado, por otro lado, el autor Vega (2019) consiguió 89% al aplicar el algoritmo XGBoosting, Alsalman (2019) alcanzó 90% utilizando redes neuronales.

También para el indicador sensibilidad para medir el modelo Machine Learning se obtuvo un porcentaje de 100% utilizando el algoritmo Máquina de vectores, otros autores como Burman (2019) también consideró medir la sensibilidad de su modelo obteniendo un resultado de 90%, así mismo Vega (2019) alcanzó hasta un 95% y Hamoud (2018) un valor de 63% .Por otro lado, autores como Chahuan (2019), Canagareddy (2019), Alsalman (2019), Candia (2019), Segura (2018), Chiheb (2017) y Soto (2015) solo consideraron como métrica, la precisión.

Finalmente, para el indicador especificidad en nuestra investigación se alcanzó un porcentaje de 100% utilizando Máquina de vectores, mientras que para árboles de decisión y K-NN se obtuvo 91.92%, 84% respectivamente. Autores como Burman (2019) y Vega (2019) también utilizaron el mismo indicador de medición obteniendo resultados igual a 91% utilizando SVM y 83% con el algoritmo XGBoosting respectivamente. De lo contrario autores como Chahuan (2019), Canagareddy (2019), Alsalman (2019), Candia (2019), Segura (2018), Chiheb (2017) y Soto (2015) solo consideraron como métrica, la precisión. Sin embargo, Hamoud (2018) se diferenció por la utilización de la métrica recall para medir su modelo, el cual obtuvo un valor de 62%.

## **VI. CONCLUSIONES**



En la presente investigación realizada se ha llegado a las siguientes conclusiones:

Se pudo concluir luego de usar varios algoritmos como: árbol de decisión, SVM y K-NN que el modelo de Machine Learning que brinda la mejor precisión en el rendimiento académico de los estudiantes universitarios es la máquina de vectores con un 100%.

Así mismo, se pudo determinar luego de usar varios algoritmos como: árbol de decisión, SVM y K-NN que el modelo Machine Learning que brinda la mejor sensibilidad, es decir casos verdaderos positivos para predecir el rendimiento académico de los estudiantes universitarios es la máquina de vectores con un 100%.

También se pudo determinar que luego de usar varios algoritmos como: árbol de decisión, SVM y K-NN que el modelo Machine Learning que brinda la mejor especificidad, es decir casos verdaderos negativos para predecir el rendimiento académico de los estudiantes universitarios es la máquina de vectores con un 100%.

Por lo tanto, se puede concluir que la máquina de aprendizaje (SVM) permite predecir con alta precisión, sensibilidad y especificidad el rendimiento académico de los alumnos universitarios. Además se pudo validar el nivel de similitud entre las observaciones utilizando el índice de Kappa de Cohen, con el cual se obtuvo 1 lo que significa muy buena concordancia.

## **VII.RECOMENDACIONES**

Las recomendaciones para futuras investigaciones son las siguientes:

- Se recomienda aplicar otros algoritmos de aprendizaje automático a los datos históricos de los estudiantes con el fin de ampliar el panorama de predicción, ya que actualmente el modelo ha sido entrenado con escasa data obtenida a través de un cuestionario.
- Se recomienda realizar estudios en entidades particulares y nacionales con relación al rendimiento académico para comprender mejor la situación del estudiante durante su estadía en la Universidad.
- Se recomienda recolectar datos de los estudiantes de distintas universidades y distintas carreras con el fin de crear un modelo aplicable para todas las Universidades.
- Se recomienda crear un modelo predictivo empleando más factores relacionados al rendimiento académico, como por ejemplo factores psicológicos, salud, gestión de tiempo, etc.
- Se recomienda hacer estudios sobre la utilización de algoritmos combinados con la finalidad de mejorar el creado en base a solo un algoritmo
- Se recomienda crear un software de apoyo institucional para evaluar a los estudiantes nuevos.

## **REFERENCIAS**

- AGGARWAL, D., MITTAL, S. y BALI, V., 2019. Prediction model for classifying students based on performance using machine learning techniques. *International Journal of Recent Technology and Engineering*, vol. 8, no. 2 Special Issue 7, pp. 496-503. ISSN 22773878. DOI 10.35940/ijrte.B1093.0782S719.
- AIGNEREN, M., 2009. El cuestionario el instrumento de recolección de información de la técnica de la encuesta social. *Centro de Estudios de OPINIÓN* [En línea], pp. 1-79. Disponible en: <http://aprendeonline.udea.edu.co/revistas/index.php/ceo/article/viewFile/1696/1345>.
- ALSALMAN, Y.S., KHAMEES ABU HALEMAH, N., ALNAGI, E.S. y SALAMEH, W., 2019. Using Decision Tree and Artificial Neural Network to Predict Students Academic Performance. *2019 10th International Conference on Information and Communication Systems, ICICS 2019*, pp. 104-109. DOI 10.1109/IACS.2019.8809106.
- BARRIENTOS, R., CRUZ, N., ACOSTA, H., RABATTE, I., GOGESCOECHEA, M., PAVÓN, P. y BLÁZQUEZ, S., 2009. Árboles De Decisión Como Herramienta En El Diagnóstico Médico. *Artículo Original* [en línea], pp. 20-24. Disponible en: [https://www.uv.mx/rm/num\\_anteriores/revmedica\\_vol9\\_num2/articulos/arboles.pdf](https://www.uv.mx/rm/num_anteriores/revmedica_vol9_num2/articulos/arboles.pdf)
- BBVA., 2019. Machine Learning, que es y cómo funciona. [En línea]. Disponible en: <https://www.bbva.com/es/machine-learning-que-es-y-como-funciona/>
- BERNAL, C. *Metodología de la investigación*. Tercera edición. Colombia: Pearson Education, 2010. 320p. ISBN 978-958-699-128-5
- BODDY, Clive. Sample size for qualitative research [En línea]. *Qualitative Market Research*. Reino Unido: Emerald Group Publishing Limited, Vol. 19 No. 4, pp. 426-432, 2016. Disponible en: [doi.org/10.1108/QMR-06-2016-0053](https://doi.org/10.1108/QMR-06-2016-0053)
- BURMAN, I. y SOM, S., 2019. Predicting Students Academic Performance Using Support Vector Machine. *Proceedings - 2019 Amity International Conference on*

Artificial Intelligence, AICAI 2019, pp. 756-759. DOI 10.1109/AICAI.2019.8701260.

CANAGAREDDY, D., SUBARAYADU, K. y HURBUNGS, V., 2019. A Machine Learning Model to Predict the Performance of University Students [En línea]. S.l.: Springer International Publishing. ISBN 9783030182397. Disponible en: [http://dx.doi.org/10.1007/978-3-030-18240-3\\_29](http://dx.doi.org/10.1007/978-3-030-18240-3_29).

CANDIA OVIEDO, D.I., 2019. Predicción del rendimiento académico de los estudiantes de la UNSAAC a partir de sus datos de ingreso utilizando algoritmos de aprendizaje automático. Universidad Nacional de San Antonio Abad del Cusco [en línea], pp. 141. Disponible en: <http://repositorio.unsaac.edu.pe/handle/UNSAAC/4120>.

CARMONA, E., 2016. Abstract Support Vector Machine 1 Introducción. [en línea], no. November, pp. 1-27. Disponible en: [https://www.researchgate.net/publication/263817587\\_Tutorial\\_sobre\\_Maquinas\\_de\\_Vectores\\_Soporte\\_SVM](https://www.researchgate.net/publication/263817587_Tutorial_sobre_Maquinas_de_Vectores_Soporte_SVM).

CHARRIS, L., HENRIQUEZ, C., HERNANDEZ, S., JIMENO, L., GUILLEN, O. y MORENO, S., 2018. Análisis comparativo de algoritmos de árboles de decisión en el procesamiento de datos biológicos. Investigación y Desarrollo en TIC [en línea], no. 1, pp. 10. Disponible en: <http://revistas.unisimon.edu.co/index.php/identific/article/view/3158/3905>.

CHAUHAN, N., SHAH, K., KARN, D. y DALAL, J., 2019. Prediction of Student's Performance Using Machine Learning. SSRN Electronic Journal, ISSN 1556-5068. DOI 10.2139/ssrn.3370802.

CHIHEB, F., BOUMAHDI, F., BOUARFA, H. y BOUKRAA, D., 2017. Predicting students' performance using decision trees: Case of an Algerian University. Proceedings of the 2017 International Conference on Mathematics and Information Technology, ICMIT 2017, vol. 2018-January, pp. 113-121. DOI 10.1109/MATHIT.2017.8259704.

- CHILCA, Alva (2017). Autoestima, hábitos de estudio y rendimiento académico en estudiantes universitarios. [En línea] [Fecha de consulta: 20 abril 2021]. Disponible en: <https://revistas.usil.edu.pe/index.php/pyr/article/view/145/377>
- FLORES-ORTIZ, E., RIVERA-CORONEL, H. y SÁNCHEZ-CANCINO, F., 2016. Bajo rendimiento académico: Más allá de los factores socio psicopedagógicos. Revista Digital Internacional de Psicología y Ciencia Social, vol. 2, no. 1, pp. 95-104. DOI 10.22402/j.rdipycs.unam.2.1.2016.60.95-104.
- FUNDACION BBVA., 2019. U-Ranking 2019 [En línea] [Fecha de consulta: 20 abril 2021]. Disponible en: <https://www.fbbva.es/noticias/un-33-de-los-alumnos-no-finaliza-el-grado-que-inicio-y-un-21-abandona-sin-terminar-estudios-universitarios/>
- GARBANZO VARGAS, G.M., 2012. Factores asociados al rendimiento académico en estudiantes universitarios, una reflexión desde la calidad de la educación superior pública. Revista Educación, vol. 31, no. 1, pp. 43. ISSN 0379-7082. DOI 10.15517/revedu.v31i1.1252.
- GARCÍA, C. y GÓMEZ, I., 2006. Algoritmos de aprendizaje: knn & kmeans. Universidad Carlos III de Madrid [en línea], Disponible en: <http://www.it.uc3m.es/jvillena/irc/practicas/08-09/06.pdf>.
- GARCÍA JIMÉNEZ, M.V., IZQUIERDO, J.M.A. y JIMÉNEZ BLANCO, A., 2000. La predicción del rendimiento académico: Regresión lineal versus regresión logística. Psicothema, vol. 12, no. SUPPL. 2, pp. 248-252. ISSN 02149915.
- GARCÍA PICHARDO, V.H., 2005. Algoritmo ID3 en la detección de ataques en aplicaciones web. [en línea], pp. 77. Disponible en: <https://repositorio.tec.mx/handle/11285/567132>.
- GARCÍA, T., 2003. El cuestionario como instrumento de investigación/evaluación. Página del proyecto de apoyo para profesionales de la formación (PROMETEO) de la Junta de Andalucía [En línea], pp. 28. Disponible en: [http://www.univsantana.com/sociologia/El\\_Cuestionario.pdf](http://www.univsantana.com/sociologia/El_Cuestionario.pdf).
- HAMOUD, A.K., HASHIM, A.S. y AWADH, W.A., 2018. Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis.

International Journal of Interactive Multimedia and Artificial Intelligence, vol. 5, no. 2, pp. 26. ISSN 1989-1660. DOI 10.9781/ijimai.2018.02.004.

HERNANDEZ S, Roberto. Metodología de la Investigación. 4ta. ed. México: McGraw-Hill, 2006. 736p. ISBN 970-10-5753-8

HINESTROZA, D., 2018. El Machine Learning a través de los tiempos y los aportes a la humanidad. [En línea]. Disponible en: <https://repository.unilibre.edu.co/bitstream/handle/10901/17289/EL%20MACHINE%20LEARNING.pdf?sequence=1&isAllowed=y>

IBM. [En línea] [Fecha de consulta: 20 abril 2021]. Disponible en: [https://www.ibm.com/ar-es/analytics/machine-learning?p1=Search&p4=43700052827921639&p5=b&gclid=Cj0KCQjw4ImEBhDFARIsAGOTMj-i0qqIQvMAIluCcHKQf1080isY7H8dWgVdG-1NVojw5OZNVQULJIsaAoMgEALw\\_wcB&gclsrc=aw.ds](https://www.ibm.com/ar-es/analytics/machine-learning?p1=Search&p4=43700052827921639&p5=b&gclid=Cj0KCQjw4ImEBhDFARIsAGOTMj-i0qqIQvMAIluCcHKQf1080isY7H8dWgVdG-1NVojw5OZNVQULJIsaAoMgEALw_wcB&gclsrc=aw.ds)

JIMENEZ, H., 2000. Competencia social: intervención preventiva en la escuela. Infancia y Sociedad. [En línea]. Disponible en: [https://www.researchgate.net/publication/237036207\\_El\\_rendimiento\\_academico\\_concepto\\_investigacion\\_y\\_desarrollo](https://www.researchgate.net/publication/237036207_El_rendimiento_academico_concepto_investigacion_y_desarrollo)

JOAQUIN, A (2017). Máquinas de Vector Soporte (Support Vector Machines, SVMs). [En línea] [Fecha de consulta: 20 abril 2021]. Disponible en: [https://www.cienciadedatos.net/documentos/34\\_maquinas\\_de\\_vector\\_soporte\\_support\\_vector\\_machines#M%C3%A1quinas\\_de\\_Vector\\_Soporte](https://www.cienciadedatos.net/documentos/34_maquinas_de_vector_soporte_support_vector_machines#M%C3%A1quinas_de_Vector_Soporte)

LEON, H., 2018. Desarrollo De Un Modelo Algorítmico Basado En Árboles De Decisión Para La Predicción De La Permanencia De Un Paciente En Un Proceso Psicoterapéutico. [En línea], pp. 95. Disponible en: <https://core.ac.uk/download/pdf/154890058.pdf>.

LOPEZ GUARIN, C.E., GUZMAN, E.L. y GONZALEZ, F.A., 2015. A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining. Revista Iberoamericana de Tecnologías del Aprendizaje, vol. 10, no. 3, pp. 119-125. ISSN 19328540. DOI 10.1109/RITA.2015.2452632.



- MANTEROLA, C., GRANDE, L., OTZEN, T., GARCÍA, N., SALAZAR, P. y QUIROZ, G., 2018. Confiabilidad, precisión o reproducibilidad de las mediciones. Métodos de valoración, utilidad y aplicaciones en la práctica clínica. Revista chilena de infectología, vol. 35, no. 6, pp. 680-688. ISSN 0716-1018. DOI 10.4067/s0716-10182018000600680.
- MATICH, D.J., 2001. Redes Neuronales: Conceptos Básicos y Aplicaciones. Historia [En línea], pp. 55. Disponible en: <ftp://decsai.ugr.es/pub/usuarios/castro/Material-Redes-Neuronales/Libros/match-redesneuronales.pdf>.
- MITCHELL, Tom (1997). Aprendizaje automático, McGraw Hill. [En línea] [Fecha de consulta: 20 abril 2021]. Disponible en: <http://www.cs.cmu.edu/afs/cs.cmu.edu/user/mitchell/ftp/mlbook.html>
- MOINE, J.M., 2013. Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo. XVII Congreso Argentino De Ciencias De La Computación, vol. XVII CACIC, pp. 111.
- MORERA, A., 2018. Introducción a los modelos de redes neuronales artificiales El perceptron simple y multicapa. [En línea]. Disponible en: <https://zagan.unizar.es/record/69205/files/TAZ-TFG-2018-148.pdf>
- MUEEN, A., ZAFAR, B. y MANZOOR, U., 2016. Modeling and Predicting Students' Academic Performance Using Data Mining Techniques. International Journal of Modern Education and Computer Science, vol. 8, no. 11, pp. 36-42. ISSN 20750161. DOI 10.5815/ijmecs.2016.11.05.
- NAVARRO, R.E., 2014. El rendimiento académico: concepto, investigación y desarrollo Red Iberoamericana de Investigación Sobre Cambio y Eficacia Escolar., no. January 2003.
- ORIHUELA MAITA, G.Y., 2019. Aplicación de Data Science para la Predicción del Rendimiento Académico de los Estudiantes de la Facultad de Ingeniería de Sistemas de la Universidad Nacional del Centro del Perú. Universidad Nacional del centro de Perú, pp. 114.

- PUARUNGROJ, W., BOONSIRISUMPUN, N., PONGPATRAKANT, P. y PHROMKHOT, S., 2018. Application of data mining techniques for predicting student success in English exit exam. ACM International Conference Proceeding Series, pp. 1-6. DOI 10.1145/3164541.3164638.
- QUEZADA LUCIO, N., 2018. K-Vecino más próximo en una aplicación de clasificación y predicción en el Poder Judicial del Perú. *Pesquimat*, vol. 21, no. 1, pp. 11. ISSN 1560-912X. DOI 10.15381/pes.v21i1.15077.
- RAMAPHOSA, K.I.M., ZUVA, T. y KWUIMI, R., 2018. Educational Data Mining to Improve Learner Performance in Gauteng Primary Schools. 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems, icABCD 2018, pp. 1-6. DOI 10.1109/ICABCD.2018.8465478.
- REYES SALDAÑA, J. y GARCÍA FLORES, R., 2005. El proceso de descubrimiento de conocimiento en bases de datos. *Ingenierías*, vol. 8, no. 26, pp. 37-47. ISSN 1405-0676.
- RIVAS-ASANZA, W., MAZON-OLIVO, B. y MEJÍA-PEÑAFIEL, E., 2018. Capítulo 1: Generalidades de las redes neuronales artificiales. *Redes neuronales artificiales aplicadas al reconocimiento de patrones [En línea]*, no. June, pp. 11-35. Disponible en: <http://repositorio.utmachala.edu.ec/handle/48000/12499>.
- ROKASH, L y MAIMON, O, (2015). *Data Mining With Decision Trees Theory and Applications*. World Scientific. Segunda Edición. 2015. [En línea] [Fecha de consulta: 20 Abril 2021]. Disponible en: [https://doc.lagout.org/Others/Data%20Mining/Data%20Mining%20with%20Decision%20Trees\\_%20Theory%20and%20Applications%20%282nd%20ed.%29%20%5BRokach%20%26%20Maimon%202014-10-23%5D.pdf](https://doc.lagout.org/Others/Data%20Mining/Data%20Mining%20with%20Decision%20Trees_%20Theory%20and%20Applications%20%282nd%20ed.%29%20%5BRokach%20%26%20Maimon%202014-10-23%5D.pdf)
- SALAS-DOMINGUEZ, M.I. y MUÑOZ-DÍAZ, I., 2019. Análisis de concordancia de atributos en color de piezas galvanizadas. *Revista de Tecnologías en procesos Industriales*, vol. 3, no. 6, pp. 1-6. DOI 10.35429/jtip.2019.6.3.1.6.
- SAMUEL, A.L., 2000. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, vol. 44, no. 1-2, pp. 207-219. ISSN 00188646. DOI 10.1147/rd.441.0206.

- SCHUBERT, Olga, [et al.]. Quantitative proteomics: challenges and opportunities in basic and applied research [En línea]. Nat Protocols: 12, 2017, pp. 1289–1294. Disponible en: <https://doi.org/10.1038/nprot.2017.040>
- SEGURA-MORALES, M. y LOZA-AGUIRRE, E., 2018. Using Decision Trees for Predicting Academic Performance Based on Socio-Economic Factors. Proceedings - 2017 International Conference on Computational Science and Computational Intelligence, CSCI 2017, pp. 1132-1136. DOI 10.1109/CSCI.2017.197.
- SHAKIL AHAMED, A.T.M., MAHMOOD, N.T. y RAHMAN, R.M., 2017. Prediction of Academic Performance During Adolescence Based on Socioeconomic, Psychological and Academic Factors. Studies in Computational Intelligence, vol. 710, pp. 71-80. ISSN 1860949X. DOI 10.1007/978-3-319-56660-3\_7.
- SIES., 2016. Avance curricular en educación superior [En línea] [Fecha de consulta: 20 abril 2021]. Disponible en: <https://bibliotecadigital.mineduc.cl/bitstream/handle/20.500.12365/1958/mono-710.pdf?sequence=1&isAllowed=y>
- SOTO, G., 2015. Uso De Técnicas De Machine Learning Para Predecir El Rendimiento Académico De Los Estudiantes De La Carrera De Ingeniería Civil En Informática De La Universidad Del Bio Bio, Chillan.
- USKOV, V, 2019. Machine Learning – based Predictive Analytics of Student Academic Performance in STEM Education. IEEE Global Engineering Education Conference (EDUCON), 2019. pp. 1370-1376.
- VARGAS CORDERO, Z.R., 2009. La Investigación aplicada: Una forma de conocer las realidades con evidencia científica. Revista Educación, vol. 33, no. 1, pp. 155. ISSN 0379-7082. DOI 10.15517/revedu.v33i1.538.
- ZAPATA-TAPASCO, A., PÉREZ-LONDOÑO, S. y MORA-FLÓREZ, J., 2014. Método basado en clasificadores k-NN parametrizados con algoritmos genéticos y la estimación de la reactancia para localización de fallas en sistemas de distribución. Revista Facultad de Ingeniería, no. 70, pp. 220-232. ISSN 01206230.

## **ANEXOS**

## ANEXO N° 2: Matriz de Consistencia

Problema	Objetivo	Hipótesis	Variable	Dimensiones	Indicadores	Metodología
P.G: ¿En qué medida Machine Learning permitirá predecir el rendimiento académico de los estudiantes universitarios?	O.G: Aplicar Machine Learning para predecir el rendimiento académico de los estudiantes universitarios	H.G El Machine Learning permite predecir el rendimiento académico de los estudiantes universitarios	Variable independiente:  Machine Learning			
P.E.1: ¿En qué medida Machine Learning permite predecir con precisión el rendimiento académico de los estudiantes universitarios?	O.E.1: Determinar en qué porcentaje Machine Learning permite predecir con precisión el rendimiento académico de los estudiantes universitarios	H.E.1: El Machine Learning permite predecir con precisión el rendimiento académico de los estudiantes universitarios	Variable dependiente:  predecir el rendimiento académico	Métricas de precisión	Accuracy = $(TP+TN/TP+TN+FP+FN) * 100$	Tipo de investigación: Aplicada  Diseño de investigación: Experimental de tipo pre-experimental
P.E.2: ¿En qué medida Machine Learning permite predecir con sensibilidad el rendimiento académico de los estudiantes universitarios?	O.E.2: Determinar en qué porcentaje Machine Learning permite predecir con sensibilidad el rendimiento académico de los estudiantes universitarios	H.E.2: El Machine Learning permite predecir con sensibilidad el rendimiento académico de los estudiantes universitarios			Sensibilidad = $(TP/TP+FN) * 100$	
P.E.3: ¿En qué medida Machine Learning permite predecir con especificidad el rendimiento académico de los estudiantes universitarios?	O.E.3: Determinar en qué porcentaje Machine Learning permite predecir con especificidad el rendimiento académico de los estudiantes universitarios	H.E.3: El Machine Learning permite predecir con especificidad el rendimiento académico de los estudiantes universitarios			Especificidad = $(TN/TN+FP) * 100$	

### ANEXO N° 3: Operacionalización de Variables

Variable	Definición conceptual	Definición Operacional	Dimensiones	Indicadores	Escala de medición
<b>Independiente:</b>  Machine Learning	“Es una rama de la inteligencia artificial que permite que las máquinas aprendan sin ser expresamente programadas para ello”. BBVA (2019)				
<b>Dependiente:</b>  Predicción rendimiento académico	<p>“Predicción de desempeño estudiantil, es una de las áreas clave de la aplicación de EDM (Minería de datos educativo), que se encarga de predecir el desempeño del estudiante en educación basadas en factores subyacentes que se dan como entrada”. KHAN, I (2019)</p> <p>“La predicción tiene como objetivo visualizar el valor de una variable llamada variable predicha de algún conjunto conocido de valores llamados predictores de datos”. Burman (2019)</p>	Para medir la predicción del rendimiento académico se dan usos de las métricas de precisión, la cual será obtenida a través de una herramienta de Machine Learning.	Métricas de precisión	<p>Accuracy = <math>(TP+TN/TP+TN+FP+FN) * 100</math></p> <p>Sensibilidad = <math>(TP/TP+FN) * 100</math></p> <p>Especificidad = <math>(TN/TN+FP) * 100</math></p>	Razón

ANEXO N° 4: Comparación de algoritmos de aprendizaje automático

ML Algorithm	Main focus on task type	Parametric	Training speed	Prediction speed	Automatically learns features
Linear Regression	Regression	Yes	Fast	Fast	No
Logistic Regression	Classification	Yes	Fast	Fast	No
Decision Tree Classification	Either	No	Fast	Depends on value of "n"	Yes
KNN	Either	No	Fast	Fast	No
ANN	Either	No	Slow	Fast	Yes
Naïve Bayes Classification	Classification	Yes	Fast (except feature extraction)	Fast	No
Random Forest Classification	Either	No	Slow	Mode-rate	Yes
SVM Classification	Either	No	Slow	Fast	Yes

Fuente: Uskov (2019)

**ANEXO N° 5: INSTRUMENTO DE OBSERVACIÓN PARA EL ALGORITMO SVM  
- FICHA DE REGISTRO**

Tipo de Prueba	Post Test
Investigador	Jeancarlos Garcia Dionisio
Fecha de inicio	

Algoritmo	SVM
-----------	-----

Matriz de confusión:

		Predicción	
		Positive	Negative
Observación	Positive	(TP) True Positive	(FN) False Negative
	Negative	(FP) False Positive	(TN) True Negative

Métricas a Evaluar:

Ítem	Indicador	Medida	Fórmula	Precisión
1	Precisión (exactitud)	Razón	$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} * 100$	100%
2	Especificidad	Razón	$Especificidad = \frac{TN}{TN+FP} * 100$	100%
3	Sensibilidad	Razón	$Sensibilidad = \frac{TP}{TP+FN} * 100$	100%

Otras Métricas:

Ítem	Indicador	Medida	Fórmula	Precisión
1	F 1 SCORE	Razón	$2 * (\text{Recall} * \text{Precisión}) / (\text{Recall} + \text{Precisión})$	1



ANEXO N° 6: INSTRUMENTO DE OBSERVACIÓN PARA EL ALGORITMO  
ÁRBOL DE DECISIÓN - FICHA DE REGISTRO

Tipo de Prueba	Post Test
Investigador	Jeancarlos Garcia Dionisio
Fecha de inicio	

Algoritmo	Árbol de decisión
-----------	-------------------

Matriz de confusión:

		Predicción	
		Positive	Negative
Observación	Positive	(TP) True Positive	(FN) False Negative
	Negative	(FP) False Positive	(TN) True Negative

Métricas a Evaluar:

Ítem	Indicador	Medida	Fórmula	Precisión
1	Precisión (exactitud)	Razón	$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} * 100$	89.51 %
2	Especificidad	Razón	$Especificidad = \frac{TN}{TN+FP} * 100$	91.92 %
3	Sensibilidad	Razón	$Sensibilidad = \frac{TP}{TP+FN} * 100$	85.05 %

ANEXO N° 7: INSTRUMENTO DE OBSERVACIÓN PARA EL ALGORITMO KNN  
- FICHA DE REGISTRO

Tipo de Prueba	Post Test
Investigador	Jeancarlos Garcia Dionisio
Fecha de inicio	

Algoritmo	K-NN
-----------	------

Matriz de confusión:

		Predicción	
		Positive	Negative
Observación	Positive	(TP) True Positive	(FN) False Negative
	Negative	(FP) False Positive	(TN) True Negative

Métricas a Evaluar:

Ítem	Indicador	Medida	Fórmula	Precisión
1	Precisión (exactitud)	Razón	$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} * 100$	79.75 %
2	Especificidad	Razón	$Especificidad = \frac{TN}{TN+FP} * 100$	84 %
3	Sensibilidad	Razón	$Sensibilidad = \frac{TP}{TP+FN} * 100$	72.41 %

ANEXO N° 8: VALIDACIÓN DE INSTRUMENTO - EXPERTO 1

**Apellidos y Nombre del experto:** Macedo Alcantara Dayan Fernando

**Título y/o grado:** Ingeniero de Sistemas e Informática

**Fecha:** 10/05/2021

**Nombre del instrumento:** Ficha de Registro

**Autor:** Jeancarlos Garcia Dionisio

**Título de investigación:** Machine Learning para predecir el rendimiento académico de los estudiantes universitarios.

INDICADORES	CRITERIOS	Deficiente 0-20%	Regular 21-50%	Bueno 51- 70%	Muy Bueno 71-80%	Excelente 81-100%
Claridad	Promedio de Validación					95
Objetividad	Esta expresado en conducta observable.					95
Actualidad	Es adecuado al avance de la ciencia.					95
Organización	Existe una organización lógica					95
Suficiencia	Comprende los aspectos de cantidad y calidad.					95
Intencionalidad	Adecuado para valorar aspectos del sistema metodológico y científico					95
Consistencia	Está basado en aspectos teóricos, científicos acordes a la tecnología educativa.					95
Coherencia	Está basado en aspectos teóricos, científicos acordes a la tecnología educativa.					95
Metodología	Responde al propósito del trabajo bajo los objetivos a lograr					95
Pertinencia	El instrumento es adecuado al tipo de investigación.					95
<b>Promedio de Validación</b>						95

Promedio de valoración: 95

Observaciones:



FIRMA

## ANEXO N° 9: VALIDACIÓN DE INSTRUMENTO - EXPERTO 2

**Apellidos y Nombre del experto:** Borja Reyna Whiston Kendrick

**Título y/o grado:** Maestro en Ingeniería de Sistemas e Informática

**Fecha:** 11/05/2021

**Nombre del instrumento:** Ficha de Registro

**Autor:** Jeancarlos Garcia Dionisio

**Título de investigación:** Machine Learning para predecir el rendimiento académico de los estudiantes universitarios.

INDICADORES	CRITERIOS	Deficiente 0-20%	Regular 21-50%	Bueno 51- 70%	Muy Bueno 71-80%	Excelente 81-100%
Claridad	Promedio de Validación					92
Objetividad	Esta expresado en conducta observable.					92
Actualidad	Es adecuado al avance de la ciencia.					92
Organización	Existe una organización lógica					92
Suficiencia	Comprende los aspectos de cantidad y calidad.					92
Intencionalidad	Adecuado para valorar aspectos del sistema metodológico y científico					92
Consistencia	Está basado en aspectos teóricos, científicos acordes a la tecnología educativa.					92
Coherencia	Está basado en aspectos teóricos, científicos acordes a la tecnología educativa.					92
Metodología	Responde al propósito del trabajo bajo los objetivos a lograr					92
Pertinencia	El instrumento es adecuado al tipo de investigación.					92
<b>Promedio de Validación</b>						92

Promedio de valoración: 92%

Observaciones: Ninguna



\_\_\_\_\_  
FIRMA

## ANEXO N° 10: VALIDACIÓN DE INSTRUMENTO - EXPERTO 3

**Apellidos y Nombre del experto:** Mendoza Rivera Ricardo Dario

**Título y/o grado:** Maestro en Ingeniería de Sistemas e Informática,  
**ESPECIALIDAD:** Sistemas de Información

**Fecha:** 11/05/2021

**Nombre del instrumento:** Ficha de Registro

**Autor:** Jeancarlos Garcia Dionisio

**Título de investigación:** Machine Learning para predecir el rendimiento académico de los estudiantes universitarios.

INDICADORES	CRITERIOS	Deficiente 0-20%	Regular 21-50%	Bueno 51- 70%	Muy Bueno 71-80%	Excelente 81-100%
Claridad	Promedio de Validación					95
Objetividad	Esta expresado en conducta observable.					95
Actualidad	Es adecuado al avance de la ciencia.					95
Organización	Existe una organización lógica					94
Suficiencia	Comprende los aspectos de cantidad y calidad.					95
Intencionalidad	Adecuado para valorar aspectos del sistema metodológico y científico					95
Consistencia	Está basado en aspectos teóricos, científicos acordes a la tecnología educativa.					95
Coherencia	Está basado en aspectos teóricos, científicos acordes a la tecnología educativa.					94
Metodología	Responde al propósito del trabajo bajo los objetivos a lograr					95
Pertinencia	El instrumento es adecuado al tipo de investigación.					95
<b>Promedio de Validación</b>						

Promedio de valoración: 94.8%

Observaciones:



\_\_\_\_\_  
 FIRMA

## ANEXO N° 11: VALIDACIÓN DE INSTRUMENTO - EXPERTO 4

**Apellidos y Nombre del experto:** Daza Vergaray Alfredo

**Título y/o grado:** DrSc en Ingeniería de Sistemas

**Fecha:** 11/05/2021

**Nombre del instrumento:** Ficha de Registro

**Autor:** Jeancarlos Garcia Dionisio

**Título de investigación:** Machine Learning para predecir el rendimiento académico de los estudiantes universitarios.

INDICADORES	CRITERIOS	Deficiente 0-20%	Regular 21-50%	Bueno 51- 70%	Muy Bueno 71-80%	Excelente 81-100%
Claridad	Promedio de Validación					99
Objetividad	Esta expresado en conducta observable.					99
Actualidad	Es adecuado al avance de la ciencia.					97
Organización	Existe una organización lógica					99
Suficiencia	Comprende los aspectos de cantidad y calidad.					95
Intencionalidad	Adecuado para valorar aspectos del sistema metodológico y científico					98
Consistencia	Está basado en aspectos teóricos, científicos acordes a la tecnología educativa.					95
Coherencia	Está basado en aspectos teóricos, científicos acordes a la tecnología educativa.					98
Metodología	Responde al propósito del trabajo bajo los objetivos a lograr					95
Pertinencia	El instrumento es adecuado al tipo de investigación.					97
<b>Promedio de Validación</b>						

Promedio de valoración: 97.2%

Observaciones:




---

FIRMA

## ANEXO N° 12: INSTRUMENTO PARA RECOLECCIÓN DE DATOS

Item	Factor	Aspecto	Valor
1	Determinante personal	¿Género biológico?	Masculino / Femenino
2		¿Cuál es tu edad?	Dato Numérico
3		¿A qué distrito perteneces?	
4		¿A qué edad ingresó a la carrera?	Dato numérico
5		¿Tienes beca completa o media beca en la universidad?	Beca completa / Media Beca / Ninguna
6		¿A qué carrera profesional perteneces?	
7	Auto concepto	¿Qué tan responsable se considera?	muy irresponsable/ irresponsable/ indiferente/ responsable/ muy responsable
8		¿Considera que tiene las aptitudes suficientes para culminar el ciclo de manera exitosa?	muy irresponsable/ irresponsable/ indiferente/ responsable/ muy responsable
9		¿Se considera una persona inteligente?	muy irresponsable/ irresponsable/ indiferente/ responsable/ muy responsable
10	Motivación	¿Es de su agrado la carrera profesional elegida?	SI / NO
11		¿Cuántas horas al día dedica a estudiar?	Dato Numérico entre 0 - 24
12		Si pudieras escoger entre estudiar o no estudiar. ¿Estudiarías?	SI / NO
13	Socio cultural	¿Cuántos integrantes hay en su grupo familiar?	Dato Numérico
14		¿Cuántos integrantes de su grupo familiar trabajan?	Dato Numérico
15		¿Cuántos integrantes de su grupo familiar estudian?	Dato Numérico
16		¿Cuántos integrantes de su grupo familiar son pensionados?	Dato Numérico
17		¿Eres dependiente o independiente de tu padre?	Dependiente / Independiente
18		¿Eres dependiente o independiente de tu madre?	Dependiente / Independiente
19	Educación de los padres	¿Qué nivel de educación posee su madre?	Primaria / Secundaria / Técnico / Superior
20		¿Qué nivel de educación posee su padre?	Primaria / Secundaria / Técnico / Superior
21	Inteligencia Emocional	¿Cómo reacciona frente a un problema en su entorno social?	Reacciona con violencia física o verbal / Intenta conversar sobre la situación / Con indiferencia
22		¿De qué forma influye el stress en su vida académica?	Muy negativamente / Negativamente / No influye / Positivamente / Muy positivamente
23		¿Qué nivel de empatía presenta hacia sus compañeros?	Sin empatía / Muy poco empático / Indiferente / Poco empático / Muy empático
24	Económico	¿Cuántas horas a la semana trabaja?	Dato numérico
25		¿Cuál es el ingreso mensual aproximado en su grupo familiar?	Dato numérico
26	Escuelas de origen.	¿En qué tipo de establecimiento completó sus estudios básicos?	Público / Privado
27		¿En qué tipo de establecimiento completó sus estudios medios?	Público / Privado
28	Variable objetivo	¿Cuál es tu promedio ponderado?	Dato numérico

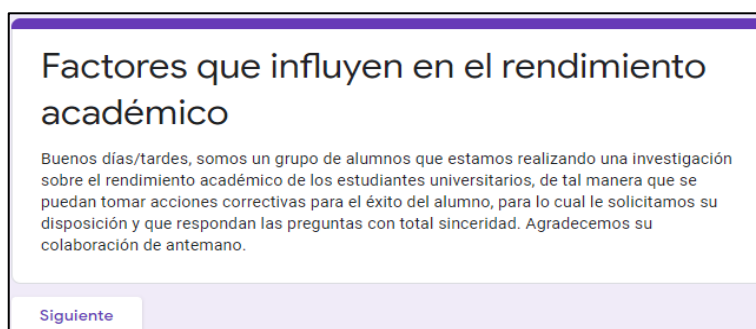
## Desarrollo del modelo Machine Learning para predecir el rendimiento académico de los estudiantes universitarios

El presente documento detallará las tareas llevadas a cabo para el desarrollo del modelo Machine Learning para predecir el rendimiento académico de los estudiantes universitarios, haciendo uso de método KDD, la cual está dividida en 5 etapas. A continuación, se dará más detalle de lo que se realizó en cada etapa.

### Etapa 1: Etapa de selección de datos

El fin de esta investigación es predecir el rendimiento académico de los estudiantes, de tal manera que se logre identificar aquellos estudiantes con probabilidades de bajo rendimiento académico, buscando que la universidad tome las mejores decisiones para el éxito del alumno.

Para la obtención de los datos de alumno, se realizó un cuestionario que consta de 28 preguntas los cuales estuvieron relacionadas con determinantes personales (6 preguntas), auto-concepto (3 preguntas), la motivación (3 preguntas), factores socio-culturales (6 preguntas), educación de los padres (2 preguntas), inteligencia emocional (3 preguntas), factores económicos (2 preguntas), escuela origen (2 preguntas) y variable objetiva (1 pregunta), en el ANEXO N° 13 se puede observar en detalle las preguntas. Estas preguntas fueron transcritas a google formulario y enviadas a estudiantes de Ingeniería de Sistemas e Informática de la Universidad, obteniendo un total de 87 registros.



**Factores que influyen en el rendimiento académico**

Buenos días/tardes, somos un grupo de alumnos que estamos realizando una investigación sobre el rendimiento académico de los estudiantes universitarios, de tal manera que se puedan tomar acciones correctivas para el éxito del alumno, para lo cual le solicitamos su disposición y que respondan las preguntas con total sinceridad. Agradecemos su colaboración de antemano.

[Siguiete](#)

Figura 11: Cuestionario utilizando Google Forms

Fuente: Elaboración propia



## Etapa 2: Etapa de pre-procesamiento / limpieza de datos

Luego de ejecutar el cuestionario, se pudo recolectar 87 registros estudiantiles, los cuales fueron exportados a la herramienta SPSS Statistic.

En primer lugar, se realizó la definición de las variables, las cuales están asociadas a las preguntas del cuestionario, se definió que los valores sean de tipo numérico, la anchura de 8, en la columna “Decimales” se utilizó para la variable Ingreso\_mensual\_familiar 1 decimal y para la variable Promedio\_ponderado 2 decimales. Luego en la columna “Valores”, se añadió las alternativas por cada pregunta y su equivalencia en números, para la columna “Perdidos” se utilizó el valor 99. También en la columna “Medida” se seleccionó entre ordinal, escalar y nominal según corresponde y finalmente en la columna “Rol” se seleccionó el tipo de dato, es decir si es de entrada, destino, etc. En nuestro caso se eligió para todas las variables, tipo Entrada.

	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	Genero	Numérico	8	0	Género biológico	{1, Masculin...	Ninguno	8	Izquierda	Ordinal	Entrada
2	Edad	Numérico	8	0	¿Cuál es tu ed...	{99, No cont...	99	8	Derecha	Escala	Entrada
3	Distrito	Numérico	8	0	¿A qué distrito ...	{1, 26 Octu...	99	8	Izquierda	Nominal	Entrada
4	Edad_ingreso_carrera	Numérico	8	0	¿A qué edad in...	Ninguno	99	8	Derecha	Escala	Entrada
5	Beneficio	Numérico	8	0	¿Tienes benefic...	{1, Beca)...	Ninguno	8	Izquierda	Nominal	Entrada
6	Carrera_profesional	Numérico	8	0	¿A qué carrera ...	{1, Ingenier...	Ninguno	16	Izquierda	Nominal	Entrada
7	Actitud_responsable	Numérico	8	0	¿Qué tan respo...	{1, muy irre...	Ninguno	8	Izquierda	Nominal	Entrada
8	Aptitud_culminar_ciclo_exitoso	Numérico	8	0	¿Considera que...	{1, muy en ...	Ninguno	8	Izquierda	Nominal	Entrada
9	Persona_inteligente	Numérico	8	0	¿Se considera ...	{1, muy en ...	Ninguno	8	Izquierda	Nominal	Entrada
10	Gusto_carrera_profesional_elegida	Numérico	8	0	¿Es de su agr...	{1, Si)...	Ninguno	8	Derecha	Nominal	Entrada
11	Horas_dia_estudio	Numérico	8	0	¿Cuántas hora...	Ninguno	Ninguno	8	Derecha	Escala	Entrada
12	Estudiar_noEstudiar	Numérico	8	0	Si pudieras esc...	{1, Si)...	Ninguno	8	Izquierda	Nominal	Entrada
13	Cantidad_integrantes_familia	Numérico	8	0	¿Cuántos integ...	{99, No cont...	99	8	Derecha	Escala	Entrada
14	Cantidad_integrantes_familia_trabajan	Numérico	8	0	¿Cuántos integ...	Ninguno	99	8	Derecha	Escala	Entrada
15	Cantidad_integrantes_familia_estudian	Numérico	8	0	¿Cuántos integ...	Ninguno	99	7	Derecha	Escala	Entrada
16	Cantidad_integrantes_familia_pensionados	Numérico	8	0	¿Cuántos integ...	Ninguno	99	8	Derecha	Escala	Entrada
17	Dependiente_Independiente_padre	Numérico	8	0	¿Eres dependie...	{1, independ...	Ninguno	8	Izquierda	Nominal	Entrada
18	Dependiente_Independiente_madre	Numérico	8	0	¿Eres dependie...	{1, independ...	Ninguno	8	Izquierda	Nominal	Entrada
19	Nivel_educacion_madre	Numérico	8	0	¿Qué nivel de e...	{1, primaria)...	Ninguno	8	Izquierda	Nominal	Entrada
20	Nivel_educacion_padre	Numérico	8	0	¿Qué nivel de e...	{1, primaria)...	Ninguno	8	Izquierda	Nominal	Entrada
21	Reaccion_problema_entorno_social	Numérico	8	0	¿Cómo reaccio...	{1, Reaccio...	Ninguno	8	Izquierda	Nominal	Entrada
22	Influencia_stress_vida_academica	Numérico	8	0	¿De qué forma ...	{1, muy neg...	Ninguno	8	Izquierda	Nominal	Entrada
23	Nivel_empatia_compañeros	Numérico	8	0	¿Qué nivel de e...	{1, sin emp...	Ninguno	8	Izquierda	Nominal	Entrada
24	Horas_semana_trabaja	Numérico	8	0	¿Cuántas hora...	Ninguno	Ninguno	8	Derecha	Escala	Entrada
25	Ingreso_mensual_familiar	Numérico	8	1	¿Cuál es el ingr...	{99999,0, N...	99999,0	8	Derecha	Escala	Entrada
26	Tipo_establecimiento_estudio_primario	Numérico	8	0	¿En que tipo d...	{1, público)...	Ninguno	8	Izquierda	Nominal	Entrada
27	Tipo_establecimiento_estudio_secundario	Numérico	8	0	¿En que tipo d...	{1, público)...	Ninguno	8	Izquierda	Nominal	Entrada
28	Promedio_ponderado	Numérico	8	2	¿Cuál es tu pro...	{99,00, No c...	99,00	8	Derecha	Escala	Entrada

Figura 12: Definición de variables – SPSS Statistics

Fuente: Elaboración propia

A continuación, se da detalle sobre el valor asignado por cada variable:

Genero	
Femenino	2
Masculino	1
Beneficio	
Beca	
	1
Media Beca	2
Ninguna	3
Carrera_profesional	
Ingeniería de Sistemas	1
Ingeniería Industrial	2
Contabilidad	3
Mecánica	4
Actitud_responsable	
Muy irresponsable	1
Irresponsable	2
Indiferente	3
Responsable	4
Muy responsable	5
Aptitud_culminar_ciclo_exitoso	
Muy en desacuerdo	1
En desacuerdo	2
No sabe	3
De acuerdo	4
Muy de acuerdo	5
Persona_inteligente	
Muy en desacuerdo	1
En desacuerdo	2
No sabe	3
De acuerdo	4
Muy de acuerdo	5
Gusto_carrera_profesional_elegida	
Si	1
No	2
Estudiar_noEstudiar	
Si	1
No	2
Dependiente_Independiente_padre	
Independiente	1
Dependiente	2
Dependiente_Independiente_madre	
Independiente	1
Dependiente	2
Nivel_educación_madre	
Primaria	1
Secundaria	2
Técnico	3
Superior	4
Nivel_educación_padre	
Primaria	1

Secundaria	2
Técnico	3
Superior	4
Reacción_problema_entorno_social	
Reacciona con violencia física o verbal	1
Intenta conversar sobre la situación	2
Con indiferencia	3
Influencia_stress_vida_academica	
Muy negativamente	1
Negativamente	2
No influye	3
Positivamente	4
Muy positivamente	5
Nivel_empatia_compañeros	
Sin empatía	1
Muy poco empático	2
Indiferente	3
Poco empático	4
Muy empático	5
Tipo_establecimiento_estudio_primario	
Público	1
Privado	2
Tipo_establecimiento_estudio_secundario	
Público	1
Privado	2

En segundo lugar, de los datos exportados de google formulario a Excel, equivalentes a un total de 87 registros estudiantiles. Se hizo un análisis de los datos, corroborando que no existiera datos nulos, también se homogenizaron las respuestas con relación a las siguientes preguntas:

- ¿Cuál es tu edad?, la respuesta debe ser de tipo numérico

Se corrigió el registro N°66

- ¿A qué distrito perteneces?

Se homogenizaron las respuestas para todos los registros

- ¿A qué edad ingresó a la carrera?, la respuesta debe ser de tipo numérico

Se corrigieron los registros N°8, N°61, N°63, N°66, N°77

- ¿A qué carrera profesional perteneces?

Se homogenizaron las respuestas para todos los registros

- ¿Cuántas horas al día dedica a estudiar?, la respuesta debe ser de tipo numérico

Se corrigieron los registros N°22, N°50, N°52, N°66, N°86

- ¿Cuántas horas a la semana trabaja?, la respuesta debe ser de tipo numérico

Se corrigieron los registros N°22, N°50, N°52

	A	B	C	D	E	F	G	H	I	J	K
	Género biológico	¿Cuál es tu edad?	¿A qué distrito pertenece?	¿A qué edad ingresó a la universidad?	¿Tienes beneficio en la universidad?	¿A qué carrera profesional estás estudiando?	¿Qué tan responsable se considera que eres?	¿Consideras que tienes las habilidades necesarias para el trabajo?	¿Se considera una persona responsable?		
59	Femenino	20	Chimbote	18	Ninguna	Ing. de Sistemas e Inform.	responsable	de acuerdo	de acuerdo	Si	
60	Masculino	19	Chimbote	17	Ninguna	ING. de Sistemas e inform.	responsable	de acuerdo	no sabe	Si	
61	Masculino	22	Nuevo Chimbote	2017	Ninguna	ing de Sistemas	indiferente	no sabe	de acuerdo	Si	
62	Masculino	22	Santa	18	Ninguna	Ingeniería de sistemas e	responsable	de acuerdo	de acuerdo	Si	
63	Masculino	21	Nuevo Chimbote	2018	Ninguna	Ingeniería de sistemas e i	responsable	de acuerdo	muy de acuerdo	Si	
64	Masculino	18	Chimbote	17	Ninguna	Ing Sistemas	responsable	de acuerdo	de acuerdo	Si	
65	Masculino	24	Chimbote	20	Ninguna	Ingeniería de Sistemas e i	responsable	de acuerdo	de acuerdo	Si	
66	Masculino	20 años	Nuevo Chimbote	19 años	Ninguna	Ingeniería de Sistemas e i	responsable	de acuerdo	de acuerdo	Si	
67	Masculino	20	Nuevo Chimbote	17	Ninguna	Ingeniería de Sistemas e i	responsable	de acuerdo	de acuerdo	Si	
68	Masculino	19	CHIMBOTE	17	Ninguna	Ingeniería de Sistemas e i	responsable	de acuerdo	no sabe	Si	
69	Masculino	22	Chimbote	19	Ninguna	Ingeniería de Sistemas e i	responsable	de acuerdo	de acuerdo	Si	
70	Masculino	18	Nuevo Chimbote	16	Ninguna	Ingeniería de Sistemas e i	indiferente	no sabe	de acuerdo	Si	
71	Masculino	20	casma	19	Ninguna	ingeniería en sistemas e i	responsable	de acuerdo	de acuerdo	Si	
72	Masculino	21	SAMANCO	18	Ninguna	INGENIERÍA DE SISTEMAS	indiferente	de acuerdo	no sabe	Si	
73	Masculino	31	Nuevo Chimbote	19	Ninguna	Ingeniería de Sistemas	responsable	de acuerdo	muy de acuerdo	Si	
74	Femenino	26	Nuevo Chimbote	18	Ninguna	Ingeniería de Sistemas e i	indiferente	de acuerdo	de acuerdo	Si	
75	Masculino	19	Nuevo Chimbote	16	Ninguna	Ingeniería de sistemas e i	responsable	de acuerdo	en desacuerdo	Si	
76	Masculino	23	CHIMBOTE	19	Ninguna	no sabe de Sistemas e i	responsable	no sabe	de acuerdo	Si	
77	Masculino	21	Casma	17 años	Ninguna	Ingeniería de Sistemas e i	indiferente	de acuerdo	de acuerdo	Si	
78	Masculino	19	Chimbote	17	Ninguna	ING Sistemas	indiferente	de acuerdo	de acuerdo	Si	
79	Masculino	21	Nuevo Chimbote	18	Ninguna	Ingeniería de Sistemas e i	responsable	de acuerdo	de acuerdo	Si	
80	Masculino	22	Nuevo Chimbote	19	Ninguna	Ingeniería de Sistemas e i	responsable	de acuerdo	de acuerdo	Si	
81	Masculino	21	Chimbote	17	Ninguna	ING. SISTEMAS E INFOR	responsable	de acuerdo	muy de acuerdo	Si	
82	Masculino	21	Chimbote	17	Ninguna	Ing. Sistemas e informatic	responsable	de acuerdo	muy de acuerdo	Si	
83	Masculino	18	Cosishco	17	Ninguna	Ingeniería de sistemas e i	responsable	de acuerdo	de acuerdo	Si	
84	Masculino	19	Nuevo Chimbote	18	Ninguna	Ingeniería de Sistemas e i	muy responsable	de acuerdo	de acuerdo	Si	
85	Masculino	20	Nuevo Chimbote	16	Ninguna	Ingeniería de Sistemas e i	responsable	de acuerdo	de acuerdo	Si	

Figura 13: Datos importados de Google Forms a Excel

Fuente: Elaboración propia

Luego de realizar la tarea mencionada anteriormente se importó los datos al SPSS Statistic

	Genero	Edad	Distrito	Edad_ingr_eso_carrera	Beneficio	Carrera_profesional	Actividad_responsable	Actividad_ultima	Persona_inteligente	Quito_carrera_profesional	Horas_dia_estudio	Estudiar_noEstudiar	Cantidad_integrantes_familia	Cantidad_integrantes_familia	Cantidad_integrantes_familia	Cantidad_integrantes_familia
1	Femenino	19	Nuevo Ch.	17	Ninguno	Ingeniería de Sistemas	muy respo.	muy de ac.	muy de ac.	Si	6.5	Si	3	1	1	1
2	Femenino	19	Nuevo Ch.	17	Ninguno	Ingeniería de Sistemas	responsable	de acuerdo	de acuerdo	Si	3.5	Si	2	1	1	1
3	Masculino	20	Nuevo Ch.	19	Ninguno	Ingeniería de Sistemas	responsable	muy de ac.	de acuerdo	Si	15.5	Si	5	2	3	3
4	Masculino	22	Casma	17	Ninguno	Ingeniería de Sistemas	indiferente	de acuerdo	de acuerdo	Si	10.5	Si	3	1	0	0
5	Masculino	18	Nuevo Ch.	17	Ninguno	Ingeniería de Sistemas	muy respo.	de acuerdo	de acuerdo	Si	3.5	Si	3	1	2	2
6	Femenino	18	Nuevo Ch.	17	Ninguno	Ingeniería de Sistemas	responsable	muy de ac.	de acuerdo	Si	3.5	Si	6	1	4	4
7	Masculino	20	Nuevo Ch.	18	Ninguno	Ingeniería de Sistemas	responsable	de acuerdo	de acuerdo	Si	2.5	Si	7	1	5	5
8	Masculino	19	Nuevo Ch.	16	Ninguno	Ingeniería de Sistemas	responsable	de acuerdo	no sabe	Si	4.5	Si	3	1	0	0
9	Femenino	18	Nuevo Ch.	17	Ninguno	Ingeniería de Sistemas	responsable	de acuerdo	de acuerdo	Si	3.5	Si	6	1	4	4
10	Masculino	20	Nuevo Ch.	19	Ninguno	Ingeniería de Sistemas	responsable	indiferente	de acuerdo	Si	4.5	Si	8	5	2	2
11	Femenino	20	Nuevo Ch.	18	Ninguno	Ingeniería de Sistemas	responsable	de acuerdo	no sabe	Si	3.5	Si	4	2	2	2
12	Masculino	19	Casma	18	Ninguno	Ingeniería de Sistemas	responsable	de acuerdo	de acuerdo	Si	4.5	Si	5	2	1	1
13	Masculino	19	Nuevo Ch.	17	Ninguno	Ingeniería de Sistemas	responsable	muy de ac.	muy de ac.	Si	2.5	Si	4	1	3	3
14	Masculino	19	Nuevo Ch.	17	Ninguno	Ingeniería de Sistemas	responsable	muy de ac.	muy de ac.	Si	3.5	Si	5	3	0	0
15	Masculino	20	Nuevo Ch.	18	Ninguno	Ingeniería de Sistemas	responsable	de acuerdo	de acuerdo	Si	4.5	Si	3	2	0	0
16	Masculino	21	Nuevo Ch.	17	Ninguno	Ingeniería de Sistemas	responsable	muy de ac.	de acuerdo	Si	16.5	Si	4	2	2	2
17	Masculino	25	Nuevo Ch.	22	Ninguno	Ingeniería de Sistemas	responsable	muy de ac.	muy de ac.	Si	2.5	Si	4	1	3	3
18	Masculino	20	Nuevo Ch.	17	Ninguno	Ingeniería de Sistemas	responsable	muy de ac.	de acuerdo	Si	4.5	Si	4	2	1	1
19	Masculino	21	Nuevo Ch.	17	Ninguno	Ingeniería de Sistemas	responsable	de acuerdo	de acuerdo	Si	2.5	Si	3	1	2	2
20	Masculino	26	Santa	20	Ninguno	Ingeniería de Sistemas	responsable	de acuerdo	de acuerdo	Si	2.5	Si	5	3	1	1
21	Masculino	20	Nuevo Ch.	16	Ninguno	Ingeniería de Sistemas	responsable	de acuerdo	de acuerdo	Si	6.5	Si	4	2	2	2
22	Masculino	21	Casma	18	Ninguno	Ingeniería de Sistemas	indiferente	muy de ac.	de acuerdo	Si	3.5	Si	4	2	2	2
23	Masculino	23	Santa	18	Ninguno	Ingeniería de Sistemas	responsable	muy de ac.	de acuerdo	Si	8.5	Si	10	3	3	3
24	Masculino	20	Nuevo Ch.	16	Ninguno	Ingeniería de Sistemas	indiferente	muy de ac.	de acuerdo	Si	6.5	Si	6	4	2	2
25	Masculino	20	Nuevo Ch.	16	Ninguno	Ingeniería de Sistemas	responsable	de acuerdo	de acuerdo	Si	6.5	Si	3	2	0	0
26	Femenino	19	Nuevo Ch.	16	Ninguno	Ingeniería de Sistemas	responsable	no sabe	no sabe	Si	10.5	Si	4	1	2	2
27	Masculino	23	Nuevo Ch.	21	Ninguno	Ingeniería de Sistemas	responsable	de acuerdo	de acuerdo	Si	4.5	Si	6	2	2	2

Figura 14: Datos importados de Excel a SPSS Statistics

Fuente: Elaboración propia

Así mismo, a través de la herramienta se realizó una cuantificación de ítems no contestados, para lo cual se realizó lo siguiente:

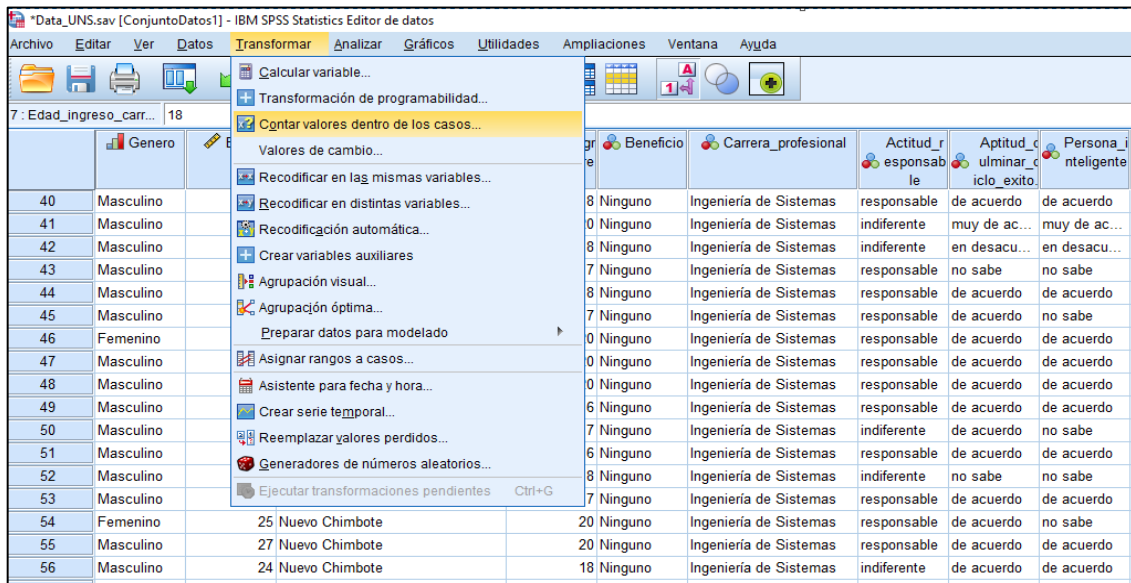


Figura 15: Cuantificación de ítems no contestados parte 1 - SPSS Statistics

Fuente: Elaboración propia

Se agrega una variable que obtendrá la cantidad de datos perdidos por cada registro

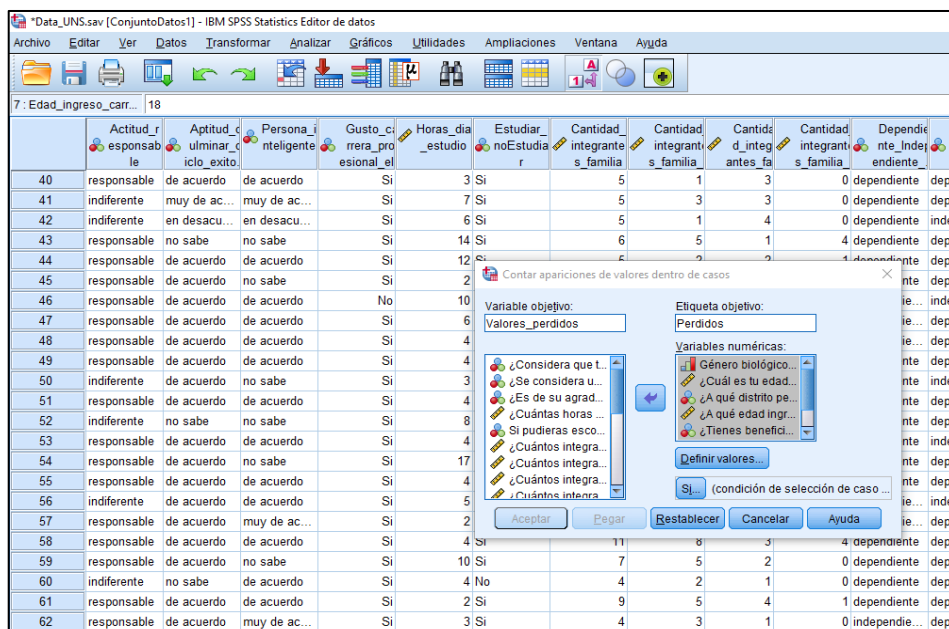


Figura 16: Cuantificación de ítems no contestados parte 2 - SPSS Statistics

Fuente: Elaboración propia

Finalmente se evidenció q no hubo ítems nulos en el total de registros (87)

	Influencia stress_y_acade	Nivel_empatia_compañeros	Horas_semana_trabaja	Ingreso_mensual_familiar	Tipo_establecimiento_estudi.	Tipo_establecimiento_estudi.	Promedio_ponderado	Valores_perdidos
40	vame...	muy empát...	0	900,0	público	público	13,00	0
41	uye	poco empá...	36	2500,0	público	privado	13,00	0
42	vame...	poco empá...	0	2000,0	privado	privado	13,00	0
43	vame...	poco empá...	0	10000,0	privado	privado	13,00	0
44	vame...	poco empá...	5	1500,0	público	público	13,00	0
45	vame...	poco empá...	0	15000,0	público	público	13,00	0
46	vame...	muy empát...	20	1800,0	público	público	12,80	0
47	vame...	poco empá...	24	1200,0	público	público	12,66	0
48	vame...	poco empá...	24	1100,0	público	público	12,16	0
49	vame...	poco empá...	20	600,0	público	privado	12,00	0
50	vame...	poco empá...	0	1500,0	privado	privado	12,00	0
51	vame...	muy empát...	20	600,0	público	privado	12,00	0
52	vame...	indiferente	0	500,0	privado	privado	12,00	0
53	uye	poco empá...	0	1800,0	privado	privado	12,00	0
54	vame...	poco empá...	48	1200,0	público	público	12,00	0
55	vame...	poco empá...	48	1500,0	público	público	11,00	0
56	vame...	poco empá...	27	1200,0	público	público	11,00	0
57	vame...	indiferente	0	1000,0	privado	público	16,00	0
58	vame...	muy empát...	0	10000,0	privado	privado	14,00	0
59	vame...	poco empá...	0	4000,0	privado	público	15,00	0
60	vame...	indiferente	0	600,0	público	público	13,00	0
61	uye	poco empá...	0	2500,0	público	público	13,00	0
62	uye	sin empatía	8	2500,0	público	público	13,60	0
63	uye	poco empá...	0	2500,0	público	público	15,60	0
64	uye	indiferente	35	1500,0	público	privado	12,00	0
65	vame...	poco empá...	0	3000,0	público	público	15,00	0
66	vame...	muy empát...	0	4000,0	privado	privado	14,00	0
67	vame...	poco empá...	0	1500,0	privado	privado	15,00	0

Figura 17: Cuantificación de ítems no contestados parte 3 - SPSS Statistics

Fuente: Elaboración propia

### Etapa 3: Etapa de transformación y reducción

En esta etapa, se utilizó la herramienta SPSS Modeler para crear el modelo predictivo Machine Learning, así mismo se deshabilitaron algunas variables no influyentes.

Inicialmente se cargó la base de datos creada a partir de la herramienta SPSS Statistic, para lo cual se utilizó un “nodo origen” de tipo “Archivo de Statistic”, para validar la importación de datos se utilizó un “nodo resultado” de tipo Tabla

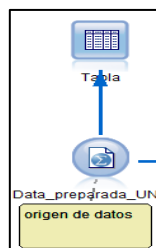


Figura 18: Carga de datos con nodo origen - SPSS Modeler

Fuente: Elaboración propia

Luego se utilizó un “nodo de operaciones con campos”, específicamente el nodo derivar, en el cual se insertó lógica para clasificar a los estudiantes de rendimiento reprobado con valor (1), regulares con valor (2), buenos (3), distinguidos (4) y excelentes (5), esto se realizó luego de identificar a la variable objetivo, para nuestro estudio es el promedio ponderado del estudiante.

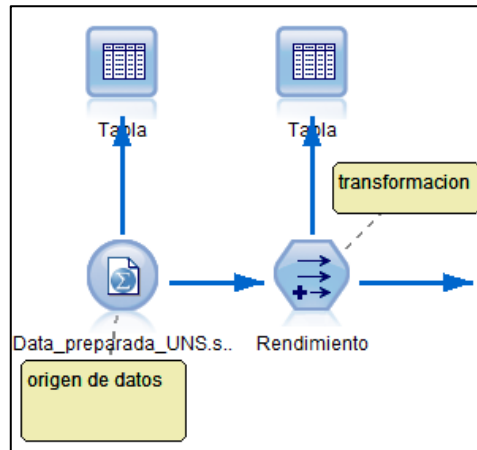


Figura 19: Transformación de datos con nodo derivar - SPSS Modeler

Fuente: Elaboración propia

Finalmente se conectó el nodo de origen de datos hacia el nodo derivar de tal manera que se realiza la transformación del promedio ponderado. A continuación, una imagen de muestra.

	al_familiar	Tipo_establecimiento_estudio_primario	Tipo_establecimiento_estudio_secundario	Promedio_ponderado	Rendimiento
1	800.000	2	2	18.000	4
2	600.000	1	1	18.000	4
3	7340.000	2	2	16.400	3
4	1200.000	2	2	16.000	3
5	2000.000	1	2	16.000	3
6	1500.000	2	2	16.000	3
7	900.000	1	1	16.000	3
8	900.000	2	2	15.210	3
9	1500.000	2	1	15.000	3
10	2000.000	2	2	15.000	3
11	3500.000	2	2	15.000	3
12	1200.000	1	1	15.000	3
13	2500.000	1	1	15.000	3
14	2500.000	1	1	14.600	3
15	1500.000	2	2	14.500	3
16	5000.000	1	1	14.000	3
17	1000.000	1	1	14.000	3
18	2500.000	2	2	14.000	3
19	1500.000	1	1	14.000	3
20	1000.000	1	2	14.000	3

Figura 20: Ejemplo de la transformación de datos - SPSS Modeler

Fuente: Elaboración propia

Luego se utilizó el “nodo Tipo”, con la finalidad de elegir nuestras variables de tipo entrada y destino, así mismo se cargaron los valores numéricos para cada variable.

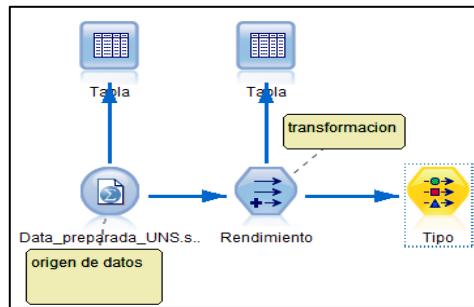


Figura 21: Aplicación del nodo tipo - SPSS Modeler

Fuente: Elaboración propia

Campo	Medida	Valores	No se enc...	Comprobar	Rol
Dependiente...	Nominal	1,2	Ninguno		Entrada
Nivel_educ...	Nominal	1,2,3,4	Ninguno		Entrada
Nivel_educ...	Nominal	1,2,3,4	Ninguno		Entrada
Reaccion_pr...	Nominal	1,2,3	Ninguno		Entrada
Influencia_st...	Nominal	1,2,3,4,5	Ninguno		Entrada
Nivel_empat...	Nominal	1,2,3,4,5	Ninguno		Entrada
Horas_sem...	Continuo	[0,70]	Ninguno		Entrada
Ingreso_me...	Continuo	[0,0,1500...	*	Ninguno	Entrada
Tipo_establ...	Nominal	1,2	Ninguno		Entrada
Tipo_establ...	Nominal	1,2	Ninguno		Entrada
Promedio_p...	Continuo	[11,0,18,0]	*	Ninguno	Entrada
Rendimiento	Nominal	2,3,4	Ninguno		Destino

Figura 22: Ejemplo del nodo tipo - SPSS Modeler

Fuente: Elaboración propia

Después se usó el “nodo Filtro”, para deshabilitar aquellas variables no influyentes.

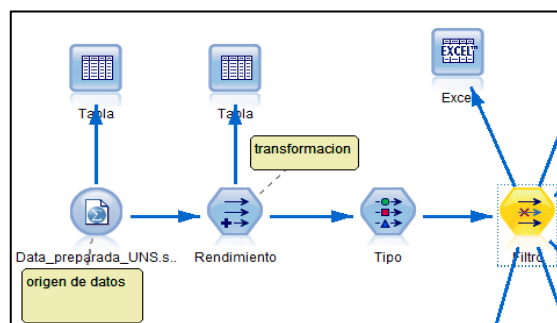


Figura 23: Aplicación del nodo filtro - SPSS Modeler

Fuente: Elaboración propia



#### Etapa 4: Minería de datos

En esta etapa, se eligió distintos algoritmos de aprendizaje automático con la finalidad de comparar resultados y obtener el mejor modelo predictivo. En esta investigación se usó árbol de decisión, máquina de vectores y K-NN. Estos algoritmos son proporcionados por la herramienta SPSS Modeler.

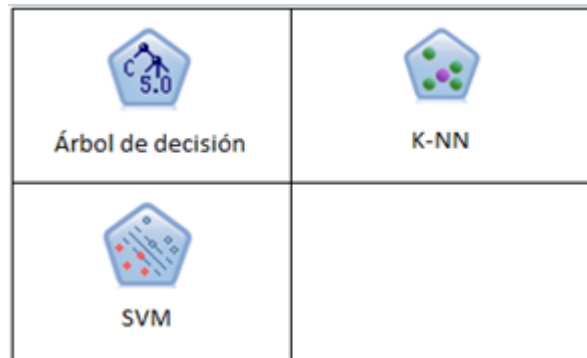


Figura 24: Algoritmos de aprendizaje automático - SPSS Modeler

Fuente: Elaboración propia

A continuación, se muestra los modelos creados por cada algoritmo de aprendizaje:

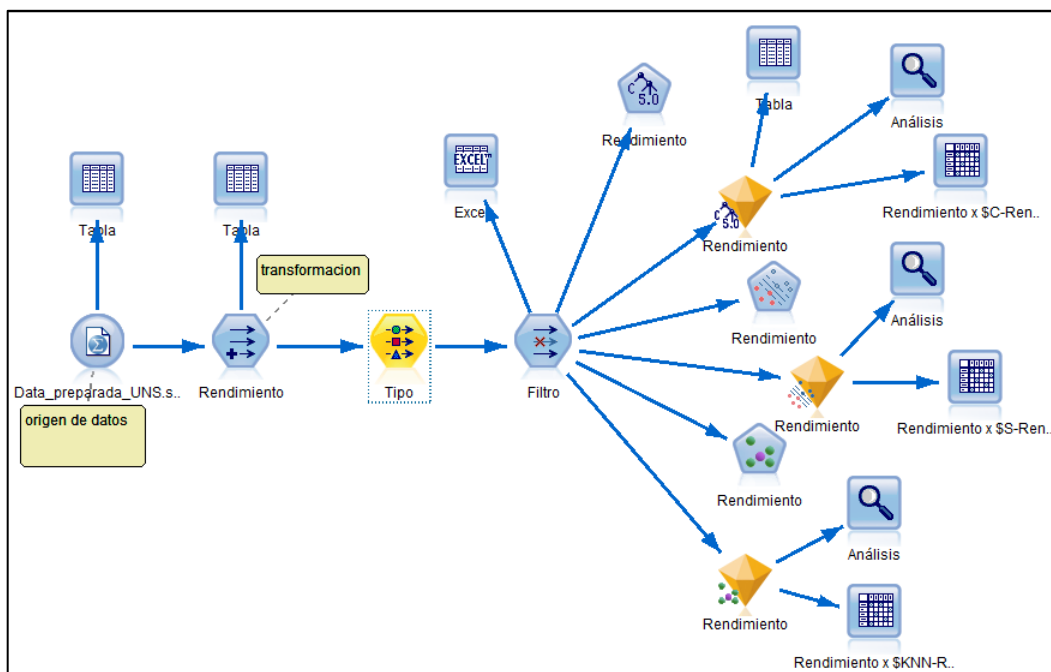


Figura 25: Proyecto de rendimiento académico - SPSS Modeler

Fuente: Elaboración propia

## Etapa 5: Interpretación

Árbol de decisión – C5, la herramienta nos proporciona información sobre que variables son de mayor relevancia con relación a la variable rendimiento. Además, nos muestra que el modelo alcanzó una precisión de 85.06% y una tasa de error de 14.94%

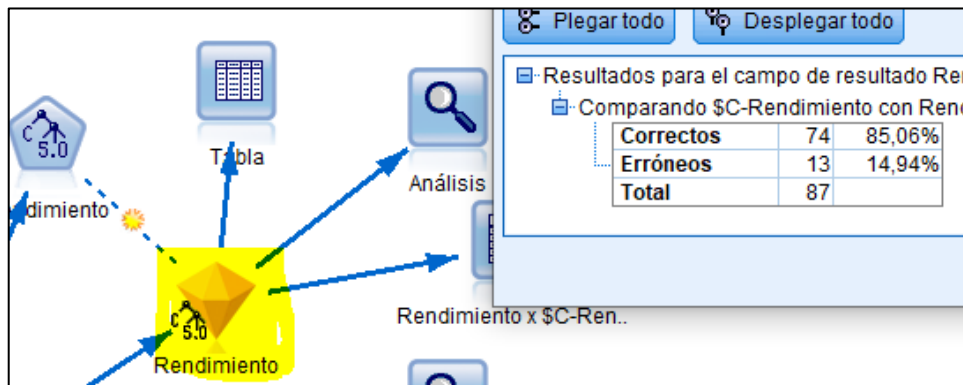


Figura 26: Precisión de modelo utilizando árbol de decisión - SPSS Modeler

Fuente: Elaboración propia

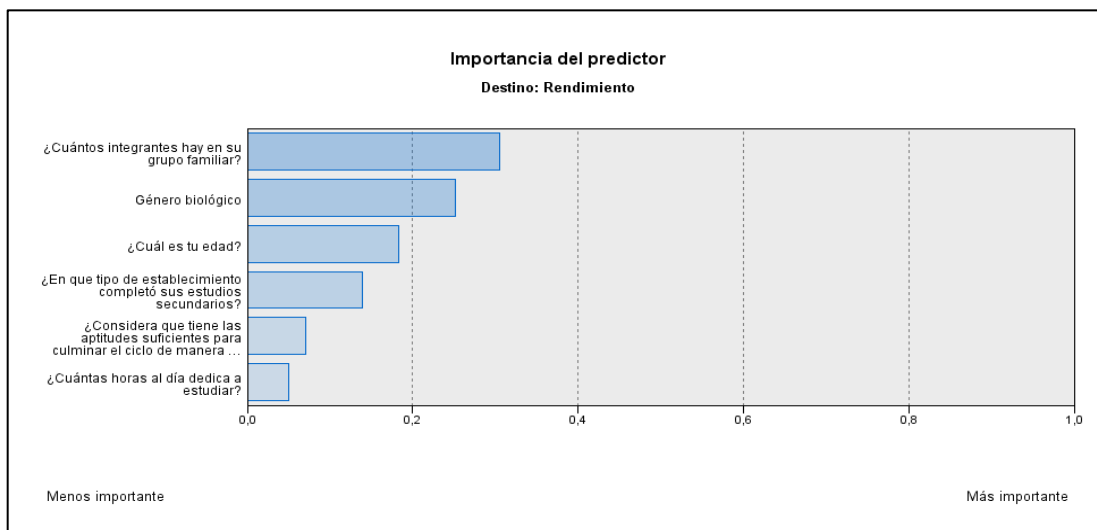


Figura 27: Variables con relevancia utilizando árbol de decisión - SPSS Modeler

Fuente: Elaboración propia

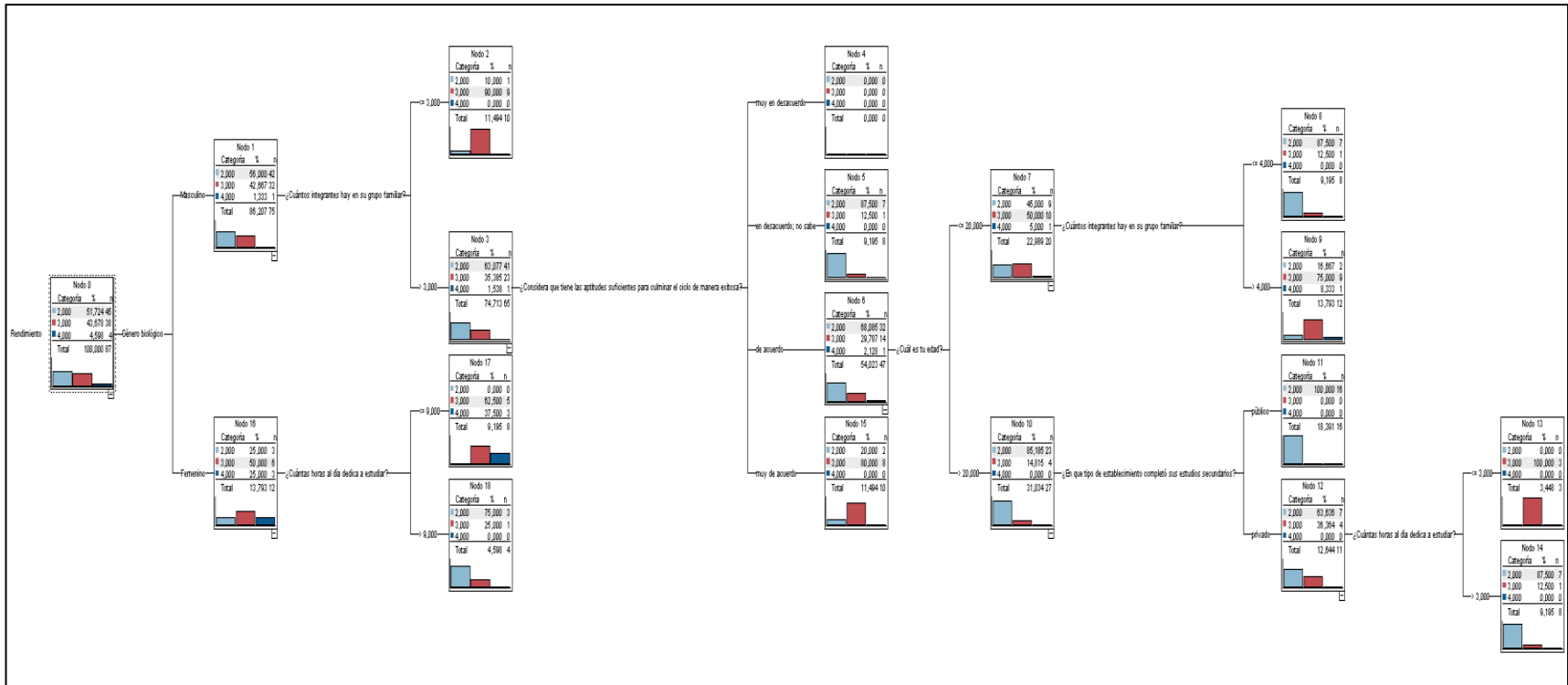


Figura 28: Árbol de decisión – SPSS Modeler

Fuente: Elaboración propia

Máquina de Vectores (SVM): A través de la herramienta nos muestra una precisión de 100 % por ende una tasa de error igual a 0

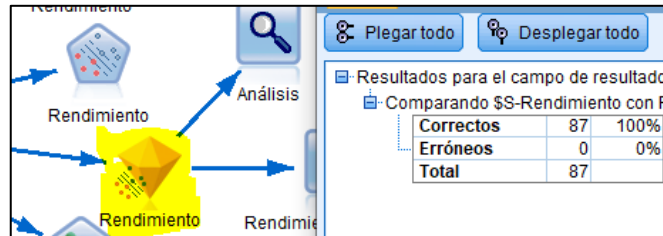


Figura 29: Precisión de modelo utilizando SVM - SPSS Modeler

Fuente: Elaboración propia

K-Vecinos: La modelo nos indica que tiene una precisión igual a 72.41% y una tasa de error de 24.59%

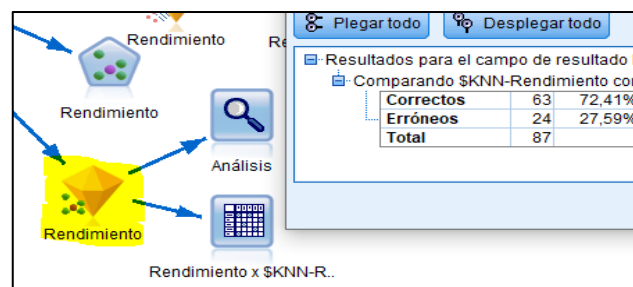


Figura 30: Precisión de modelo utilizando K vecinos - SPSS Modeler

Fuente: Elaboración propia


## Declaratoria de Originalidad del Autor

Yo, Garcia Dionisio Jeancarlos Donato, egresado de la Facultad de Ingeniería y de la Escuela Profesional Ingeniería de Sistemas de la Universidad César Vallejo Lima - Este, declaro bajo juramento que todos los datos e información que acompañan al Trabajo de Investigación / Tesis titulado: **“Machine Learning para predecir el rendimiento académico de los estudiante universitarios”**, es de mi autoría, por lo tanto, declaro que el Trabajo de Tesis:

1. No ha sido plagiado ni total, ni parcialmente.
2. He mencionado todas las fuentes empleadas, identificando correctamente toda cita textual o de paráfrasis proveniente de otras fuentes.
3. No ha sido publicado ni presentado anteriormente para la obtención de otro grado académico o título profesional.
4. Los datos presentados en los resultados no han sido falseados, ni duplicados, ni copiados.

En tal sentido asumo la responsabilidad que corresponda ante cualquier falsedad, ocultamiento u omisión tanto de los documentos como de información aportada, por lo cual me someto a lo dispuesto en las normas académicas vigentes de la Universidad César Vallejo.

Lugar y fecha,

Garcia Dionisio Jeancarlos Donato	
DNI: 70022628	Firma 
ORCID: ORCID: 0000-0001-6739-169X	