UNIVERSIDAD NACIONAL JORGE BASADRE GROHMANN

Escuela de Posgrado

DOCTORADO EN CIENCIAS DE LA EDUCACIÓN

MODELO DE MINERÍA DE DATOS PARA EVALUAR EL EFECTO DEL USO DEL AULA VIRTUAL SOBRE EL RENDIMIENTO ACADÉMICO DE LOS ESTUDIANTES DE LA FACULTAD DE INGENIERÍA DE LA UNIVERSIDAD NACIONAL JORGE BASADRE GROHMANN DE TACNA, EN TIEMPOS DE PANDEMIA, 2020

TESIS

PRESENTADA POR:

M.Sc. EDGAR AURELIO TAYA ACOSTA

Para optar el Grado Académico de:

DOCTOR EN CIENCIAS DE LA EDUCACIÓN

TACNA - PERÚ

2021

UNIVERSIDAD NACIONAL JORGE BASADRE GROHMANN

Escuela de Posgrado

DOCTORADO EN CIENCIAS DE LA EDUCACIÓN

MODELO DE MINERÍA DE DATOS PARA EVALUAR EL EFECTO DEL USO DEL AULA VIRTUAL SOBRE EL RENDIMIENTO ACADÉMICO DE LOS ESTUDIANTES DE LA FACULTAD DE INGENIERÍA DE LA UNIVERSIDAD NACIONAL JORGE BASADRE GROHMANN DE TACNA, EN TIEMPOS DE PANDEMIA, 2020

Tesis sustentada y aprobada el 07 de octubre del 2021; estando el jurado calificador integrado por:

PRESIDENTE : .....
Dr. Heber Melbin Cabrera Cruz

SECRETARIO : .....
Dr. Humberto Benito Vargas Pichón

MIEMBRO : .....
Dr. Edwin Antonio Hinojosa Ramos

ASESOR : .....
Dr. Edwin Antonio Hinojosa Ramos

AGRADECIMIENTOS

En primer lugar, agradezco a Dios nuestro creador, por guiarme y darme la luz para esta investigación.

También es justo agradecer a mi querida Alma Mater la Universidad Nacional Jorge Basadre Grohmann de la Heroica ciudad de Tacna por ser la formadora de mi capacidades y habilidades académicas y de investigación.

Quiero agradecer a mis Padres por su inmenso amor y aliento en este proceso de elaboración de la tesis.

Agradecer a mi familia, a mi hermosa esposa Leyla, a mis amados hijos Mao y Tian, por el tiempo que me permitieron alejarme de ellos, de aislarme para poder emprender esta aventura de investigación, gracias por su comprensión.

Quiero agradecer también a mi asesor y amigo Dr. Edwin Antonio Hinojosa Ramos por su excelente e importante apoyo y acompañamiento en la labor de preparación y ejecución de la presente tesis.

DEDICATORIA

El presente trabajo de investigación lo dedico a mis Padres Gladys y Edgar, a mis hijos Mao y Tian, a mi hermosa esposa Leyla, a toda mi familia y todos los amigos de verdad.

CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

La Pandemia mundial ha obligado a las universidades a implementar un sistema no presencial en las actividades académicas y la Universidad Nacional Jorge Basadre Grohmann (UNJBG) no es la excepción. Sin embargo no se tiene evidencias del éxito del uso del aula virtual respecto al rendimiento de los estudiantes, para esto se ha propuesto un Modelo basado en minería de datos para relacionar el acceso al entorno virtual de aprendizaje y el rendimiento académico en los estudiantes de la Facultad de Ingeniería en tiempos de pandemia 2020, esto se logró utilizando un marco de trabajo KDD y técnicas computacionales tanto a nivel descriptivo como de aprendizaje automático dentro del campo del Machine Learning. La presente tesis es de tipo correlacional, explicativa y predictiva, donde la población de estudio corresponde a datos de acceso al aula virtual, calificaciones y datos socioeconómicos de todos los estudiantes de la UNJBG tomando como muestra a la Facultad de Ingeniería en el periodo académico 2020-I. Se desarrolló en el marco de trabajo KDD y se siguió cada una de las etapas. Para la captura de datos se solicitó a la UNJBG acceso a los datos de uso del aula virtual en nuestro caso desde la plataforma Moodle, luego se solicitó acceso a los datos de registro académico para obtener los rendimientos académicos de los estudiantes, se hizo el análisis de los datos, procediendo a limpiar, preprocesar y transformar para que cumpla con las exigencias del modelo que se propuso. Inicialmente se ha descrito los datos para entenderlos y centrar el análisis. Luego se procedió a aplicar operaciones de combinación de los datos y de esta forma poder determinar la relación entre las distintas variables, como por ejemplo el número de accesos al aula virtual y el rendimiento académico. Para el aprendizaje automático se dividió los datos en dos partes: un grupo de datos para entrenamiento correspondiente al 80% y un grupo de datos para el testeo de 20% del total de datos. Se utilizó la matriz de confusión para evaluar la precisión de la clasificación, logrando

implementar un modelo basado en el algoritmo Gradient Boosted Trees que fue el que mejor desempeño tuvo al clasificar el rendimiento académico con valores de dos clases (SATISFACTORIO y DEFICIENTE) con una precisión de 91,79%, también se logró desarrollar un modelo basado en el algoritmo de Random Forest que obtuvo la mayor precisión para clasificar el rendimiento académico con valores de tres clases (SATISFACTORIO, REGULAR y DEFICIENTE) con una precisión de 89,26%.

Palabras Clave: Minería de Datos, Aprendizaje automático, KDD, MLP, Deep learning.

ABSTRACT

The global pandemic has forced universities to implement a remote system in academic activities and the Jorge Basadre Grohmann National University (UNJBG) is no exception. However, there is no evidence of the success of the use of the virtual classroom with respect to student performance, for this a Model based on data mining has been proposed to relate access to the virtual learning environment and academic performance in students of the Faculty of Engineering in times of pandemic 2020, this was achieved using a KDD framework and computational techniques at both a descriptive and machine learning level within the field of Machine Learning. This thesis is correlational, explanatory and predictive, where the study population corresponds to access data to the virtual classroom, grades and socioeconomic data of all UNJBG students, taking as a sample the Faculty of Engineering in the academic period 2020 -I. It was developed in the KDD framework and each of the stages was followed. For data capture, the UNJBG was requested to access the data on the use of the virtual classroom in our case from the Moodle platform, then access to the academic record data was requested to obtain the academic performance of the students, the analysis was made of the data, proceeding to clean, preprocess and transform so that it complies with the demands of the model that was proposed. Initially, the data has been described to understand them and focus the analysis. Then we proceeded to apply data combination operations and thus be able to determine the relationship between the different variables, such as the number of accesses to the virtual classroom and academic performance. For machine learning, the data was divided into two parts: a data group for training corresponding to 80 % and a data group for testing 20 % of the total data. The confusion matrix was used to evaluate the accuracy of the classification, managing to implement a model based on the Gradient Boosted Trees algorithm, which was the one with the best performance when classifying academic

performance with values of two classes (SATISFACTORY and POOR) with precision of 91,79 %, it was also possible to develop a model based on the Random Forest algorithm that obtained the highest precision to classify academic performance with values of three classes (SATISFACTORY, REGULAR and POOR) with a precision of 89,26 %.

Keywords: Data Mining, Machine Learning, KDD, MLP, Deep learning.

RÉSUMÉ

La pandémie mondiale a contraint les universités à mettre en œuvre un système à distance dans les activités académiques et l'Université nationale Jorge Basadre Grohmann (UNJBG) ne fait pas exception. Cependant, il n'y a aucune preuve du succès de l'utilisation de la classe virtuelle en ce qui concerne les performances des étudiants, pour cela un modèle basé sur l'exploration de données a été proposé pour relier l'accès à l'environnement d'apprentissage virtuel et les performances académiques des étudiants de la Faculté d'ingénierie de en période de pandémie 2020, cela a été réalisé à l'aide d'un cadre KDD et de techniques de calcul à la fois au niveau descriptif et de l'apprentissage automatique dans le domaine de l'apprentissage automatique. Cette thèse est corrélationnelle, explicative et prédictive, où la population étudiée correspond aux données d'accès à la classe virtuelle, aux notes et aux données socio-économiques de tous les étudiants de l'UNJBG, en prenant comme échantillon la Faculté d'ingénierie dans la période académique 2020 -I. Il a été développé dans le framework KDD et chacune des étapes a été suivie. Pour la saisie des données, il a été demandé à l'UNJBG d'accéder aux données sur l'utilisation de la classe virtuelle dans notre cas à partir de la plateforme Moodle, puis l'accès aux données du dossier académique a été demandé pour obtenir les performances académiques des étudiants, l'analyse a été faite de les données, en procédant au nettoyage, au prétraitement et à la transformation afin qu'elles soient conformes aux exigences du modèle qui a été proposé. Initialement, les données ont été décrites pour les comprendre et cibler l'analyse. Ensuite, nous avons procédé à l'application d'opérations de combinaison de données et ainsi être en mesure de déterminer la relation entre les différentes variables, telles que le nombre d'accès à la classe virtuelle et les performances académiques. Pour le machine learning, les données ont été divisées en deux parties : un groupe de données pour l'entraînement correspondant à 80 % et un groupe de données pour tester 20 % des données totales. La matrice de confusion a été utilisée pour évaluer l'exactitude de la classification, en réussissant à implémenter un modèle

basé sur l'algorithme Gradient Boosted Trees qui était celui qui avait les meilleures performances lors de la classification des performances académiques avec des valeurs de deux classes (SATISFAISANT et MAUVAIS) avec une précision de 91,79 %, il a également été possible de développer un modèle basé sur l'algorithme Random Forest qui a obtenu la plus grande précision pour classer les performances académiques avec des valeurs de trois classes (SATISFAISANT, RÉGULIER et MAUVAIS) avec une précision de 89,26 %.

Mots-clés : Data Mining, Machine Learning, KDD, MLP, Deep learning

INTRODUCCIÓN

La educación superior universitaria constituye una actividad importante en el desarrollo de cualquier comunidad, ya que se caracteriza por la generación de nuevo conocimiento, la formación profesional y la vinculación responsable con el entorno. Es decir podemos afirmar que la universidad es la conciencia de la sociedad y la extensión universitaria el apoyo fundamental en la formación de dicha conciencia (Serna Alcantara, 2007).

La educación universitaria no presencial, ha experimentado en los últimos años un incremento importante de alrededor del 5 % convirtiéndose en una propuesta importante en el quehacer académico universitario, prestigiosas universidades están ofreciendo formación a distancia en todos sus niveles, tanto de pregrado y posgrado: MOOCs (Massive Online Open Courses). (Martínez, 2017). Además surge como un forma de hacer frente a las brechas de desigualdad de oportunidades, aportando a la inclusión de aquellas personas que por diversos motivos no han podido seguir sus estudios y debieron abandonar la universidad en el formato convencional de esta (Escanés et al., 2014)

Existen sin embargo ideas contradictorias y una vigente discusión sobre la pertinencia de la educación no presencial, generando debates en todos los estamentos, tendiendo en general un pronóstico de coexistencia donde se potenciarían mutuamente, aunque existen ventajas y desventajas como la sincronía y a sincronía (Martínez, 2017)

Con la llegada de la pandemia mundial provocada por el virus Sars-COV2, ha remecido los cimientos de las universidades del mundo y sorprendió en distinta medida a ellas, a unas más que a otras. En el contexto latinoamericano

nuestras universidades en su mayoría no estaban preparadas para atender esta demanda tan repentina de formación no presencial. En el caso del Perú existen limitaciones de todo tipo, especialmente referidas a la capacitación docente, la infraestructura tecnológica y la conectividad que han impedido poder afrontar de la mejor manera esta contingencia.(Velazque Rojas et al., 2020)(Sohrabi et al., 2020)

Este nuevo escenario ha desencadenado diversos problemas colaterales, entre ellos, la afectación al rendimiento académico, la deserción y algunos fenómenos que son interesantes estudiar como el nivel de uso de las aulas virtuales, la gestión de recursos y actividades en los entornos virtuales de aprendizaje.

Esta pandemia ha afectado considerablemente en varias partes del mundo al proceso educativo, especialmente a la educación universitaria, generando niveles altos de deserción estudiantil generada por el bajo rendimiento académico producto del entorno difícil que se presenta (García et al., 2016) y (Esteban et al., 2016).

En este contexto se ha incrementado exponencialmente la generación de datos transaccionales en las diferentes plataformas o entornos de aprendizaje virtual(Shah et al., 2021)(Devasia et al., 2016) que ahora son utilizados para la educación no presencial determinada por restricciones impuestas por los gobiernos en busca de disminuir la sobrecarga del sector salud (America, n.d.).

Además, este incremento de los valores transaccionales de accesos a los EVAs nos brinda muchos datos que son importantes de analizar y relacionar para poder aprender de ellos y predecir el comportamiento futuro, esto gracias a la minería de datos.(Hardman et al., 2013).

Por otro lado creemos que la predicción del rendimiento académico es una de las tareas más importantes dentro del campo del Learning Analytics (LA) y el Educational Data Mining (EDM)(Chatti et al., 2012)(Peña-Ayala, 2014). Ya que nos permitirá conocer el comportamiento académico de los estudiantes y de esta forma se pueden convertir en herramientas de alerta temprana y poder tomar decisiones de ajuste y fortalecimiento para disminuir el fracaso académico de los estudiantes en tiempos de pandemia.

Los algoritmos de aprendizaje automático permiten el tratamiento intensivo de datos, además de aprender de ellos para luego poder predecir comportamientos futuros, en especial de variables relacionadas al acceso, permanencia e interacción con los entornos virtuales de aprendizaje (Ramírez, 2018).

En este trabajo de investigación se ha utilizado algoritmos de clasificación como Naive Bayes, Árboles de Decisión, Random Forest, Gradient Boosted Trees, regresión logística y SVM, que nos permitieron proponer un modelo con la mayor precisión para clasificar automáticamente el progreso académico de los estudiantes universitarios.

En el presente trabajo también se ha utilizado el KDD (Knowledge Discovery from Data) (Moulet y Kodratoff, 1995) como marco de trabajo para diseñar un modelo de minería de datos que nos ha permitido capturar los datos, analizarlos, limpiarlos, transformarlos, relacionarlos, visualizarlos. Del mismo modo se ha utilizado algoritmos de Machine Learning para poder predecir el comportamiento del rendimiento académico en función a un conjunto de variables relacionadas al acceso del aula virtual y las condiciones socioeconómicas de los estudiantes de la Universidad Nacional Jorge Basadre Grohmann.

En el capítulo I se ha descrito la realidad problemática origen de esta tesis, así como los objetivos e hipótesis. En el capítulo II se ha revisado el estado del arte a través de antecedentes nacionales e internacionales obtenidos de publicaciones de alto impacto, además realizamos un compendio de la teoría necesaria para conceptualizar nuestra investigación. En el capítulo III, se ha descrito y detallado la metodología empleada en el marco de KDD. En el capítulo IV se ha desarrollado el marco filosófico que servirá como base epistemológica del trabajo de investigación. En el capítulo V se muestran los resultados obtenidos tanto a nivel descriptivo, del modelo de minería de datos e inferencial. En el capítulo VI se realiza el contraste de los resultados con los obtenidos en algunas de las investigaciones más importantes de nuestros antecedentes descritos en la primera parte de la presente tesis, con un sentido crítico y analítico. Finalmente se plantea nuestras conclusiones y recomendaciones resultantes de la tesis.

CAPÍTULO I

PLANTEAMIENTO DEL PROBLEMA

1.1. DESCRIPCIÓN DEL PROBLEMA

En el actual contexto muchas actividades han migrado sus operaciones a escenarios no presenciales que han configurado condiciones propicias para la utilización de aplicaciones informáticas de todo tipo(José, 2020). Esto conlleva en consecuencia a un incremento exponencial de datos obtenidos por estas aplicaciones, lo que nos presenta una oportunidad imperdible de utilizarlos para realizar procesos de analítica de datos, que puedan permitirnos extraer conocimiento de ellos y podamos tomar mejores decisiones, sobre todo en el campo de la educación superior universitaria (Formia, 2012). Esto datos son de enorme utilidad si se realiza un análisis científico profundo, con las técnicas algorítmicas adecuadas procedentes de las disciplinas de Machine Learning, Data Mining, Big Data o Data Analytics, y se diseñan modelos capaces de describir el impacto de distintos fenómenos(Caraballo, 2020).

En el caso de la Universidad Nacional Jorge Basadre Grohmann (UNJBG) resulta interesante y fundamental realizar este estudio ya que no existe ninguna investigación que evalúe esta relación entre uso del aula virtual y rendimiento académico en tiempos de pandemia y de esta manera reformular las políticas de educación no presencial.

Entender este efecto ayudará a diseñar políticas específicas, de educación no presencial universitaria, que mitiguen las trágicas consecuencias que podrían causar no entender bien este efecto.

1.2. FORMULACIÓN DEL PROBLEMA

¿Cuál será el efecto del uso del aula virtual en el rendimiento académico de los estudiantes de la Facultad de Ingeniería de la UNJBG, en tiempos de pandemia?

1.2.1. Problemas específicos

- ¿Cómo será el análisis de los datos de uso del aula virtual de los estudiantes de la Facultad de Ingeniería de la UNJBG en el periodo académico 2020-I, utilizando un modelo de minería de datos?

- ¿Cómo será el análisis de los datos del rendimiento académico de los estudiantes de la Facultad de Ingeniería de la UNJBG en el periodo académico 2020-I, utilizando un modelo de minería de datos?

- ¿Cómo será el grado de asociación entre el nivel de uso del aula virtual y el rendimiento de los estudiantes de la Facultad de Ingeniería de la UNJBG, en el periodo 2020-I?

1.3. JUSTIFICACIÓN DE LA INVESTIGACIÓN

Es innegable que esta pandemia mundial ha modificado nuestros hábitos y costumbres y nada volverá a ser igual(Sohrabi et al., 2020).

Eso significa que las actividades de todo tipo en la sociedad deberán migrar a procesos no presenciales de preferencia, ya que posiblemente se presenten variaciones o mutaciones del virus que hagan necesarios ciclos de presencialidad y no presencialidad cada vez más frecuentes y menos distanciados.

En tal sentido la Universidad Nacional Jorge Basadre Grohmann debe preparase y generar políticas claras y efectivas para la adaptación y el aseguramiento de la calidad de los procesos formativos y de investigación que

tiene, todo esto dentro del marco legal y restricciones sanitarias para asegurar la vida de las personas.

En el campo de la educación no presencial es importante y necesario conocer que efecto tiene el nivel de uso del aula virtual en el desempeño académico de los estudiantes de la UNJBG en periodos académicos pandémicos.

En consecuencia, este trabajo de investigación se justifica plenamente ya aporta nuevas evidencias científicas sobre este fenómeno y es sumamente necesario e indispensable para nuestra institución.

La Ciencia de Datos y sus diversas técnicas como: Machine Learning, Mineria de datos, Big data, ha tenido buenos resultados a nivel nacional e internacional en diversos campos de estudio en especial para estudiar el fenómeno del COVID y su implicancia en todos los campos(Shi et al., 2021)(Davalbhakta et al., 2020)(Chakraborty et al., 2017) y es una tendencia mundial e innovadora para realizar investigaciones inéditas y que tengan un verdadero aporte a la ciencia.

1.4. OBJETIVOS

1.4.1. Objetivo general

Determinar el efecto del uso del aula virtual en el rendimiento académico de los estudiantes de la Facultad de Ingeniería de la UNJBG en tiempos de pandemia.

1.4.2. Objetivos específicos

Tenemos los siguientes objetivos específicos que pretendemos desarrollar en la presente tesis:

- Utilizar un modelo de minería de datos para analizar los datos correspondientes al uso del aula virtual de la UNJBG en el periodo académico 2020-I.

- Utilizar un modelo de minería de datos para analizar los datos correspondientes al rendimiento académico de los estudiantes de la Facultad de Ingeniería de la UNJBG en el periodo académico 2020-I.

- Utilizar un modelo de minería de datos para determinar el grado de correlación entre el nivel de uso del aula virtual y el rendimiento académico de los estudiantes de la Facultad de Ingeniería de la UNJBG en el periodo académico 2020-I.

1.5. HIPÓTESIS

1.5.1. Hipótesis general

El uso del aula virtual tendrá un efecto significativamente bajo en el rendimiento académico de los estudiantes de la Facultad de Ingeniería de la UNJBG en tiempos de pandemia.

1.5.2. Hipótesis específicas

- Será posible utilizar un modelo de minería de datos para analizar los datos correspondientes al uso del aula virtual en la UNJBG en el periodo 2020-I.

- Será posible utilizar un modelo de minería de datos para analizar los datos correspondientes al rendimiento académico de los estudiantes de la Facultad de Ingeniería la UNJBG en el periodo 2020-I.

- El grado de correlación entre el nivel de uso del aula virtual y el rendimiento académico de los estudiantes de la Facultad de Ingeniería de la UNJBG en el periodo 2020-I será significativamente bajo.

1.6. VARIABLES

1.6.1. Identificación de variables

Una variable es una propiedad que puede fluctuar y cuya variación es susceptible de medirse u observarse(Guevara et al., 2020).

En esta tesis desarrollaremos la relación entre dos variables:

Variable independiente : Uso del aula virtual.
Variable dependiente : Rendimiento académico.

1.6.2. Caracterización de las variables

A continuación, se clasificará las variables:

Por su naturaleza:
Uso del aula virtual: Cuantitativa
Rendimiento académico: Cuantitativa
Por su escala de medición:
Uso del aula virtual: Cardinal de un intervalo
Rendimiento académico: Cardinal de un intervalo.

1.6.3. Definición operacional de las variables

Variable	Definición conceptual	Dimensiones	Indicadores	Unidad	Escala
VI: Uso del aula virtual de la UNJBG	**Los datos de acceso al aula virtual por**	**Temporal**	**-Número de accesos por fecha.**	**Accesos**	**Numérica**

	parte de los estudiantes.		**-Número de accesos por hora.**	**Accesos**	**Numérica**
		Usuario	**-Número de accesos por usuario.**	**Accesos**	**Numérica**
			-Número de accesos por usuario afectado.	**Accesos**	**Numérica**
		Componente	**-Número de accesos por contexto del sistema.**	**Accesos**	**Numérica**
			-Número de accesos por componente.	**Accesos**	**Numérica**
		Evento	**-Número de accesos por nombre del evento.**	**Accesos**	**Numérica**
			-Número de accesos por descripción del evento.	**Accesos**	**Numérica**
		Localización	**-Número de accesos por origen.**	**Accesos**	**Numérica**
			-Número de accesos por dirección IP.	**Accesos**	**Numérica**
VD: Rendimiento académico	**El promedio final**	**Rendimiento académico**	**-Nota final.**	**Calificación**	**Vigesimal**

1.7. LIMITACIONES

Con respecto al área geográfica este estudio no tendrá las limitaciones propias de espacio, ya que se trabajará íntegramente con datos y herramientas de software, teniendo en cuenta las restricciones en el contexto de esta pandemia provocada por el Sars-Cov2.

En términos temporales esta tesis está enfocada al estudio de los datos correspondientes al período académico 2020-I principalmente, aunque se ha utilizado data del período académico 2019-I también para tener un marco de referencia y realizar análisis comparativos de realidades pre-pandémicas y pandémicas en un escenario cien por ciento no presencial.

Con respecto a los métodos y técnicas empleadas el estudio se ha centrado en técnicas de conexión, extracción y preprocesamiento de datos, así como en técnicas de análisis de datos, todo esto en el marco de trabajo del KDD (Knowledge Discovery in Database).

Con respecto al financiamiento se ha recurrido a fondos propios para la ejecución de la tesis.

En términos de la obtención de los datos, tenemos la dificultad de que los datos están dispersos en distintas dependencias, además de que las estructuras de datos son distintas de los distintos conjuntos de datos.

1.8. DESCRIPCIÓN DE LAS CARACTERÍSTICAS DE LA INVESTIGACIÓN

1.8.1. Tipo de estudio

Por la naturaleza del estudio esta tesis es aplicada.

1.8.2. Nivel de investigación

Esta tesis se configura como Correlacional, explicativa y predictiva.

CAPÍTULO II

MARCO TEÓRICO

2.1. ANTECEDENTES DE ESTUDIO

En este apartado de la tesis se realizó una revisión de las investigaciones más importantes del mundo referidas al tema, así es que ha recolectado trabajos muy interesantes de los mejores repositorios indexados que se pudo tener acceso como: IEEE Xplore, ACM Digital Library, Springer Link entre otros.

El trabajo titulado “Clarify of the Random Forest Algorithm in an Educational Field”(Ahmed y Hikmat Sadiq, 2018) se enfoca en utilizar el algoritmo de clasificación de Random Forest para extraer información útil de un dataset de estudiantes y realizar la predicción de su progreso académico. Para esto el dataset fue trabajado con Weka Tools y el modelo propuesto por los autores. Los resultados muestran que la precisión del método propuesto es 83,56 % y la precisión de la técnica en WEKA Tool obtuvo una precisión de 80,82 %. Además, el método propuesto permite ser ejecutado varias veces con diferentes resultados debido a la naturaleza del Random Forest, que cada vez obtiene una muestra aleatoria del dataset, esto permite escoger la mejor ejecución.

En el trabajo titulado “Predicting Student Academic Performance using Support Vector Machine and Random Forest” (Alamri et al., 2020) los autores realizaron una interesante comparación de sus modelos de clasificación binaria y regresión en la predicción de la performance académica de los estudiantes en las asignaturas de matemática y Portugués (nivel educativo de secundaria) del modelo propuesto por ellos respecto a la técnica presentada en el trabajo “Using Data Mining to predict secondary school student performance”(Cortez y Silva, 2008) obteniendo con la técnica de Support Vector Machine (SVM) 92,43 % de precisión con respecto al trabajo de Cortéz que tenía una precisión de 86,3 % y

con la técnica de Random Forest obtuvieron un 91,59 % respecto a 91,2 % del otro trabajo.

Asi mismo, en el trabajo titulado “Application of machine learning on student data for the appraisal of academic performance” (Alloghani et al., 2019) se observa que este trabajo está enfocado en aplicar técnicas de minería de datos para analizar el progreso académico de los estudiantes de 14 países del medio este de Europa incluyendo tres países del norte de África, el dataset contiene dieciséis atributos, doce de los cuales son categóricos y los cuatro restantes numéricos y se formaron 10 grupos de 10 grados distintos. Fueron usados tres algoritmos predictivos: árboles de decisión, redes neuronales y Naive bayes. Como resultado se obtuvo que el clasificador CART () 98,6 % de los estudiantes como miembros del grado G-2. De los grupos G-07 y G-08 el modelo predijo con una precisión entre 96 % y 96,6 % respectivamente. También el trabajo demuestra que la precisión del algoritmo de Naves Bayes es de 87,1 %, de las redes neuronales es de 93,1 % mientras que el algoritmo de árboles de decisión tuvo una precisión de 92,7 %.

El trabajo titulado “Student Performance Prediction using Multi-Layers Artificial Neural Networks: A Case Study on Educational Data Mining” (Altaf et al., 2019) utiliza redes neuronales para clasificar el comportamiento académico a partir de datos obtenidos del LMS Moodle de alrededor de 900 estudiantes de 10 clases universitarias. Lo interesante de este estudio es que entrena cada red neuronal con cada clase, lo que demuestra la importancia de la predicción individual con respecto a la precisión de la predicción. También se determinó que las características más importantes en la clasificación fueron: el grado y el total de sesiones de aprendizaje. En lo que se refiere al rendimiento de las predicciones con redes neuronales se observó una precisión de: 74,3 %, 75,8 %,97,1 %,80 %,90 %,65,2 %,95 %,92,1 %,83,1 % y 90 % para los cursos de Introducción a la computación, Fundamentos de programación, Ética Profesional, Habilidades de comunicación, Cálculo multivariable, Análisis

numérico, Inteligencia artificial, Ingeniería Web, Programación de sistemas, Programación Visual respectivamente.

En el trabajo titulado “Student Performance Predictor using Multiclass Support Vector Classification Algorithm” (Athani et al., 2018), se tomó los datos de las calificaciones de escuelas portuguesas y se las agrupó en cinco niveles desde la A hasta la F, donde A es el grupo de estudiantes con las mejores notas y F el grupo de estudiantes que desaprobaron. Se implementaron varios algoritmos para clasificar como: Máquinas de Soporte Vectorial Multiclase y redes neuronales utilizando Weka Tools. Se pudo evidenciar que las Máquinas de Soporte Vectorial Multiclase fueron las que una mejor precisión usando cross validation con un valor de 89 %.

El trabajo titulado “Interpretable Deep Learning for University Dropout Prediction”(Baranyi et al., 2020) diseña un modelo de predicción basado en redes neuronales profundas para predecir el rendimiento académico final de los estudiantes de la Universidad de Tecnología y Economía de Budapest con la finalidad de identificar los estudiantes con riesgo de deserción, la precisión que lograron fue de 72,4 % (AUC=0,771) lo que demostró que el aprendizaje profundo es adecuado para la predicción de abandonos.

El trabajo titulado “Features Exploration for Grades Prediction using Machine Learning”(Bouchard et al., 2020) presenta el procesamiento de una data muy grande de estudiantes de la Junta escolar de Quebec y aplica algoritmos de clasificación para predecir la nota final de los estudiantes, se probaron con varias características. En promedio se tuvo alrededor de 75 % de precisión.

En el trabajo titulado “Learning Models for Student Performance Prediction”(Cavazos y B, 2018), se analizan registros de escuelas mexicanas en tres periodos: 2014-2017,2015-2018 y 2016-2019 en 24 asignaturas

evidenciando que los aspectos familiares y motivacionales son características importantes cuando se tiene acceso a los históricos de notas. Se obtuvieron los siguientes resultados para la predicción de calificaciones cuando se tiene información histórica da calificaciones: con regresión lineal se obtiene un MAE de 8,9 en el 3er bimestre, de 8,01 en el cuarto bimestre y 10,24 en el 5 bimestre, con redes neuronales se obtiene un MAE de 13,99 en el 3er bimestre, 9,32 en el cuarto bimestre, 11,85 en el 5to bimestre y con SVM un MAE de 8,19 en el 3er bimestre, 6,15 en el 4to bimestre y 8,71 en el 5to bimestre.

En el trabajo titulado “Predicting academic performance of university students from multi-sources data in blended learning”(Chango et al., 2019), se utilizan diferentes algoritmos de clasificación para predecir el rendimiento académico de estudiantes de ingeniería en entornos mixtos (presenciales y no presenciales) de aprendizaje de 65 estudiantes del primer año de la carrera de Ingeniería eléctrica de la Universidad de Córdoba (España). En general se obtuvieron buenos rendimientos de los algoritmos: la precisión entre 73 % y 82 %, la medida F entre 0.72 y 0.82 y el ROC entre 0.80 y 0.97, pero encuentran que el PART (Algoritmo parcial de árboles de decisión) definitivamente es el mejor algoritmo.

En el trabajo titulado “Student Performance Prediction Model for Early-Identification of At-risk Students in Traditional Classroom Settings”(Chanlekha y Niramitranon, 2018) se enfocan en comparar el rendimiento de modelos de predicción para identificar que estudiantes tienen tendencia en obtener bajas notas. Los datos fueron obtenidos de la oficina de registro central de la Universidad de Kasetsart que contiene notas e información demográfica de la Facultad de Ingeniería de 10 años entre 2008 y 2017. Se obtuvieron los siguientes resultados en términos de precisión y el valor r(correlación): para el curso de Cómputo y programación con el algoritmo de árboles de decisión fue de 48,56 % y 56,92 % respectivamente, con el algoritmo de Naive bayes fue de 35,21 % y 20,16 % respectivamente, para el algoritmo de Random forest fue de

48,77 % y 62,06 % respectivamente, para el algoritmo de SVM fue de 48,15 % y 47,43 % respectivamente. Para el curso de Matemáticas para ingeniería I con el algoritmo de árboles de decisión fue de 54,71 % y 83 % respectivamente, para el algoritmo de Naive Bayes fue de 50,79 % y 53,71 % respectivamente, para el algoritmo de Random Forest 54,89% y 81,92% respectivamente, para la SVM 55,27 % y 85,35 % respectivamente, para redes neuronales 56,20 % y 83,18 % respectivamente. Para el curso de Ingeniería Mecánica I con el algoritmo de árboles de decisión fue de 61,01 % y 59,8 % respectivamente, para el algoritmo de Naive Bayes fue de 58 % y 62,81 % respectivamente, para el algoritmo de Random Forest 60,81 % y 59,8 % respectivamente, para la SVM 60,4 % y 65,33 % respectivamente, para redes neuronales 60,81 % y 64,32 % respectivamente. Para el curso de Ingeniería Mecánica II con el algoritmo de árboles de decisión fue de 53,02 % y 78,05 % respectivamente, para el algoritmo de Naive Bayes fue de 48,66 % y 0 % respectivamente, para el algoritmo de Random Forest 52,68 % y 76,83 % respectivamente, para la SVM 51,68 % y 80,49 % respectivamente, para redes neuronales 56,04 % y 76,83 % respectivamente y para el curso de Análisis y Diseño de Algoritmos con el algoritmo de árboles de decisión fue de 65 % y 100 % respectivamente, para el algoritmo de Naive Bayes fue de 57,5 % y 100 % respectivamente, para el algoritmo de Random Forest 62,5 % y 100 % respectivamente, para la SVM 57,5 % y 1 % respectivamente, para redes neuronales 62,5 % y 100 % respectivamente.

En el trabajo titulado “Predicting student performance using data from an Auto-grading system”(H. Chen y Ward, 2020) el grupo de investigación a cargo experimenta construyendo modelos de regresión lineal y árboles de decisión con características de la data obtenida del sistema de autoevaluación Marmoset de la Universidad de Waterloo donde incluye ratios de aprobación, resultados de pruebas, número de tareas enviadas e intervalos entre envíos. Los algoritmos de clasificación los usaron para predecir las categorías de los estudiantes y la regresión para predecir las notas de los exámenes intermedios y finales, lo que aquí en la UNJBG vendría a ser primera y segunda evaluación. Se utilizaron 428

registros de estudiantes quienes tienen notas de exámenes intermedios y finales de la ECE 150 del otoño del año 2016. Obteniendo para la regresión un p-valor menor de 2,2e-16 que es mucho más pequeño que 0,05 que indica que están relacionadas las variables: intervalo de tiempo entre entregas y la nota intermedia. De igual forma se obtuvieron resultados para la nota final donde se evidenció relación entre el intervalo de tiempo de entregas y la nota final.

2.2. BASES TEÓRICAS

2.2.1. Rendimiento académico

El rendimiento académico en nuestra investigación se refiere a la performance del estudiante en término de calificaciones de las diferentes asignaturas, promediadas en una sola nota final por cada estudiante en el período académico correspondiente.

2.2.2. Learning Analytics

Se refiere a la interpretación de una gran cantidad de datos producidos y obtenidos de diversas fuentes con la finalidad de evaluar el rendimiento académico, predecir la performance y detectar problemas potenciales. Desde su primera mención, Learning Analytics(Johnson et al., 2012) ha ganado una relevancia cada vez mayor. En diversas publicaciones se identificó el análisis del aprendizaje como una de las tendencias más importantes en el aprendizaje y la enseñanza mejorados por la tecnología. Por lo tanto, no es de extrañar que Learning Analytics sea el tema de muchos artículos científicos. La investigación y mejora de Learning Analytics implica hacer el desarrollo, el uso y la integración de nuevos procesos y herramientas para mejorar el desempeño de la enseñanza y el aprendizaje de estudiantes individuales y de profesores. Learning Analytics se centra específicamente en el proceso de aprendizaje(Long y Siemens, 2014). Debido a sus conexiones con la enseñanza y el aprendizaje digitales, Learning

Analytics es un campo de investigación interdisciplinario con conexiones con el campo de la investigación de la enseñanza y el aprendizaje, la informática y la estadística(Adams Becker et al., 2013). Los datos disponibles se recopilan, analizan y los conocimientos adquiridos se utilizan para comprender el comportamiento de los estudiantes y proporcionarles apoyo adicional(Gašević et al., 2015).

2.2.3. Universidad y Pandemia

La educación superior se enfrenta un escenario de incertidumbre y cambio. Además de los cambios provocados por la pandemia mundial provocada por el virus Sars-Cov2, así como cambios políticos y sociales, la competencia a nivel universitario aumenta. Las universidades comparten los mismos desafíos que las empresas: la necesidad de aumentar la eficiencia financiera y operativa, expandir el impacto local y global, establecer nuevos modelos de financiamiento durante un clima económico cambiante y responder a las demandas de una mayor responsabilidad para asegurar el éxito organizacional a todos los niveles(Jones, 2019). La Educación Superior debe superar estas cargas externas de manera eficiente y dinámica, pero también comprender las necesidades del alumnado, que representa tanto el contribuyente como el donante de este sistema (Long y Siemens, 2014).

2.2.4. Regresión lineal

Este algoritmo modela la relación que existe entre dos variables ajustando una ecuación lineal a los datos. Este modelo es muy usado cuando intentamos predecir una variable basada en otra (o que depende de otra), en nuestro caso el rendimiento académico respecto a los accesos a al aula virtual. Para esto debemos revisar el cumplimiento de seis condiciones para que el modelo tenga buenos resultados:

- Las dos variables deben ser medidas en términos continuos.
- Las variables deben tener una relación lineal entre ellas.
- Los datos no deben tener outliers significativos
- Los registros de datos deben ser independientes.
- Los registros de datos deberían cumplir la homocedasticidad.
- Los errores residuales de la regresión lineal deben distribuirse normalmente.

Una vez verificadas las condiciones, podemos aplicar el modelo de regresión de acuerdo a la ecuación 1 (Prenkaj et al., 2020a):

$$Y = a + bX \qquad (1)$$

Donde:
Y= es la variable dependiente (a predecir)
X= es la variable explicativa o independiente
b= es la pendiente de la recta
a= es la intersección

2.2.5. Support Vector Machines

El objetivo de este modelo es separar las clases de acuerdo con el margen máximo: es decir determina un clasificador lineal que maximiza la distancia entre este y los puntos más cercanos de cada clase. Este clasificador se denomina hiperplano de separación óptimo. Entonces si tenemos un set de datos de entrenamiento $(x_i, y_i)\ i \in [1, l]$ donde $x_i \in R^n$ y $y_i \in \{-1, +1\}$, SVM se encarga de optimizar el siguiente problema descrito por la ecuación 2 (Prenkaj et al., 2020a):

$$min_{w,b,\varepsilon} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \varepsilon_i \qquad (2)$$

Sujeto a $y_i(w^T\varphi(x_i)+b) \geq 1-\varepsilon_i$ y $\varepsilon_i \geq 0$. Los registros x_i de entrenamiento se mapean en un espacio dimensional superior mediante la función φ. El algoritmo SVM encuentra el hiperplano que separe con el margen máximo en este gran espacio dimensional. Finalmente, $C > 0$ representa el parámetro de penalidad del término de error.

2.2.6. Naive Bayes

Este algoritmo está basado en el teorema de Bayes. Asigna la clase más probable a un determinado registro de datos denotado por su vector de características. Según la teoría bayesiana, se supone que las características de entrada son independientes e idénticamente distribuidas para una clase en particular, lo que hace que el clasificador Bayes ingenuo sea muy adaptable a datos de alta dimensionalidad lo que hace que algoritmo Naive Bayes sea muy adecuado para datos de alta dimensionalidad. Sea $\boldsymbol{X} = \{X_1, \dots, X_n\}$ un vector de características y sean las $x_k$los valores de cada característica X_j. Además, sean las letras mayúsculas en negrita que representen un vector de características y sean las letras minúsculas en negrita que representen un vector pero de valores de características. El clasificador Bayesiano $h^*(\boldsymbol{x})$ utiliza como funciones discriminantes las probabilidades posteriores de la clase dado un vector de características. En otras palabras $f_i^*(\boldsymbol{x}) = P(C = i|\boldsymbol{X} = \boldsymbol{x})$ donde C es una variable aleatoria no observada que describe la clase de una instancia. Por consiguiente, aplicando la regla de Bayes, nosotros tenemos el enunciado que describe la ecuación 3:

$$P(C = i|\boldsymbol{X} = \boldsymbol{x}) = \frac{P(\boldsymbol{X} = \boldsymbol{x}|C = i)P(C=i)}{P(\boldsymbol{X}=\boldsymbol{x})} \qquad (3)$$

Recordemos que $P(\boldsymbol{X} = \boldsymbol{x})$es idéntica para todas las clases por lo tanto puede omitirse. Por tanto, las funciones discriminantes de Bayes son las que muestra la ecuación 4:

$$f_i^* = P(\boldsymbol{X} = \boldsymbol{x}|C = i)P(C = i) \qquad (4)$$

De aquí deducimos que el clasificador Bayesiano es igual a $h^*\big((x)\big) = argmin_i P(\boldsymbol{X} = \boldsymbol{x}|C = i)P(C = i)$, el cual encuentra la hipótesis de probabilidad máxima a posteriori (MAP) de la instancia $\boldsymbol{x}$. Debido a que la estimación de $P(\boldsymbol{X} = \boldsymbol{x}|C = i)$ es difícil cuando el espacio de características es altamente dimensional, en la práctica asumimos que las características son independientes dada la clase. Esto genera el clasificador Naive Bayes definido por la ecuación 5:

$$f_i^{NB}(\boldsymbol{x}) = P(C = i) \times \prod_{j=1}^{n} P\big(X_j\big|C = i\big) \qquad (5)$$

donde X_j asume cada valor x_k perteneciente a este dominio (Prenkaj et al., 2020a).

2.2.7. Árboles de decisión

En minería de datos, los árboles de decisión son modelos predictivos que puedes ser usado para modelos de regresión o para clasificación. Sin embargo, al considerar problemas de clasificación, un árbol de decisión es llamado árbol de clasificación. El árbol de decisión es un clasificador que realiza particiones recursivas del espacio de características. Consiste en nodos que tienen exactamente una entrada excepto el nodo raíz. Los nodos son divididos en dos categorías: nodos internos, que dividen el espacio de características en dos o más sub-espacios, y las hojas que también se llaman decisores. Además, cada hoja corresponde a una única clase que representa la etiqueta de resultado más

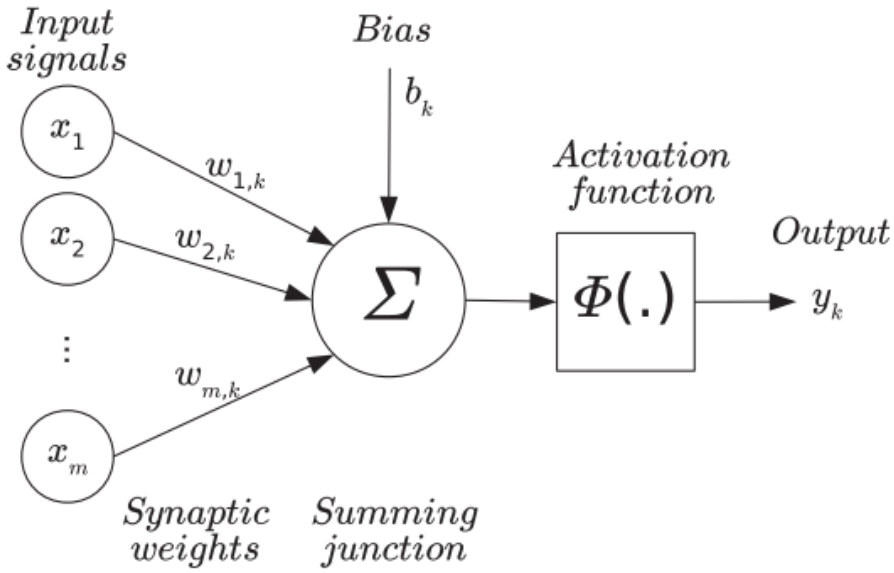
apropiada. Una instancia se clasifica rastreando sus características a través de los nodos generados por el árbol hasta los nodos hoja. En concreto, la búsqueda parte de la raíz y, según las características observadas, se elige una rama. Posteriormente, se elige el nodo correspondiente a la rama previamente seleccionada y se realiza el mismo procedimiento de elección de rama. Esto continúa hasta que se alcanza un nodo terminal hoja (Prenkaj et al., 2020a).

2.2.8. Redes Neuronales

Una red neuronal artificial, está formada por unidades de procesamiento simples llamadas neuronas interconectadas entre sí. Una neurona es la unidad de procesamiento fundamental de una red neuronal. Consta de cuatro elementos como se muestra en la Figura 1.

Figura 1

Elementos básicos de una neurona



Fuente: Extraído de “A Survey of Machine Learning Approaches for SDP in Online Courses” (Prenkaj et al., 2020b)

Donde:

- Un conjunto de enlaces ponderados conectores (sinapsis). La señal de entrada x_j cuando se transfiere es multiplicada por sus correspondientes pesos (número real) en el enlace de conexión $w_{k,j}$.
- Un combinador lineal $\sum_{j=1}^{m} w_{k,j} \times x_j$ el cual suma la entrada ponderada.
- Un término de sesgo b_k puede ser añadida a esta sumatoria. Este elemento es opcional, por lo tanto, podría valer cero.
- Una función de activación diferenciable (función de activación) que se aplica a la salida de la neurona. Limita la salida de la red neuronal. Las funciones de activación más utilizadas en la literatura son sigmoide, tangente hiperbólica y unidad lineal rectificada (RLU).

La forma más simple de una red neuronal es una única red de retroalimentación (perceptrón). Este tipo de red tiene tres capas: entrada, oculta y salida. Generalmente, las capas de entrada y salida no se cuentan, por lo que esta red es una red de una sola capa. Una generalización del feedforward único es una red de feedforward multicapa (multilayer perceptron). En este caso, hay más de una capa oculta. Las neuronas que pertenecen a las capas ocultas, denominadas neuronas ocultas, pueden adquirir más información global y resolver tareas más complejas. Las señales de entrada de una capa consisten únicamente en las señales de salida de la capa anterior. Además, todas las capas están completamente conectadas entre sí (Prenkaj et al., 2020a).

Una red neuronal aprende ajustando los pesos, los bordes y los parámetros de la función de activación seleccionada. El algoritmo utilizado para el entrenamiento es backpropagation (Hecht-Nielsen, 1989). Los pesos se inicializan aleatoriamente entre un rango específico de valores. En un entorno de aprendizaje supervisado, al predecir la etiqueta de resultado de una determinada instancia de entrenamiento, la red neuronal genera un error que es la diferencia entre la etiqueta actual y_k y la predicha $\tilde{y}_k$. La función de pérdida (continuamente

diferenciable) es la suma de los cuadrados de los errores generados $\varepsilon = \sum_k (\tilde{y}_k - y_k)^2$. El algoritmo backpropagation optimiza los pesos de la red. Por lo tanto, la red aprende a asignar entradas a salidas minimizando la función de pérdida en cada paso de entrenamiento. Específicamente, la propagación hacia atrás usa los valores de error para calcular el gradiente de la función de pérdida para encontrar el mínimo. Esta es la razón por la que la función de activación debe ser diferenciable para actualizar los pesos. Los pesos de una red neuronal son análogos a los coeficientes en un modelo de regresión lineal. Sin embargo, debido a que el número de ponderaciones en comparación con el número de coeficientes en un modelo de regresión es mayor, la interpretación de las ponderaciones en una red neuronal es ardua.

2.2.9. Métodos compuestos (ensamblados)

En lugar de depender del poder predictivo de un solo modelo, una estrategia de conjunto combina más modelos para obtener mejores resultados(Opitz y Maclin, 1999). Los modelos que construyen el método de conjunto se indican como componentes. Observe que los componentes pueden ser heterogéneos: por ejemplo, asumiendo que tenemos un conjunto de dos componentes, uno podría ser una red neuronal y el otro un árbol de decisión. Cada componente se entrena por separado. Al predecir la etiqueta de resultado de una instancia x_i, cada componente j da su predicción $\tilde{y}_{j,i}$ y se emplea un método de consenso para seleccionar la etiqueta de predicción general $\tilde{y}$. La forma más sencilla de fusionar los diferentes $\tilde{y}_{j,i}$ es seleccionando la etiqueta que los componentes han producido mejor. Esto también se conoce como consenso de votación por mayoría. Los conjuntos más populares consisten en métodos de embolsado y refuerzo. El primero se dedica a entrenar a cada clasificador en una redistribución aleatoria del conjunto de entrenamiento. El conjunto de entrenamiento de los clasificadores se genera seleccionando aleatoriamente con reemplazo tantos ejemplos como haya en el conjunto de entrenamiento original. Por el contrario, en la estrategia de impulso, el conjunto

de entrenamiento utilizado para cada clasificador se basa en el rendimiento de los clasificadores anteriores. Los ejemplos predichos incorrectamente por clasificadores anteriores ocurren con más frecuencia en el conjunto de entrenamiento del componente actual. Por lo tanto, el impulso produce nuevos clasificadores que se especializan en predecir la etiqueta de resultado de aquellas instancias que son difíciles de clasificar. Hemos identificado Random Forests y AdaBoost como los principales métodos de conjunto en la literatura. Sin embargo, algunos trabajos también se concentran en métodos conjuntos alternativos y personalizados.

2.2.10. Random Forest

Según Breiman (2005), los bosques aleatorios son una combinación de árboles de decisión de manera que cada uno depende de los valores de un vector aleatorio muestreado de forma independiente y con la misma distribución para todos los árboles del bosque. Específicamente, un bosque aleatorio es un clasificador que consta de una colección de clasificadores estructurados en árbol tal como lo muestra la ecuación 6.

$$\{h(x, \theta_k), k = 1, \dots\} \qquad (6)$$

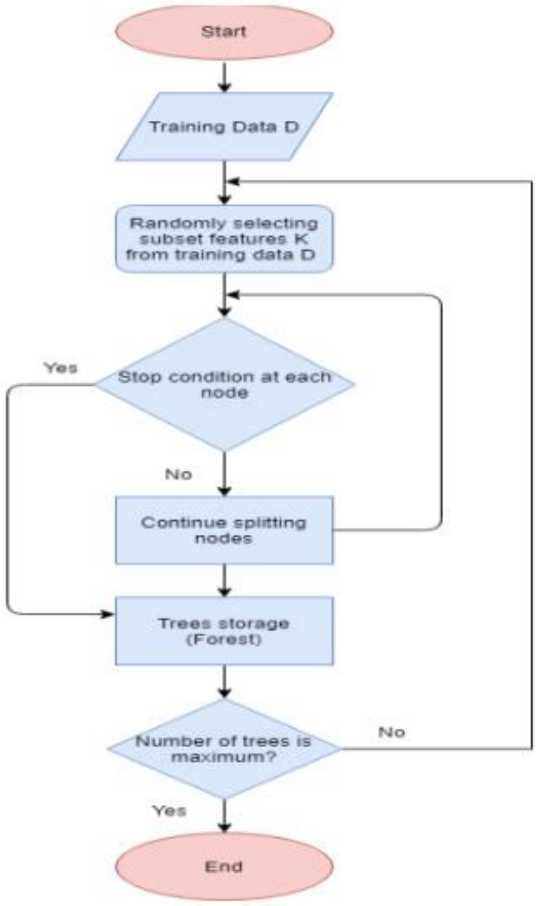
donde $\{\theta_k\}$ son vectores estocásticos, independientes e idénticamente distribuidos (i.i.d) y cada árbol genera un voto unitario para la clase más popular en la entrada x. En el bosque aleatorio, cada árbol está completamente desarrollado, lo que significa que no se emplean métodos de poda. El usuario selecciona la cantidad de características a considerar en cada nodo y la cantidad de árboles para crecer. Por lo tanto, en cada nodo, solo se buscan las entidades seleccionadas para obtener la mejor división. Cada nueva instancia se transmite a cada uno de los N árboles. El bosque elige la clase que tiene más N votos para ese caso (Pal, 2005).

Un método de aprendizaje automático que tiene la capacidad de realizar tareas de regresión y clasificación. Un clasificador de bosque aleatorio hace crecer una serie de árboles de decisión que se entrenan en diferentes partes del mismo conjunto de entrenamiento para mejorar la tasa de clasificación y superar el problema de sobreajuste.(Mahboob et al., 2017)

Random Forest elige los atributos al azar para crear un número K de árboles cada vez con diferentes atributos sin podar. En el árbol de decisión, los datos de prueba se probarán en el único árbol construido, a diferencia de Random Forest, mientras que los datos de prueba se probarán en todos los árboles construidos y luego se asignará la salida más frecuente a esa instancia (Mishra et al., 2014). Generalmente, si el bosque tiene más árboles, entonces será más robusto. Random Forest Classifier tiene la misma idea, la mejor precisión la darán las técnicas de Random Forest si tiene una mayor cantidad de árboles en el bosque. Además, los valores perdidos pueden ser manejados por Random Forest, Random Forest nunca se preocupa por la cantidad de árboles en el bosque, nunca se ajusta demasiado y Random Forest también puede manejar valores categóricos.

Figura 2

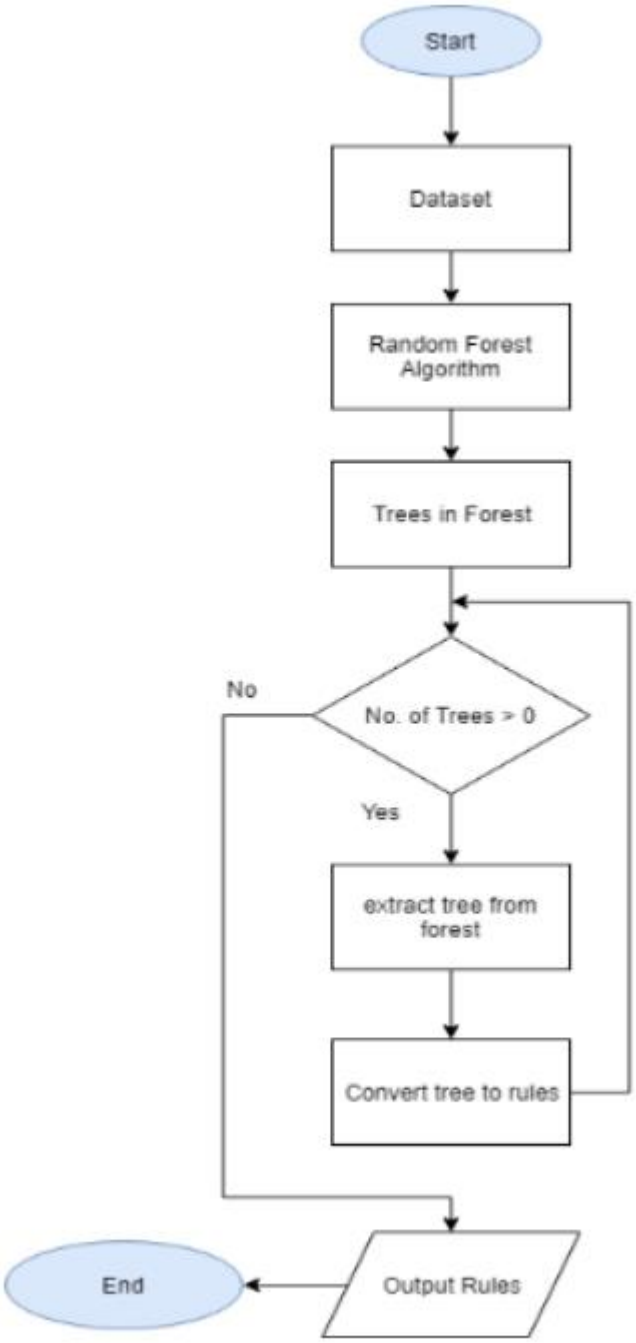
Algoritmo de Random Forest



Fuente: Mahboob, T., Irfan, S., y Karamat, A. (2017).

Figura 3

Proceso de extracción de reglas



Fuente: Mahboob, T., Irfan, S., y Karamat, A. (2017)

2.2.11. DM (Data Mining)

La minería de datos es el área de investigación científica centrada en el desarrollo de modelos para descubrir dentro de datos que provienen de entornos educativos, y el uso de esos métodos para comprender mejor el comportamiento de los estudiantes, docentes y los entornos en los que aprenden. El reciente advenimiento de los repositorios públicos de datos educativos ha hecho posible que los investigadores investiguen una amplia variedad de cuestiones científicas utilizando la minería de datos. En este artículo, se analizan cinco categorías de métodos de minería de datos educativos, así como las aplicaciones clave para las que se han utilizado métodos de minería de datos educativos (Baker, 2010; Han et al., 2012).

2.2.12. EDM (Educational Data Mining)

Cuando los datos provienen de un entorno educativo, estamos tratando con un subdominio de minería de datos llamado Minería de datos educativos o EDM. Este es un campo de investigación que aplica minería de datos, estadísticas y aprendizaje automático a datos derivados de entornos educativos. Busca extraer información significativa de grandes cantidades de datos sin procesar que se pueden utilizar para mejorar y comprender los procesos de aprendizaje (Asif et al., 2017; Baker y Yacef, 2009).

Entre los cinco enfoques más importantes de EDM: predicción, agrupamiento, minería de datos, descubrimiento dentro de modelos y descubrimiento de datos (Baker, 2010). La presente tesis combina dos enfoques: predicción y clasificación.

2.2.13. Predicción

En la predicción, el objetivo es predecir la clase o etiqueta de un conjunto de datos. Un área de aplicación clave importante de la predicción en EDM es

predecir los resultados académicos de los estudiantes. La investigación dentro de esta área se ha llevado a cabo en diferentes niveles de granularidad: a nivel de sistema de tutoría, a nivel de curso, a nivel de grado, etc. A nivel de sistema de tutoría inteligente por ejemplo, EDM predice los resultados de los exámenes de los estudiantes integrando información de tiempo y la cantidad de ayuda que un estudiante necesita para resolver problemas (Feng et al., 2006); También existen sistemas para predecir si es probable que un estudiante realice correctamente el próximo ejercicio de su examen y, de ser así, el sistema de tutoría debería omitirlo(Pardos et al., 2007). A nivel de asignaturas, existen propuestas que predicen el éxito / fracaso y la calificación de los estudiantes en un curso utilizando variables socioeconómicas como edad, sexo, estado civil, nacionalidad, domicilio, beca, habilidades diferentes, tipo de acceso a la universidad, tipo de alumno (regular, movilidad, extraordinario), situación del alumno (ordinario, empleado, deportista, etc.), años de matrícula, cursos retrasados, tipo de dedicación (tiempo completo, tiempo parcial) y situación de la deuda(Strecht et al., 2015); también encontramos trabajos que predicen las calificaciones de los estudiantes en un curso de programación considerando diferentes factores como los antecedentes matemáticos de los estudiantes, la aptitud de programación, las habilidades para resolver problemas, el género, la experiencia previa, la calificación de matemáticas de la escuela secundaria, la localidad, la experiencia previa en programación de computadoras y uso del e-learning (ElGamal, 2013); por otro lado hay iniciativas de investigación que predicen el rendimiento de los cursos sobre la base del rendimiento de los estudiantes en los cursos de requisitos previos y los exámenes parciales (Huang y Fang, 2013); otros investigaron la idoneidad de la información cuantitativa, cualitativa y de las redes sociales sobre el uso de foros, así como la idoneidad de los algoritmos de clasificación clásicos y los algoritmos de agrupamiento para predecir el éxito o el fracaso de los estudiantes en un curso (Romero et al., 2013); algunos otros investigadores proporcionan una solución de intervención temprana para cursos difíciles basada en la actividad de los estudiantes en un

sistema de gestión del aprendizaje (Arnold y Pistilli, 2012). Varios estudios predicen la aprobación / reprobación o el rendimiento académico general de los estudiantes (calificaciones totales / o parciales) al final de un programa de grado.

2.2.14. Clustering

En la agrupación, el objetivo es agrupar objetos en clases de objetos similares. Aunque la agrupación en clústeres se ha utilizado en minería de datos para una amplia variedad de tareas, una subárea interesante es agrupar a los estudiantes para estudiar patrones de comportamientos típicos. Por ejemplo existe un trabajo que encuentra comportamientos típicos en foros como trabajadores de alto nivel, es decir, estudiantes que leen todos los mensajes y publican muchos mensajes en el foro, o curiosos, es decir, estudiantes que leen todos los mensajes sin publicar ninguno (Cobo et al., 2012); otros trabajos identifican grupos de estudiantes con desempeño similar desde el jardín de infancia hasta el final de la escuela secundaria (Bowers, 2010); mientras que otros trabajos agrupan los datos de interacción de los estudiantes para construir perfiles de estudiantes (Talavera y Gaudioso, 2004).

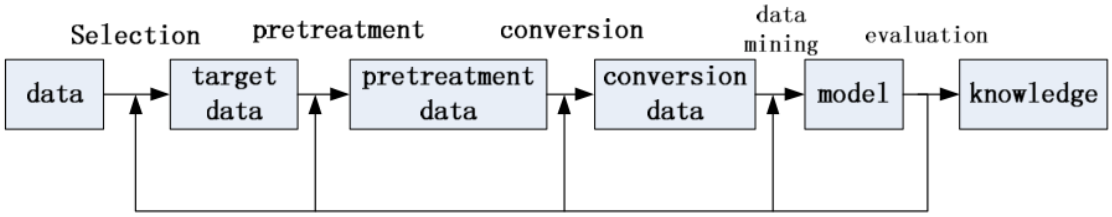
2.2.15. Descubrimiento de Conocimiento en Base de Datos (KDD)

El descubrimiento de Conocimiento en Base de Datos (KDD) se desarrolla desde un enfoque de investigación que involucra diversas áreas como bases de datos, aprendizaje automático, reconocimiento de patrones, estadística, teoría de la información, inteligencia artificial, visualización de datos (Riquelme et al., 2006). La principal característica de KDD es extraer el conocimiento de la base de datos y del almacén de datos. Este conocimiento es información implícita, previamente desconocida, potencialmente útil y de fácil comprensión. KDD es una rama de la informática desarrollada gradualmente en los últimos años y un nuevo intento en el campo de la inteligencia artificial. KDD se ha utilizado con éxito en los aspectos industriales, agrícolas, militares, financieros y comerciales,

y se ha convertido en uno de los enfoques actuales de la informática. Actualmente, KDD se utiliza frecuente como marco de trabajo en tareas de investigación descripción, evaluación del conocimiento y representación del conocimiento. El algoritmo de descubrimiento de conocimiento efectivo es la clave. Específicamente, ese es el conocimiento de la minería, como reglas de asociación, agrupamiento de datos, reglas de clasificación, patrón secuencial, modelo similar, patrón caótico, etc. con el método anterior y sus tecnologías de integración en todo tipo de base de datos del mundo real (relaciones, interpretación, temporal, espacial, distribuida, orientada a objetos). (Y. Chen et al., 2011), la secuencia de procesos del KDD es la que se visualiza en la figura 4.

Figura 4

Actividades de KDD



Fuente: Extraído de "Knowledge Discovery Technology Based on Access Information Mining on Knowledge Warehouse" (Y. Chen et al., 2011)

2.3. Definición de Términos

2.3.1. Matriz de confusión

Supongamos que tenemos r categorías $C_1, C_2, \dots, C_r$ (en nuestro caso SATISFACTORIO y DEFICIENTE) y n muestras de testeo son observadas de las categorías C_j para $j = 1, \dots, r$. Todas las muestras de testeo fueron clasificadas en dichas categorías por cierto método de clasificación y dicha clasificación se resume en una tabla de contingencia denominada matriz de error

o matriz de contingencia. Los elementos $(i,j)\ x_{ij}$ representan el número de muestras de testeo que en realidad pertenece a C y se clasifica en C para $i,j = 1,\dots,r$. De esta forma las columnas y filas de la tabla de contingencia son, respectivamente, correspondiente a la referencia (índice j) y datos clasificados (índice i) de acuerdo a la Tabla 2 (García-balboa et al., 2018):

Tabla 1

Estructura de una matriz de confusión con r-categorías.

Data Clasificada	**Data referencial**			
	C_1	C_2	…	C_r
$\boldsymbol{C_1}$	X_{11}	X_{12}	…	X_{1r}
$\boldsymbol{C_2}$	X_{21}	X_{22}	…	X_{2r}
…	…	…	…	…
$\boldsymbol{C_r}$	X_{r1}	X_{r2}	…	X_{rr}

Fuente: Extraído de "HOMOGENEITY TEST FOR CONFUSION MATRICES : A METHOD AND AN EXAMPLE"(García-balboa et al., 2018)

2.3.2. Precisión

Es la dispersión del conjunto de valores que se obtuvieron de las clasificaciones. Cuanto menor es la dispersión, mayor es la precisión. Se la puede representar como la proporción entre el número de predicciones correctas y el total de predicciones (Jalota y Agrawal, 2019).

2.3.3. Coeficiente de Correlación (r)

Corresponde la relación lineal existente entre dos conjuntos de variables, sus valores varían entre -1 y 1, cuando no existe correlación su valor cae en 0 (Martínez-Ortega y Tuya-Pendás, 2009) así como se muestra en la ecuación 7:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\ \sqrt{n\sum y_i^2 - (\sum y_i)^2}} \qquad (7)$$

donde: x_i, y_i: Valores Individuales de Datos,

n: Numero de Datos

CAPÍTULO III

MARCO METODOLÓGICO

3.1. TIPO Y DISEÑO DE LA INVESTIGACIÓN

El presente estudio es de tipo Correlacional explicativo. Se hizo una descripción de los datos de accesos y rendimiento académico de los períodos académicos 2019-I y 2020-I, para poderlas comparar y determinar diferencias.

En esta investigación para describir las diferencias entre los accesos al aula virtual del período 2019-I y 2020-I, de la misma forma se describió los datos del rendimiento académico de los períodos académicos 2019-I y 2020-I.

La investigación correlacional se encarga de determinar del grado de asociación existente entre dos o más variables en una misma muestra de sujetos, en nuestro caso de estudiantes.

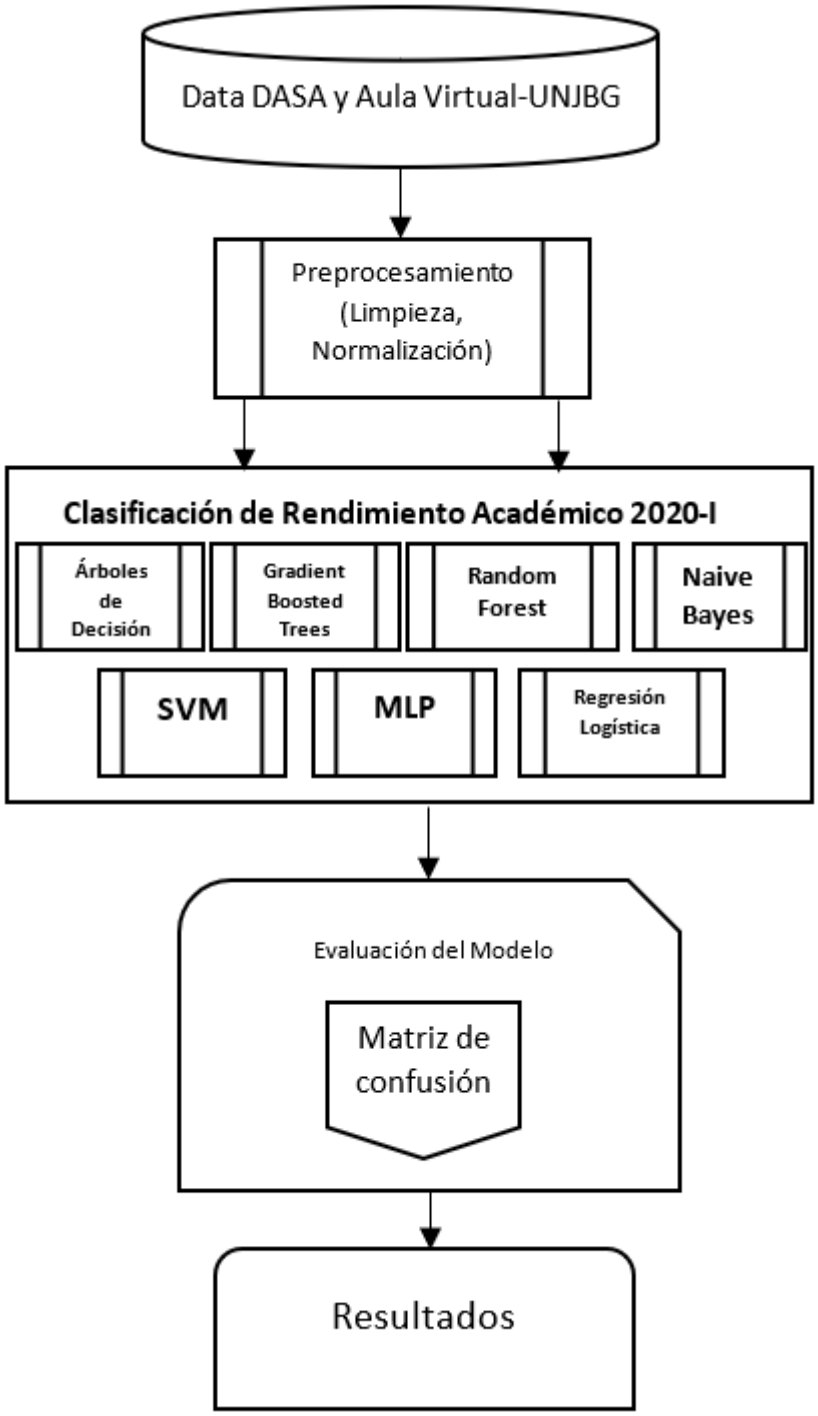
En el presente estudio se desea saber si existe asociación entre el acceso al aula virtual y el rendimiento académico en el periodo académico 2020-I.

En esencia, la investigación correlacional busca identificar probables relaciones entre las variables estudiadas: accesos al aula virtual y rendimiento académico.

En términos simples la metodología de este trabajo de investigación se resume en la figura 5.

Figura 5

Modelo propuesto para la investigación



.

3.2. POBLACIÓN Y/O MUESTRA

La población de estudio tuvo varios componentes distribuidos entre los registros de todas las calificaciones correspondientes a todos los estudiantes matriculados de la UNJBG en los períodos académico 2019-I y 2020-I que hacen un total de 130 981 registros, otro componente de nuestra población fueron los registros correspondientes a los accesos al Aula Virtual de todos los estudiantes de la UNJBG en los periodos académicos 2019-I (1 148 710 registros) y 2020-I (19 116 793 registros) que hacen un total de 20 265 503 registros, otro componente de la población fueron los 14 430 registros de los datos socioeconómicos de los estudiantes de la UNJBG, datos que fueron proporcionados por la Oficina de Registro Central de la UNJBG, del Comité de Aula Virtual y de la Dirección Académica de Bienestar Universitario de la misma Universidad correspondiente al período académico 2020-I.

Se utilizó muestreo estratificado, que es un tipo de técnica de muestreo aleatorio donde no todo se realiza al azar. Los elementos se dividen en grupos grandes que tienen una característica común. Donde de cada grupo se lista, ordena y genera una muestra por MAS (Muestro Aleatorio Simple) que puede ser del mismo tamaño o de tamaños diferentes (Rodríguez y Mendivelso, 2018). Se pudo seleccionar los registros correspondientes a la Facultad de Ingeniería en un trabajo de filtrado y combinación de los tres componentes (Calificaciones, accesos al aula virtual y datos socioeconómicos, agrupados por cada estudiante) que ascienden a un total de 3406 registros.

3.3. OPERACIONALIZACIÓN DE VARIABLES

Variable Independiente: número de accesos al aula virtual

Variable Dependiente: rendimiento académico

3.4. INDICADORES

Número de accesos al aula virtual: periodos 2019-I, 2020-I

Rendimiento académico:

- Desempeño académico 2019-I
- Desempeño académico 2020-I

3.5. TÉCNICAS E INSTRUMENTOS PARA RECOLECCIÓN DE DATOS

Como todo trabajo relacionado a la minería de datos la primera fase es la limpieza de datos, en este sentido se realizaron tareas de eliminación de valores nulos, agrupamiento, transformación de tipos de datos, trabajo con funciones de cadena de texto, combinaciones y filtrado. También se utilizaron consultas SQL para extraer registros de la base de datos del aula virtual.

3.6. PARTICIPANTES Y CONJUNTOS DE DATOS

Para esta tesis hemos utilizado los registros de 3 406 estudiantes matriculados en el período académico 2020-I de la Facultad de Ingeniería. Estos registros los podemos organizar en dos grandes rubros: la información académica: que incluyen el nombre de la escuela profesional, las notas por cada componente (conocimiento, desempeño y producto), la asignatura, la facultad, fecha de nacimiento, talla, peso, dirección, distrito, modalidad de ingreso, preparación, situación vivienda, tutores. La Tabla 3 muestra la lista de los principales atributos utilizados, sus tipos de datos y otros detalles relacionados al aspecto académico. En la Tabla 4 se muestra los atributos de los datos obtenidos de la información socioeconómica, en la Tabla 5 podemos apreciar los atributos de los datos de accesos al aula virtual. Estos registros fueron filtrados y se procedió a eliminar la información de apellidos, nombres, DNI y código de estudiantes, ya que no son de interés del estudio y además para proteger la identidad de los registros y de esta manera para cumplir con los requisitos de las obligaciones de privacidad de datos de la universidad.

Tabla 2

Atributos del conjunto de datos académico

Atributo	Tipo de Dato	Detalles
Id	*Integer*	1,2,3,…. (En total fueron…)
IdMateria	Integer	Entre 5842 y 14767
Materia	String	Taller de emprendimiento, Diseño de sistemas, etc…
Turno	String	Mañana, Tardes, Noche
Clase	String	A,B,C,D,E,F,L
IdGrado	Integer	1,2,3,…,14 (son los ciclos)
NotaFinal	*String*	Entre 1 y 20
Unidad Evaluación	String	Primer parcial, Segundo parcial, Tercer parcial, Cuarto parcial, Unico parcial, Sustitutorio.
TipoEvaluación	String	Conocimiento, Producto, Desempeño, Examen Parcial, Gabinete, Internado, Laboratorio, No especificado, Otras evaluaciones,Participación, Práctica, Producto, Sustitutorio, teoría, trabajo de investigación.
TipoPromedio	String	Artimético, Ponderado
Peso	Double	4
Nota	String	Entre 0 y 20
TipoActa	String	Acta Final
PlanCurricular	String	XXXX-F1,XXXX-F2, planes de estudio F1 y F2 de la UNJBG
Especialidad	String	Denominaciones de los 34 programas de estudio de la UNJBG
Facultad	String	Denominaciones de las 7 Facultades de la UNJBG
IdDocente	Integer	Nombre del Docente
GradoAcadémico	String	Denominación del Grado académico del Profesor
Código (#1)	String	Código del docente
Docente	String	Nombre del docente
Periodo	**String**	**2019-I,2019-II,2020-I,2020-II**

.

Tabla 3

Atributos del conjunto de datos de información personal

Atributo	Tipo de Dato	Detalles
Dni	String	Documento de identidad
Anio_ingreso	Integer	Año de ingreso
Modalidad	String	Modalidad de ingreso a la UNJBG
Prepa	String	Tipo de preparación para ingreso
sexo	String	S,M
Fnacimiento	Datetime	Año/Mes/Dia
Paisnac	String	
Dptonac	String	
distnac	String	
Tipocolegio	String	Público, Privado, Parroquial
Talla	Double	En metros
Peso	Double	En kilogramos
Imc	Índice de masa corporal	
Ojoderecho	String	XX/XX
Ojoizquierdo	String	XX/XX
Reli	String	Católica, Adventista, Mormones, Testigo de Jehová, Evangélica, Ninguno, Otros
Mat_viv	String	Material de la vivienda
Prop_viv	String	Tipo de propiedad de vivienda
Depende	String	Dependencia (Padre y Madre, Padre,Madre,Familiar/Tutor,el mismo alumno)
Sit_padres	**String**	**Casados, separados, ambos fallecidos, soltero (a), viudo(a), etc..**

Tabla 4

Atributos del conjunto de datos de uso de aula virtual 2020-I

Atributo	Tipo de Dato	Detalles
Dni	String	Documento de identidad
NRO_TOTAL_ACCESOS_2020-I_JUNIO	Integer	Accesos totales al aula virtual en Junio 2020
NRO_TOTAL_ACCESOS_2020-I_JULIO	Integer	Accesos totales al aula virtual en Julio 2020
NRO_TOTAL_ACCESOS_2020-I_AGOSTO	Integer	Accesos totales al aula virtual en Agosto 2020
NRO_TOTAL_ACCESOS_2020-I_SEPTIEMBRE	Integer	Accesos totales al aula virtual en Septiembre 2020
NRO_TOTAL_ACCESOS_2020-I_OCTUBRE	Integer	Accesos totales al aula virtual en Septiembre 2020
ACCESOS JUNIO 6 A 2PM	Integer	Accesos en el horario de 6 a 2pm de Junio
ACCESOS JULIO 6 A 2PM	Integer	Accesos en el horario de 6 a 2pm de Julio
ACCESOS AGOSTO 6 A 2PM	Integer	Accesos en el horario de 6 a 2pm de Agosto
ACCESOS SEPTIEMBRE 6 A 2PM	Integer	Accesos en el horario de 6 a 2pm de Septiembre
ACCESOS OCTUBRE 6 A 2PM	Integer	Accesos en el horario de 6 a 2pm de Octubre
ACCESOS JUNIO 2PM a 10PM	Integer	Accesos en el horario de 2pm a 10pm de Junio
ACCESOS JULIO 2PM a 10PM	Integer	Accesos en el horario de 2pm a 10pm de Julio
ACCESOS AGOSTO 2PM a 10PM	Integer	Accesos en el horario de 2pm a 10pm de Agosto
ACCESOS SEPTIEMBRE 2PM a 10PM	Integer	Accesos en el horario de 2pm a 10pm de Septiembre

ACCESOS OCTUBRE 2PM a 10PM	Integer	Accesos en el horario de 2pm a 10pm de Octubre
ACCESOS JUNIO 10PM a 6AM	Integer	Accesos en el horario de 10pm a 6am de Junio
ACCESOS JULIO 10PM a 6AM	Integer	Accesos en el horario de 10pm a 6am de Julio
ACCESOS AGOSTO 10PM a 6AM	Integer	Accesos en el horario de 10pm a 6am de Agosto
ACCESOS SEPTIEMBRE 10PM a 6AM	Integer	Accesos en el horario de 10pm a 6am de Septiembre
ACCESOS OCTUBRE 10PM a 6AM	Integer	Accesos en el horario de 10pm a 6am de Octubre
TOTAL ACCESOS A CONTEXTO ARCHIVO	Integer	Accesos totales a contexto archivo.
TOTAL ACCESOS A CONTEXTO ARCHIVO JUNIO	Integer	Accesos contexto archivo Junio.
TOTAL ACCESOS A CONTEXTO ARCHIVO JULIO	Integer	Accesos contexto archivo Julio.
TOTAL ACCESOS A CONTEXTO ARCHIVO AGOSTO	Integer	Accesos contexto archivo Agosto.
TOTAL ACCESOS A CONTEXTO ARCHIVO SEPTIEMBRE	Integer	Accesos contexto archivo Septiembre.
TOTAL ACCESOS A CONTEXTO ARCHIVO OCTUBRE	Integer	Accesos contexto archivo Octubre.
TOTAL ACCESOS A CONTEXTO FOROS	Integer	Accesos totales a contexto foros.
TOTAL ACCESOS A CONTEXTO FOROS JUNIO	Integer	Accesos contexto foros Junio.
TOTAL ACCESOS A CONTEXTO FOROS JULIO	Integer	Accesos contexto foros Julio.
TOTAL ACCESOS A CONTEXTO FOROS AGOSTO	Integer	Accesos contexto foros Agosto.
TOTAL ACCESOS A CONTEXTO FOROS SEPTIEMBRE	Integer	Accesos contexto foros Septiembre.
TOTAL ACCESOS A CONTEXTO FOROS OCTUBRE	Integer	Accesos contexto foros Octubre.

TOTAL ACCESOS A CONTEXTO CURSO	Integer	Accesos totales a contexto curso.
TOTAL ACCESOS A CONTEXTO CURSO JUNIO	Integer	Accesos contexto curso Junio.
TOTAL ACCESOS A CONTEXTO CURSO JULIO	Integer	Accesos contexto curso Julio.
TOTAL ACCESOS A CONTEXTO CURSO AGOSTO	Integer	Accesos contexto curso Agosto.
TOTAL ACCESOS A CONTEXTO CURSO SEPTIEMBRE	Integer	Accesos contexto curso Septiembre.
TOTAL ACCESOS A CONTEXTO CURSO OCTUBRE	**Integer**	**Accesos contexto curso Octubre.**

3.7. Herramientas

Para trabajar con los datos anteriormente descritos, se han realizado múltiples modificaciones para preparar el conjunto de datos para el análisis. Para ello se utilizó Microsoft Excel 2016, Python Integrated Development Environment versión 3.7.6., además KNIME versión 4.3.3.(Konstanz Information Miner) además esta última plataforma, se ha utilizado para entrenar el conjunto de datos con diferentes algoritmos de clasificación y evaluarlos para seleccionar el algoritmo de clasificación de aprendizaje automático más preciso, también se utilizó Tableau 2021.2.0 para visualizar el conjunto de datos y realizar tareas de storytelling.

3.8. ANÁLISIS DE DATOS Y PROCEDIMIENTOS

Como se ilustra en la figura 4 y en la figura 5 se han seguido tres fases principales en el marco de trabajo del KDD para responder a las preguntas de investigación. Las siguientes secciones explicarán estas fases con más detalle.

3.9. FASE DE PREPROCESAMIENTO

Como se observó en la parte metodológica; la población de datos era inmensa. Se pudo encontrar datos que contenían atributos sin valor, instancias faltantes, tipos de datos de atributos inadecuados, valores fuera de rango y otros problemas que plantean la necesidad de prepararlos primero antes de alimentarlos a la fase de análisis. Por lo tanto, los conjuntos de datos pasaron por las siguientes etapas de preparación:

3.10. LIMPIEZA DEL CONJUNTO DE DATOS

En primer lugar, se eliminaron los atributos irrelevantes de este estudio (como: código de estudiante, nombres, apellidos y DNI). Después de eso, los estudiantes con registros incompletos, como aquellos que no tenían detalles de calificaciones en la mayoría de sus cursos o aquellos que no tenían registros de cursos, fueron excluidos de la lista. Hasta esa etapa, el número restante de alumnos y sus atribuciones era de 3 406 registros, respectivamente.

3.11. CODIFICACIÓN DE CARACTERÍSTICAS

En esta etapa, los tipos de datos de todos los atributos se han cambiado a atributos numéricos por muchas razones. Primero, algunos algoritmos de aprendizaje automático, que han demostrado ser eficientes para tratar con conjuntos de datos pequeños, como la Red neuronal artificial de perceptrones múltiples (MLP-NN) (Ingrassia y Morlini, 2005) (Pasini, 2015), requiere tipos numéricos de atributos. Y el algoritmo Support Vector Machine, que también se utilizó, fue diseñado para funcionar de manera eficiente con atributos numéricos. Además, como mejor práctica al tratar con MLP-NN, en general, los atributos deben estar en forma numérica y estar normalizados para lograr los mejores resultados de clasificación. Mediante la normalización, los valores de los atributos se cambiaron y normalizaron en rangos (ya sea [0,1]) antes de introducirlos en los modelos de clasificación. Por último, dado que se utilizó

KNIME para entrenar los algoritmos de clasificación y ejecuta sus operaciones en RAM, el tratamiento con variables o cadenas de categorías requerirá más espacio, tiempo de ejecución y más sobrecarga de procesamiento (dado que los caracteres se convierten en combinaciones de bytes, especialmente cuando se trata de nombres de cursos largos) en comparación con atributos de tipo de datos numéricos. Es posible que este efecto sobre el rendimiento del procesamiento no se observe al tratar con la muestra pequeña, sin embargo, siempre es importante cumplir con las mejores prácticas para lograr resultados de análisis exitosos. En nuestro caso al procesar inicialmente alrededor de 20 millones de datos de accesos al aula virtual de la UNJBG, tuvimos problemas de hardware, lo que dificultó mucho el procesamiento.

CAPÍTULO IV

MARCO FILOSÓFICO

En esta sección desarrollaremos reflexiones de varios expertos en el campo de la Inteligencia artificial, y la información como medio de aprendizaje y predicción y sustento filosófico de nuestra investigación.

La filosofía de la inteligencia artificial es una colección de preguntas y/ reflexiones relacionadas principalmente con si la inteligencia artificial (IA) es posible o no; es decir, si es posible construir una máquina inteligente. Una preocupación secundaria es la cuestión de si es mejor pensar en los humanos y otros animales como máquinas en sí mismas(Rapaport, 1986).

La mente humana hace más que simplemente pensar en cosas. Cuando piensa, la mente se rige, al menos en parte, por las reglas de la lógica y el razonamiento inductivo, y por análisis de la fuerza de la evidencia a favor y en contra de ciertas creencias, deseos y acciones. En resumen, la mente es racional, al menos de vez en cuando. Es plausible que las computadoras sean bastante buenas para calcular la fuerza probatoria, utilizando varias lógicas. Entonces, a primera vista, las computadoras podrían ser racionales, tal vez incluso más racionales que los humanos. Pero nuevamente, un análisis más detallado revela serias dificultades (Rapaport, 1986).

Una “inteligencia artificial” (IA) es un sistema no biológico que “actúa inteligentemente” en un entorno particular. Es decir, caracterizado en términos generales, "lo que el agente hace es apropiado para sus circunstancias y su objetivo, es flexible a entornos cambiantes y objetivos cambiantes, aprende de la experiencia y toma decisiones apropiadas dadas las limitaciones de percepción y el cálculo finito"(Poole, 1998).

De nuestros conceptos mundanos y técnicos, la información es actualmente uno de los más importantes, más utilizados y menos comprendidos. Hasta ahora, los filósofos han trabajado relativamente poco sobre la información y sus conceptos afines. La filosofía, entendida como exploración y análisis conceptual, necesita centrar su atención en el nuevo mundo de la información. Esta es una manera rápida de introducir el campo que, en otros contextos, se ha definido como la filosofía y la información (PI). Como introducción, creo que es razonablemente convincente (Floridi, 2003).

La filosofía contemporánea se fundamenta en esa pérdida y la consiguiente sensación de ausencia insustituible del gran programador del juego del ser. Ya en Hume y muy claramente en Kant, dar sentido al mundo es una pesada carga, llevada enteramente sobre los hombros del yo (es indicativo, por ejemplo, que Husserl revisó las Mediaciones desde una perspectiva egocéntrica que no tenía más espacio o papel para dios). La soledad del yo en un universo silencioso se hace evidente en el idealismo alemán, que puede leerse como una serie de intentos titánicos de reconstruir una semántica absoluta apoyándose en recursos muy simplificados: la mente y su dialéctica. El gran proyecto es una naturalización del yo y del yo de la naturaleza. Los aliados naturales son una poderosa filosofía de la historia y, por supuesto, la filosofía griega, como escenario del pensamiento pre-teológico. Pero, al final, el idealismo alemán es incapaz de superar el dualismo de Kant recuperando la virginidad de Grecia con respecto al lugar integral de lo mental dentro de la naturaleza, mientras que la visión científica reemplaza gradualmente a la visión histórica. La brecha entre la mente y el ser no se puede borrar viajando en el tiempo, dice Heidegger(Floridi, 2003).

Una mayoría de filósofos indudablemente juraría la importancia del procesamiento de textos para su escritura y del correo electrónico para discusiones filosóficas, o al menos para la comunicación profesional de algún tipo con sus colegas, tales como asuntos administrativos departamentales

internos, conferencias y publicaciones. detalles, solicitudes de artículos, etc. Las listas de discusión abiertas a menudo se vuelven conversadoras, discutiendo acaloradamente, infladas o demuestran una variante de la Ley de Gresham de que el dinero malo expulsa al bien. Ha resultado difícil elegir editores adecuados para las listas de discusión editadas, o incluso encontrar a alguien dispuesto a asumir la tarea, que a menudo requiere una intervención diaria, sin remuneración. En pedagogía, la disponibilidad de PC, ahora para casi todos los estudiantes universitarios, y el acceso a Internet han sido muy elogiados. Los beneficios del acceso a Internet pueden incluir el aprendizaje a distancia o, más ampliamente aceptado como un paso adelante, el acceso a parte o todo el material de la clase en línea: textos, tareas, programa de estudios y similares. Una contribución positiva a la pedagogía también podría incluir el intercambio individual de los estudiantes con el instructor, que desplaza las horas de oficina tradicionales y notoriamente escasas. Para muchos estudiantes, el correo electrónico es menos intimidante que otras formas de intercambio. (Sin embargo, desde el punto de vista del instructor, las deficiencias verbales en el correo electrónico pueden ser más decepcionantes que en las conversaciones cara a cara, y las familiaridades y coloquialismos comunes en el correo electrónico personal de los estudiantes pueden ser seriamente desagradables para los instructores y no los preparan para encuentros posteriores más formales con superiores en el lugar de trabajo o compañeros de trabajo a través de la escritura). Yo mismo trazo el límite en el chat y las diversas formas de mensajería instantánea con los estudiantes, como sospecho que hacen la mayoría de los instructores(Dipert, 2002).

CAPÍTULO V

RESULTADOS

5.1. RESULTADO DE APLICACIÓN DEL MARCO DE TRABAJO KDD

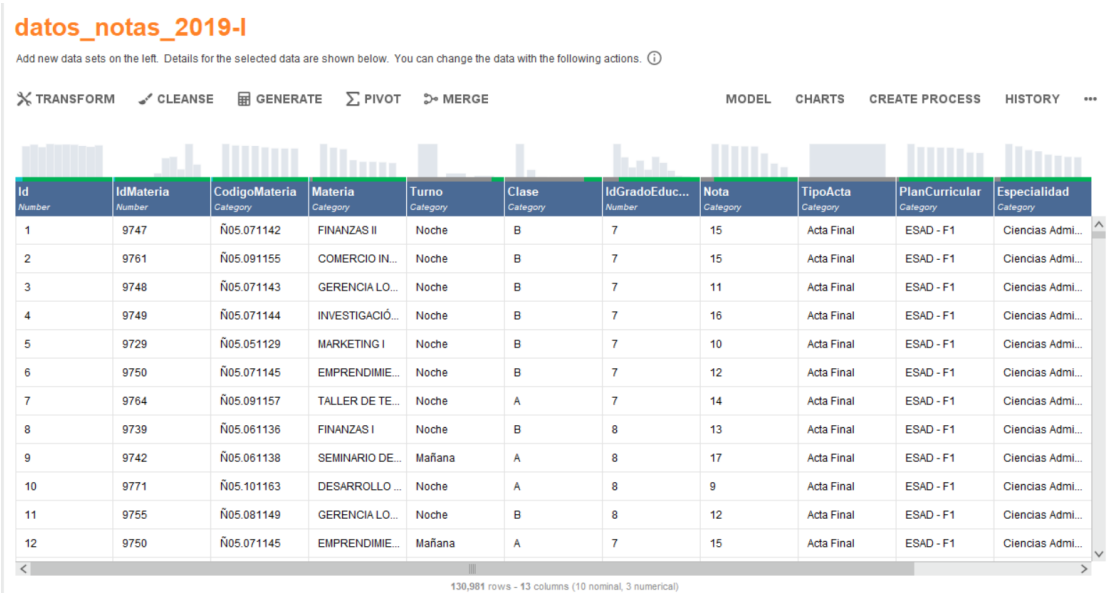
En el presente trabajo de investigación disfrutamos mucho analizando los datos, encontrando relaciones entre ellos y desarrollando un modelo basado en minería de datos para predecir el comportamiento del rendimiento académico en función a los accesos al aula virtual.

Para poder ordenar y tener una mejor idea de nuestros resultados los alinearemos al formato de KDD cuyas etapas las definimos en la figura 4.

PRIMERO: En la etapa de ***selección***, una vez identificada la data que vamos a utilizar, tuvimos un trabajo duro con la extracción de preparación de los datos del aula virtual que están en el Gestor de Base de Datos de MySQL versión 8.0.17, hicimos consultas de selección para extraer la data que necesitamos, en este caso el historial de accesos por cada estudiante. Para complementar el estudio pudimos tener acceso gracias a las facilidades que nos otorgó la Dirección Académica de Actividades y Servicios Académicos por intermedio del MSc. Nelson Mollo Condori a los registros de los estudiantes matriculados tanto en el 2019-I y el 2020-I con sus respectivas calificaciones entre 0 y 20 de los cursos en los cuales estuvieron matriculados en esos períodos.

Figura 6

Calificaciones de los estudiantes de la UNJBG en el período académico 2019-I



datos_notas_2019-I

Add new data sets on the left. Details for the selected data are shown below. You can change the data with the following actions.

TRANSFORM | CLEANSE | GENERATE | PIVOT | MERGE | MODEL | CHARTS | CREATE PROCESS | HISTORY

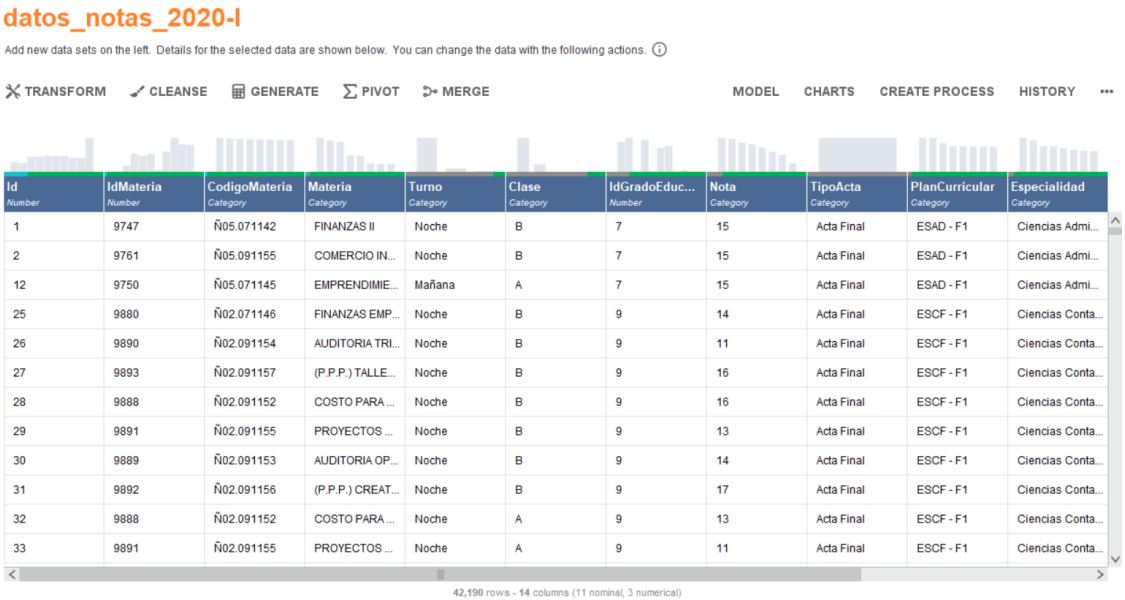
Id (Number)	IdMateria (Number)	CodigoMateria (Category)	Materia (Category)	Turno (Category)	Clase (Category)	IdGradoEduc... (Number)	Nota (Category)	TipoActa (Category)	PlanCurricular (Category)	Especialidad (Category)
1	9747	Ñ05.071142	FINANZAS II	Noche	B	7	15	Acta Final	ESAD - F1	Ciencias Admi...
2	9761	Ñ05.091155	COMERCIO IN...	Noche	B	7	15	Acta Final	ESAD - F1	Ciencias Admi...
3	9748	Ñ05.071143	GERENCIA LO...	Noche	B	7	11	Acta Final	ESAD - F1	Ciencias Admi...
4	9749	Ñ05.071144	INVESTIGACIÓ...	Noche	B	7	16	Acta Final	ESAD - F1	Ciencias Admi...
5	9729	Ñ05.051129	MARKETING I	Noche	B	7	10	Acta Final	ESAD - F1	Ciencias Admi...
6	9750	Ñ05.071145	EMPRENDIMIE...	Noche	B	7	12	Acta Final	ESAD - F1	Ciencias Admi...
7	9764	Ñ05.091157	TALLER DE TE...	Noche	A	7	14	Acta Final	ESAD - F1	Ciencias Admi...
8	9739	Ñ05.061136	FINANZAS I	Noche	B	8	13	Acta Final	ESAD - F1	Ciencias Admi...
9	9742	Ñ05.061138	SEMINARIO DE...	Mañana	A	8	17	Acta Final	ESAD - F1	Ciencias Admi...
10	9771	Ñ05.101163	DESARROLLO ...	Noche	A	8	9	Acta Final	ESAD - F1	Ciencias Admi...
11	9755	Ñ05.081149	GERENCIA LO...	Noche	B	8	12	Acta Final	ESAD - F1	Ciencias Admi...
12	9750	Ñ05.071145	EMPRENDIMIE...	Mañana	A	7	15	Acta Final	ESAD - F1	Ciencias Admi...

130,981 rows - 13 columns (10 nominal, 3 numerical)

Nota: Se obtuvo a través del Reporte de RapidMiner

Figura 7

Calificaciones de los estudiantes de la UNJBG en el período académico 2020-I.



datos_notas_2020-I

Add new data sets on the left. Details for the selected data are shown below. You can change the data with the following actions.

TRANSFORM | CLEANSE | GENERATE | PIVOT | MERGE | MODEL | CHARTS | CREATE PROCESS | HISTORY

Id (Number)	IdMateria (Number)	CodigoMateria (Category)	Materia (Category)	Turno (Category)	Clase (Category)	IdGradoEduc... (Number)	Nota (Category)	TipoActa (Category)	PlanCurricular (Category)	Especialidad (Category)
1	9747	Ñ05.071142	FINANZAS II	Noche	B	7	15	Acta Final	ESAD - F1	Ciencias Admi...
2	9761	Ñ05.091155	COMERCIO IN...	Noche	B	7	15	Acta Final	ESAD - F1	Ciencias Admi...
12	9750	Ñ05.071145	EMPRENDIMIE...	Mañana	A	7	15	Acta Final	ESAD - F1	Ciencias Admi...
25	9880	Ñ02.071146	FINANZAS EMP...	Noche	B	9	14	Acta Final	ESCF - F1	Ciencias Conta...
26	9890	Ñ02.091154	AUDITORIA TRI...	Noche	B	9	11	Acta Final	ESCF - F1	Ciencias Conta...
27	9893	Ñ02.091157	(P.P.P.) TALLE...	Noche	B	9	16	Acta Final	ESCF - F1	Ciencias Conta...
28	9888	Ñ02.091152	COSTO PARA ...	Noche	B	9	16	Acta Final	ESCF - F1	Ciencias Conta...
29	9891	Ñ02.091155	PROYECTOS ...	Noche	B	9	13	Acta Final	ESCF - F1	Ciencias Conta...
30	9889	Ñ02.091153	AUDITORIA OP...	Noche	B	9	14	Acta Final	ESCF - F1	Ciencias Conta...
31	9892	Ñ02.091156	(P.P.P.) CREAT...	Noche	B	9	17	Acta Final	ESCF - F1	Ciencias Conta...
32	9888	Ñ02.091152	COSTO PARA ...	Noche	A	9	13	Acta Final	ESCF - F1	Ciencias Conta...
33	9891	Ñ02.091155	PROYECTOS ...	Noche	A	9	11	Acta Final	ESCF - F1	Ciencias Conta...

42,190 rows - 14 columns (11 nominal, 3 numerical)

Nota: Se obtuvo a través del Reporte de RapidMiner

También se tuvo acceso a información de las fichas socioeconómicas de los estudiantes involucrados en el estudio, para esto se tuvo que instalar el AppServ e instalar en modo local MySQL, para que a través de consultas SQL, realizando muchos JOINs para lograr unir la data tal como se muestra en la figura 8.

Figura 8

Consulta SQL para seleccionar los registros de las fichas socioeconómicas y de salud de los estudiantes.

SQL Statement

```
SELECT distinct anio_ingreso,m1.modalidad,p1.item as prepa,sexo,fnacimiento
paisnac,dptonac,provnac,distnac,tipocolegio,
talla,peso,imc,ojoderecho,ojoizquierdo,r1.item as reli,m2.item as mat_viv,
t2.item as prop_viv,s1.item as depende,s2.item as sit_padres,
f3.fechaenvio as FE_Deporte,f1.fechaenvio as FE_Salud,
f2.fecha_envio as FE_Socio
FROM alumno a1
INNER JOIN fichasalud f1 ON a1.idalumno = f1.idalumno
INNER JOIN tipocolegio t1 ON a1.idtipocolegio=t1.idtipocolegio
INNER JOIN preparacion p1 ON a1.idpreparacion=p1.idpreparacion
INNER JOIN modalidad m1 ON a1.idmodalidad=m1.idmodalidad
INNER JOIN fichadeporte f3 ON a1.idalumno = f3.idalumno
INNER JOIN exfisico e2 ON f1.idfichasalud=e2.idfichasalud
INNER JOIN estadocivil e1 ON f1.estado=e1.idestadocivil
INNER JOIN fichasocio f2 ON a1.idalumno = f2.idalumno
INNER JOIN situacionpadres s2 ON f2.idsituacionpadres=s2.idsituacionpadres
INNER JOIN cargafamiliar c1 ON f2.idcargafamiliar=c1.idcargafamiliar
INNER JOIN religion r1 ON f2.idreligion=r1.idreligion
INNER JOIN vivienda v1 ON f2.idfichasocio=v1.idfichasocio
INNER JOIN material m2 ON v1.idmaterial=m2.idmaterial
INNER JOIN tenencia t2 ON v1.idtenencia=t2.idtenencia
INNER JOIN sostiene s1 ON f2.idsostiene=s1.idsostiene
```

Nota: Se obtuvo utilizando Objeto DB Query Reader de KNIME

Entre los principales datos que se obtuvieron están: el año de ingreso, la modalidad de ingreso, la preparación que tuvieron para postular, sexo, fecha de nacimiento, país de nacimiento, departamento de nacimiento, distrito, tipo colegio, talla, peso, imc, agudeza visual, religión, material de la vivienda, propiedad vivienda, dependencia y situación de los padres.

Figura 9

Datos correspondientes al aspecto psico-social de los estudiantes.

anio_ingreso Number	modalidad Category	prepa Category	sexo Category	fnacimiento Date / Time	paisnac Category	dptonac Category	provnac Category	distnac Category	tipocolegio Category	talla Number
2012	EXAMEN FASE I	Academia parti...	F	May 20, 1995	Perú	Tacna	Tacna	Ciudad Nueva	Estatal	1.570
2012	EXAMEN FASE I	Por su cuenta	M	Sep 1, 1993	Perú	Tacna	Tacna	Ciudad Nueva	Estatal	1.660
2012	EXAMEN FASE I	CEPU	M	Dec 14, 1994	Perú	Tacna	Tacna	Calana	Estatal	1.650
2012	EXAMEN FASE I	Academia parti...	M	Mar 21, 1995	Perú	Tacna	Tacna	Tacna	Estatal	1.620
2012	CEPU II	CEPU	F	Nov 30, 1994	Perú	Tacna	Tacna	Ciudad Nueva	Estatal	1.530
2012	EXAMEN FASE I	Academia parti...	F	Nov 13, 1994	Perú	Tacna	Tacna	Tacna	Estatal	1.500
2012	CEPU I	CEPU	M	Aug 22, 1994	Perú	Tacna	Tacna	Cnel. Gregorio ...	Estatal	1.790
2012	EXAMEN EXTR...	Por su cuenta	M	Feb 24, 1987	Perú	Puno	El Collao	Pilcuyo	Estatal	1.710
2012	CEPU I	CEPU	M	Apr 16, 1994	Perú	Tacna	Tacna	Ciudad Nueva	Estatal	1.720
2012	EXAMEN FASE I	CEPU	M	Feb 19, 1995	Perú	Tacna	Tacna	Cnel. Gregorio ...	Estatal	1.700
2012	EXAMEN FASE I	CEPU	M	Aug 31, 1994	Perú	Tacna	Tacna	Alto De La Alia...	Estatal	164
2012	EXAMEN EXTR...	Academia parti...	M	Jan 5, 1995	Perú	Tacna	Tacna	Tacna	Estatal	1.700

12,289 rows - 20 columns (15 nominal, 4 numerical, 1 date)

Nota: Se obtuvo utilizando RapidMiner

De igual forma seleccionamos los registros correspondientes a los accesos de los estudiantes al aula virtual tanto en el período académico 2019-I y el período académico 2020-I.

Figura 10

Datos correspondientes a los accesos de estudiantes al aula virtual en el período académico 2019-I.

Hora Date / Time	Contex... Category	Componente Category	Nombre evento Category	Descripción Category	Origen Category	Dirección IP Category
Apr 1, 2019 11:59:00 PM COT	Tarea: T...	Tarea	Se ha visualizado el estado d...	The user with id '5520' has vie...	web	201.230.37.86
Apr 1, 2019 11:59:00 PM COT	Usuario:...	Sistema	Tablero de eventos vistos	The user with id '5520' has vie...	web	201.230.37.86
Apr 1, 2019 11:58:00 PM COT	Usuario:...	Sistema	Tablero de eventos vistos	The user with id '5520' has vie...	web	201.230.37.86
Apr 1, 2019 11:57:00 PM COT	Usuario:...	Sistema	Perfil de usuario visto	The user with id '5520' viewed ...	web	201.230.37.86
Apr 1, 2019 11:57:00 PM COT	Página ...	Sistema	Curso visto	The user with id '5520' viewed ...	web	201.230.37.86
Apr 1, 2019 11:57:00 PM COT	Página ...	Sistema	Curso visto	The user with id '5520' viewed ...	web	201.230.37.86
Apr 1, 2019 11:56:00 PM COT	Sistema	Sistema	Inicio de sesión fallido	Login failed for user 'cr1390'. ...	web	190.237.100.218
Apr 1, 2019 11:54:00 PM COT	Tarea: T...	Tarea	Formulario de entrega visto.	The user with id '5520' viewed ...	web	201.230.37.86
Apr 1, 2019 11:54:00 PM COT	Sistema	Sistema	El usuario ha iniciado sesión	The user with id '5520' has log...	web	201.230.37.86
Apr 1, 2019 11:49:00 PM COT	Sistema	Sistema	Inicio de sesión fallido	Login failed for user 'cr1390'. ...	web	190.237.100.218
Apr 1, 2019 11:48:00 PM COT	Sistema	Sistema	Usuario desconectado	The user with id '5901' has log...	web	190.236.76.147
Apr 1, 2019 11:48:00 PM COT	Página ...	Sistema	Curso visto	The user with id '5901' viewed ...	web	190.236.76.147
Apr 1, 2019 11:48:00 PM COT	Curso: 1...	Sistema	Curso visto	The user with id '5901' viewed ...	web	190.236.76.147

1,148,588 rows - 7 columns (6 nominal, 1 date)

Nota*:* Se obtuvo utilizando RapidMiner.

Figura 11

Datos correspondientes a los accesos de estudiantes al aula virtual en el período académico 2020-I.

Time Date / Time	Event cont... Category	Component Category	Event name Category	Description Category	Origin Category	IP address Category
Aug 20, 0001	Course: 20-I ...	System	Course viewed	The user with id '4313' viewed th...	web	100.125.24.18
Aug 20, 0001	System	System	User login failed	Login failed for user 'qv-5011'. M...	web	100.125.24.20
Aug 20, 0001	File: SEM 2	File	Course module viewed	The user with id '2670' viewed th...	web	100.125.24.36
Aug 20, 0001	Course: 20-I ...	System	Course viewed	The user with id '7804' viewed th...	web	100.125.24.37
Aug 20, 0001	Front page	System	Course viewed	The user with id '3000' viewed th...	web	100.125.24.19
Aug 20, 0001	System	System	User logged out	The user with id '2142' has logg...	web	100.125.24.15
Aug 20, 0001	Assignment:...	Assignment	The status of the submission ha...	The user with id '3110' has view...	web	100.125.24.112
Aug 20, 0001	Assignment:...	Assignment	Course module viewed	The user with id '3110' viewed th...	web	100.125.24.112
Aug 20, 0001	Course: 20-I ...	System	Course viewed	The user with id '3044' viewed th...	web	100.125.24.12
Aug 20, 0001	Course: 20-I ...	System	Course viewed	The user with id '3110' viewed th...	web	100.125.24.115
Aug 20, 0001	Front page	System	Course viewed	The user with id '2176' viewed th...	web	100.125.24.16
Aug 20, 0001	Course: 20-I ...	System	Course viewed	The user with id '2670' viewed th...	web	100.125.24.43
Aug 20, 0001	Course: 20-I ...	System	Course viewed	The user with id '3044' viewed th...	web	100.125.24.12

19,077,265 rows - 7 columns (6 nominal, 1 date)

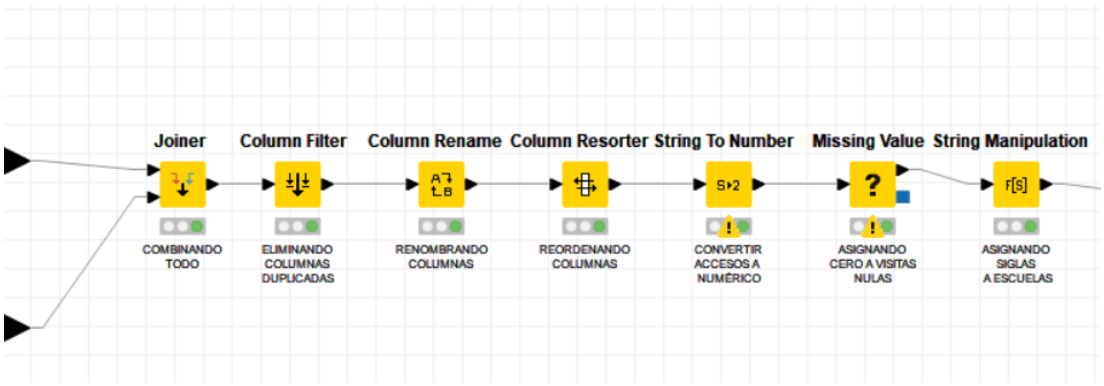
Nota: Se obtuvo utilizando RapidMiner.

Es importante resaltar que se pudo eliminar la información personal, como apellidos, nombres, código de matrícula y DNI para respetar las conductas éticas de tratamiento de datos.

SEGUNDO : En la etapa de ***pre-procesado*** de datos, se realizaron tareas de limpieza de datos de acuerdo a distintos criterios según su tipo y naturaleza como: la identificación valores desconocidos (missing values y valores empty), datos nulos, datos duplicados, combinación de datasets, filtrado por columna, renombrado de columnas, reorden de columnas, conversión de tipo de datos de cadena a número, asignación de siglas a las escuelas, para eso utilizamos técnicas estadísticas como la media, el valor más frecuente apoyándonos de la herramienta KNIME tal como se muestra en las figuras 12 y 13.

Figura 12

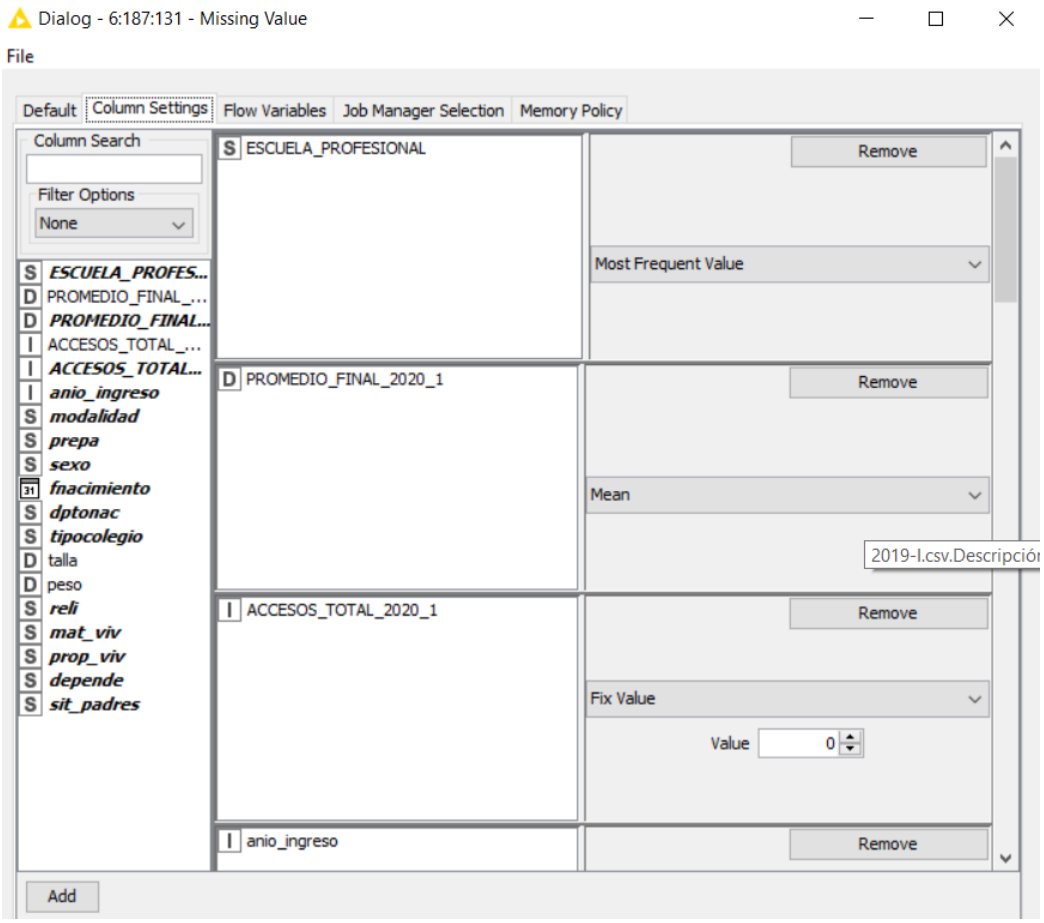
Limpieza de datos de calificaciones y accesos al aula virtual.



Nota*:* Se obtuvo utilizando KNIME

Figura 13

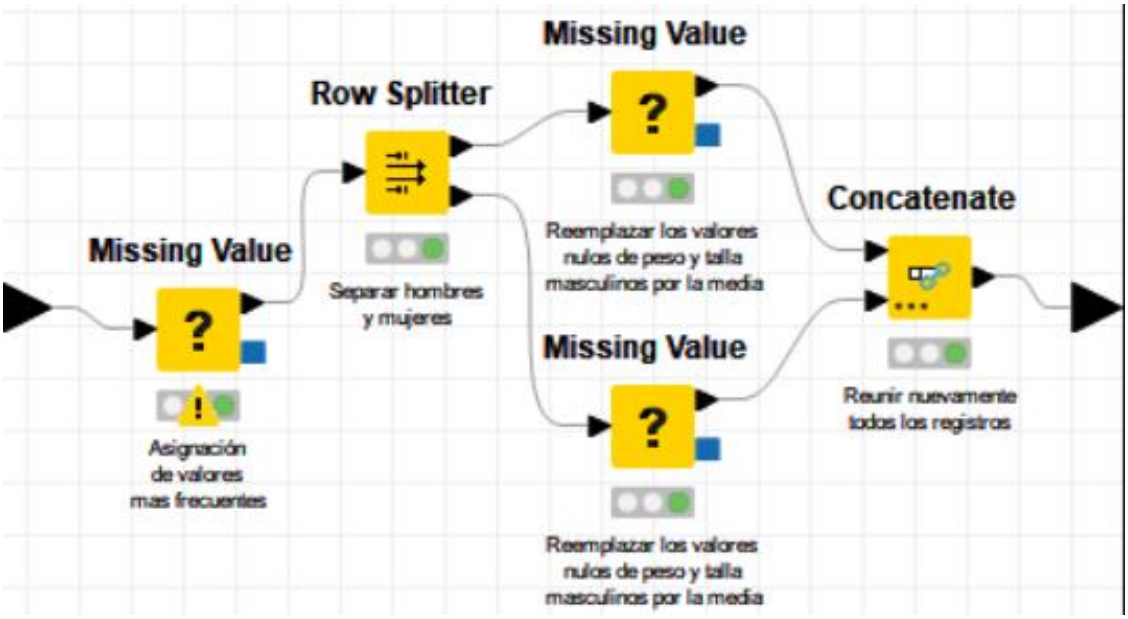
Tratamiento de valores nulos con KNIME



Nota: Se obtuvo utilizando Objeto Missing Value de KNIME

Figura 14

Tratamiento de valores nulos de peso y talla en hombres y mujeres.

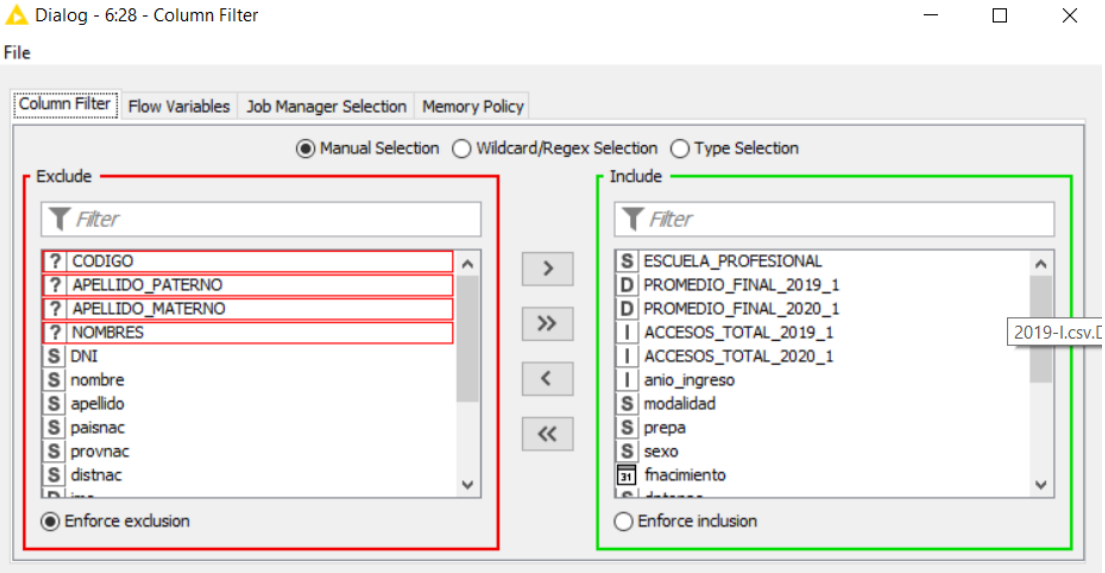


Nota: Se obtuvo utilizando Metanodo de KNIME.

TERCERO: En función al interés de estudio, que es la predicción del rendimiento académico para luego asociar con las variables de acceso al aula virtual, se realizaron tareas de transformación horizontal como: calcular el promedio de las calificaciones de los estudiantes obteniendo una sola nota como promedio del período académico 2019-I y otro promedio del período académico 2020-I, esto se llama reducción de dimensiones. También se pudo realizar una transformación vertical eliminando algunas columnas que no consideramos en la investigación, como: apellidos, nombres, dni, código universitario que aseguran la confidencialidad en el marco del manejo responsable y ético de los datos proporcionados por la Universidad ver la figura 15.

Figura 15

Filtrado por columna de los datos para asegurar la confidencialidad.



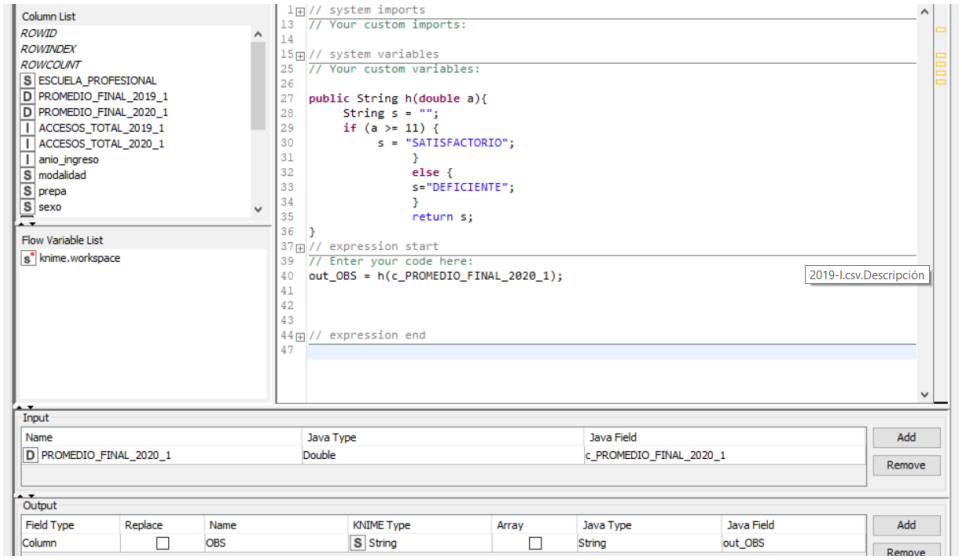
Nota: Se obtuvo utilizando Objeto Column Filter de KNIME

CUARTO: En esta etapa de minería de datos es donde se identifica las variables más determinantes en la predicción del rendimiento académico de los estudiantes, lo que permitirá asociar y relacionar las variables de acceso al aula virtual respecto a su performance académica.

Se utilizó técnicas de clasificación y clustering, para crear modelos predictivos y de esta manera estimar valores futuros para la toma de decisiones. Nuestra variable objetivo la llamamos OBS (de observación) y la preparamos transformando de dos formas los promedios finales en variables categóricas o nominales, la primera basada en la condición de que el promedio era mayor o igual que 11 su valor sería igual a ‘SATISFACTORIO’ y de lo contrario sería ‘DEFICIENTE’; la segunda forma era similar, pero ahora sería tres categorías : SATISFACTORIO’ cuando el promedio sea mayor que 11, ‘DEFICIENTE’ cuando el promedio se encuentre entre 7 y 10 y ‘MUY DEFICIENTE’ cuando el promedio se encuentre entre 0 y 6, tal como lo muestra la figura 16 y la figura 17.

Figura 16

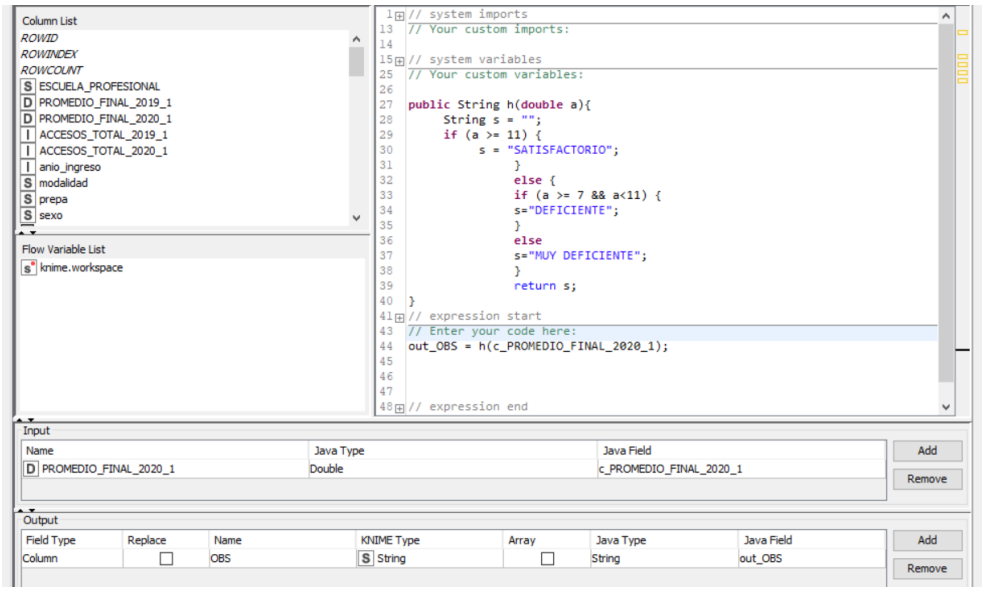
Transformación de promedios de numérico a nominal (2 niveles).



Nota: Se obtuvo utilizando Objeto Java Snipe de KNIME.

Figura 17

Transformación de promedios de numérico a nominal (3 niveles).



Nota: Se obtuvo utilizando Objeto Java Snippe de KNIME.

Una vez hechos los ajustes necesarios a los datos, se pudo resumir los resultados de la predicción de la variable objetivo (OBSERVACIÓN=SATISFACTORIO, DEFICIENTE) en la tabla 5:

Tabla 5

Resultados de clasificación con distintos algoritmos con dos valores de clase (SATISFACTORIO y DEFICIENTE)

Modelo	***Precisión general***	***Error general***	***Cohen's kappa(k)***	***Correctamente clasificados***	***Incorrectamente clasificados***
Gradient Boosted Trees	91,79 %	8,21 %	0,391	626	56
Random Forest	90,62 %	9,38 %	0,304	618	64
Decision Tree	88,94 %	11,06 %	0,281	595	74
Naive Bayes	90,91 %	9,09 %	0,348	620	62
Logistic Regression	78,59 %	21,41 %	0,225	536	146
SVM	90,47 %	9,53 %	0,115	617	65

También se hizo la predicción de la misma variable objetivo con tres valores (Satisfactorio, Deficiente y Muy deficiente), obteniendo los resultados mostrados en la tabla 6:

Tabla 6

Resultados de clasificación con tres valores de clase (SATISFACTORIO, DEFICIENTE, MUY DEFICIENTE)

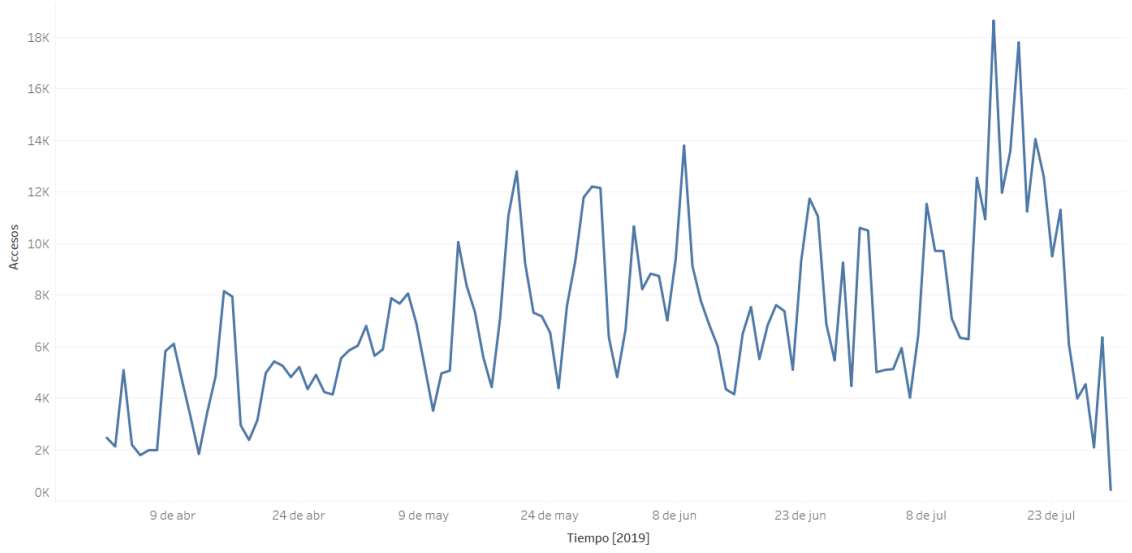
Modelo	**Precisión general**	**Error general**	**Cohen's kappa(k)**	**Correctamente clasificados**	**Incorrectamente clasificados**
Gradient Booster Trees	88,51%	11,49 %	0,324	824	107
Random Forest	89,26 %	10,74 %	0,207	831	100
Decision Tree	85,46 %	14,54 %	0,285	776	132
Naive Bayes	84,96 %	15,04 %	0,199	791	140

5.2. RESULTADOS DESCRIPTIVOS DE LAS VARIABLES

En esta sección se grafican los resultados descriptivos de las variables utilizando Tableau 2021.2.0 como herramienta para graficar los datos.

Figura 18

Accesos totales al aula virtual período académico 2019-I.



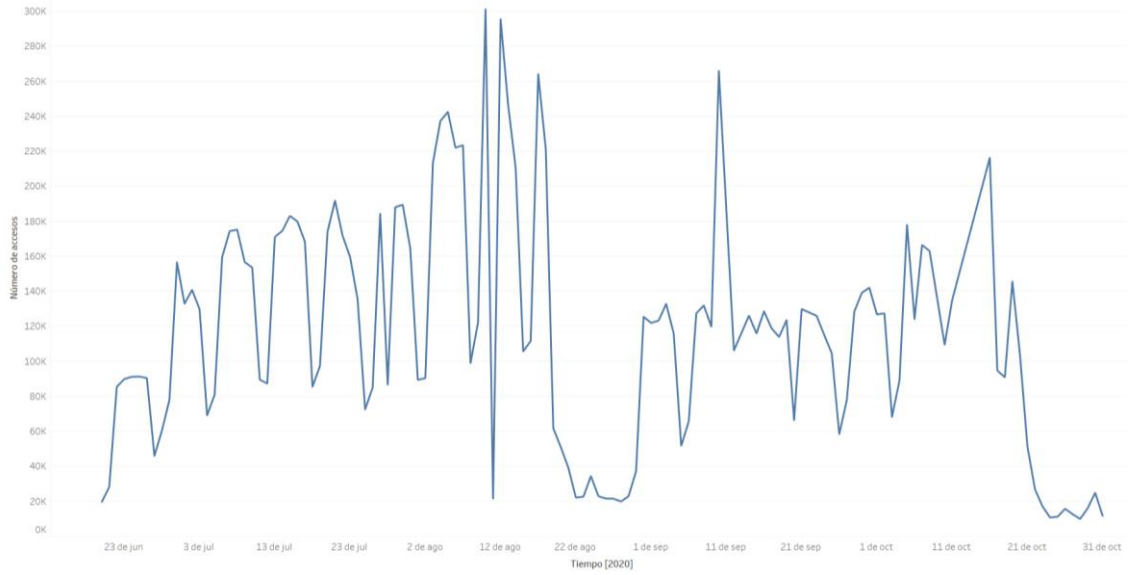
Nota: Se obtuvo utilizando Tableau

Según lo que muestra la figura 18 se puede notar que en el período del 16 al 22 de Mayo existe un ligero incremento de accesos a diferencia de los meses anteriores esto se explica porque corresponde al período de la primera evaluación del I semestre del 2019, también se nota que hay un incremento importante de accesos en el período del 12 al 24 de Julio que corresponde a la segunda evaluación y examen sustitutorio, se puede afirmar en base a esta evidencia que esas fechas debe reforzarse las medidas de seguridad que aseguren la disponibilidad de servicios del aula virtual.

En el caso de los accesos en el período 2020-I de acuerdo con la figura 19 podemos notar que:

Figura 19

Accesos totales al aula virtual período académico 2020-I.

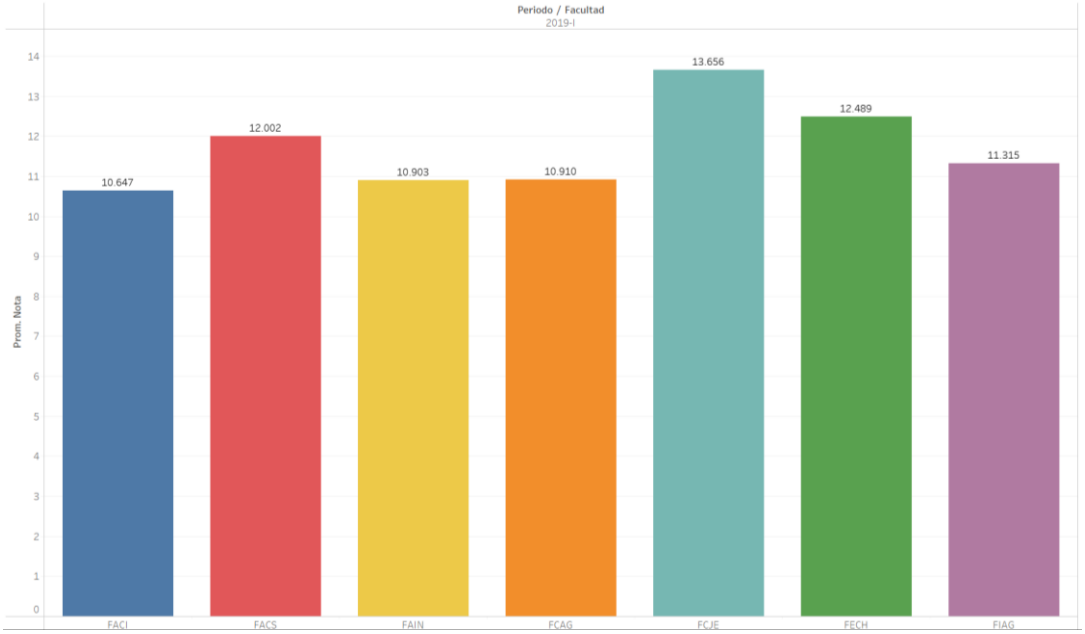


Nota*:* Se obtuvo utilizando Tableau

La figura 19 muestra claramente que hay un uso mayor de lunes a viernes disminuyendo los fines de semana, también se nota claramente un incremento importante de accesos entre el 7 al 12 de agosto, que coincide, según el Calendario académico de la UNJBG del años 2020, refrendado por Resolución de Consejo Universitario N°16862-2020-UN/JBG, con la primera evaluación, excepto el día 11 de Agosto que coincide con una caída total del sistema de aula virtual. También se nota un descenso considerable de accesos entre el 19 y el 30 de agosto, esto se asocia a la suspensión de labores académicas por alto índice de contagios entre docentes y estudiantes, de acuerdo con la resolución de Consejo Universitario N°16784-2020-UN/JBG. También se aprecia un incremento de accesos en la segunda semana de octubre que está claro que se asocia a la segunda evaluación de asignaturas. Finalmente se notó un descenso final y drástico en la semana del 20 al 31 de octubre que es semana de entrega de notas finales y matrículas del siguiente ciclo.

Figura 20

Rendimiento académico por Facultad en el período 2019-I

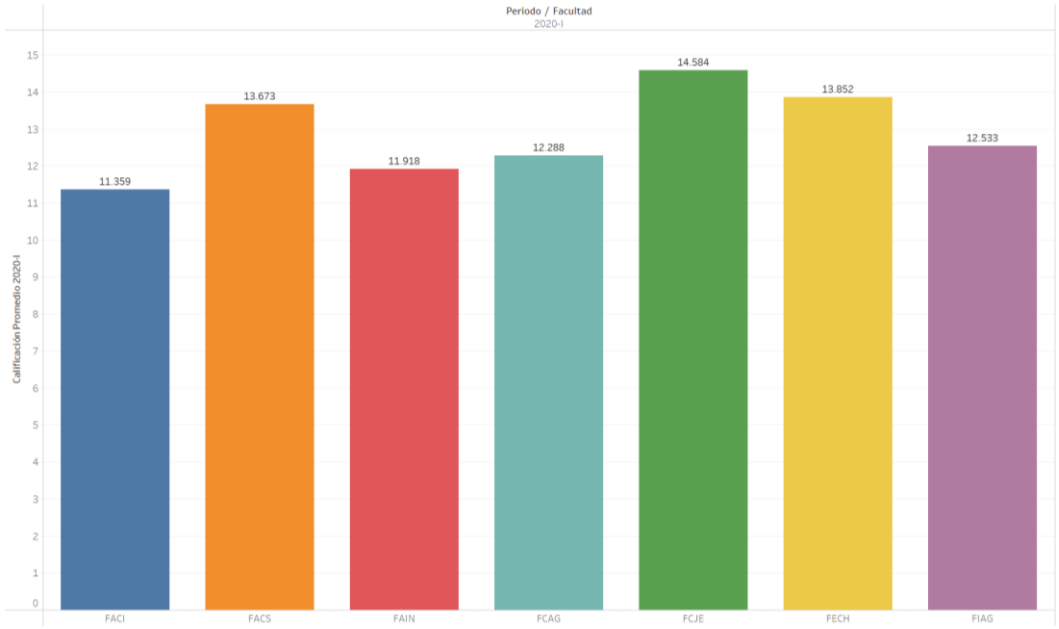


Nota: Se obtuvo utilizando Tableau

En la figura 20, se puede apreciar que en el periodo académico 2019-I, la Facultad de Ingeniería tiene un promedio ponderado de 10,903 que la posiciona como la penúltima Facultad en lo que respecta a calificaciones de los estudiantes.

Figura 21

Rendimiento académico por Facultad en el período 2020-I.

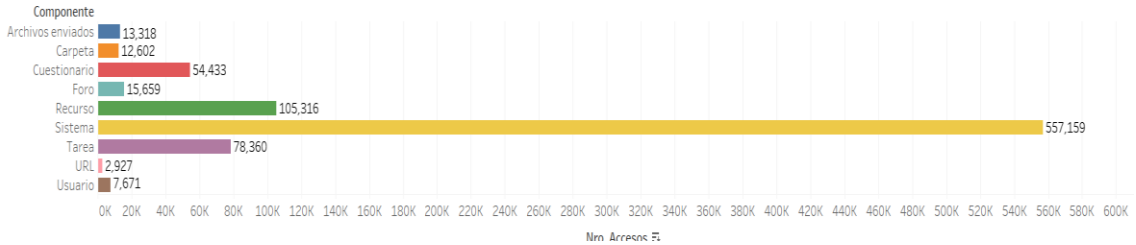


Nota: Se obtuvo utilizando Tableau.

Por otro lado, de acuerdo con la figura 21, en el período académico 2020-I, se aprecia que la Facultad de Ingeniería sube su promedio general de calificaciones alcanzando 11,918, pero se mantiene como penúltima Facultad en cuanto a rendimiento académico se refiere.

Figura 22

Accesos por componentes 2019-I



Nota: Se obtuvo utilizando Tableau.

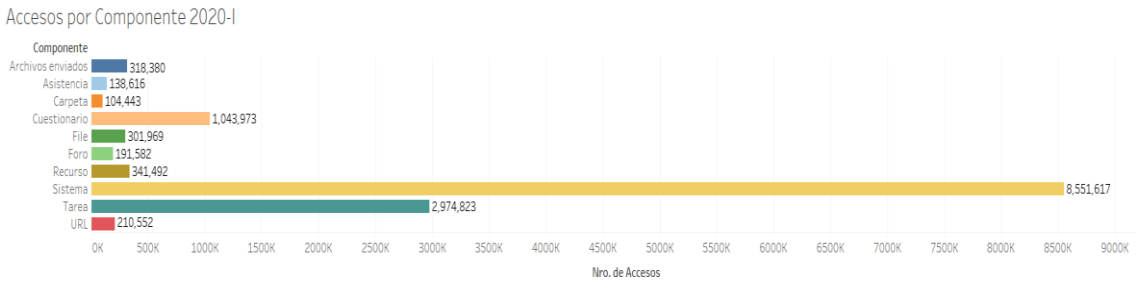
Se nota en la figura 22 con mucha facilidad que en el periodo académico 2019-I, la mayor cantidad de accesos fueron en el componente Sistema con un

total de 557 159 accesos, es decir se autenticaron para acceder a su cuenta en el aula virtual, otra cantidad de accesos importantes fueron a los recursos, es decir todos los contenidos en diversos formatos que los profesores colgaron en el aula virtual. También hay un importante número de accesos a cuestionarios con un total de 54 433 accesos, se entiende que fue para dar evaluaciones, otro grupo importante de accesos fueron a los recursos con 105 316 accesos, tareas con 78 360 acceso, foros con 15 659 accesos.

Además de acuerdo con la figura 23, se aprecia que en el caso de los accesos del 2020-I la mayor cantidad de accesos son de autenticación en el aula a través del sistema con un total de 8 551 617 accesos, otro componente importante fue el acceso a las tareas con 2 974 823 accesos, también se tuvo un total de 1 043 973 accesos a los cuestionarios entre los accesos más significativos.

Figura 23

Accesos por componentes 2020-I

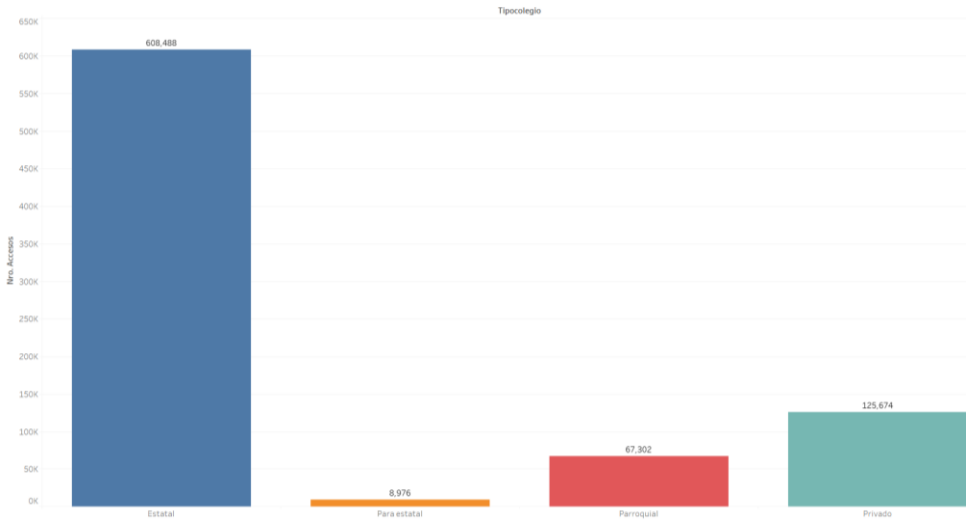


Nota: Se obtuvo utilizando Tableau

En la figura 24, se puede apreciar que en el periodo académico 2019-I, la mayor cantidad de accesos corresponde a los estudiantes de colegios públicos con un total de 608 488 accesos.

Figura 24

Accesos por tipo colegio 2019-I

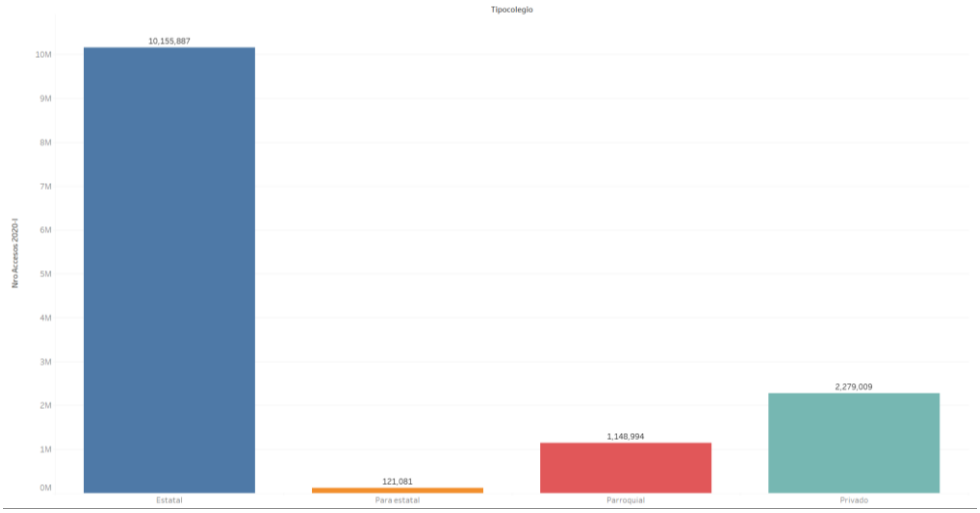


Nota: Se obtuvo utilizando Tableau

De la misma forma en la figura 25 se puede apreciar que el comportamiento de accesos por tipo colegio para el período académico 2020-I muestra una gran diferencia de estudiantes provenientes de colegios estatales.

Figura 25

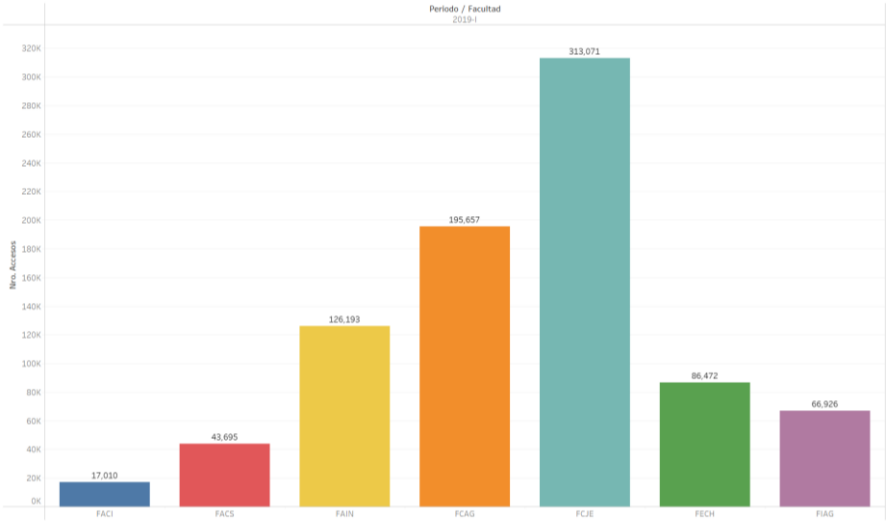
Accesos por tipo colegio 2020-I



Nota: Se obtuvo utilizando Tableau

Figura 26

Accesos por facultad en el periodo académico 2019-I

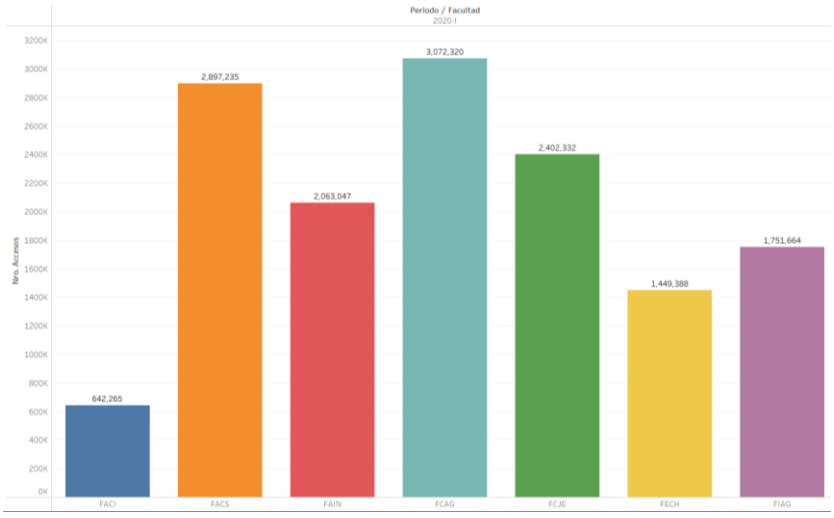


Nota: Elaboración propia utilizando Tableau

De acuerdo con la figura 26, se nota con mucha claridad que la Facultad de Ciencias Jurídicas y Empresariales alcanzó 313 071 accesos lo que representa el mayor índice de accesos al aula virtual del periodo académico 2019-I, seguida de la Facultad de Ciencias Agropecuarias con 195 657 accesos y luego la Facultad de Ingeniería con 126 193 accesos.

Figura 27

Accesos por facultad en el periodo académico 2020-I

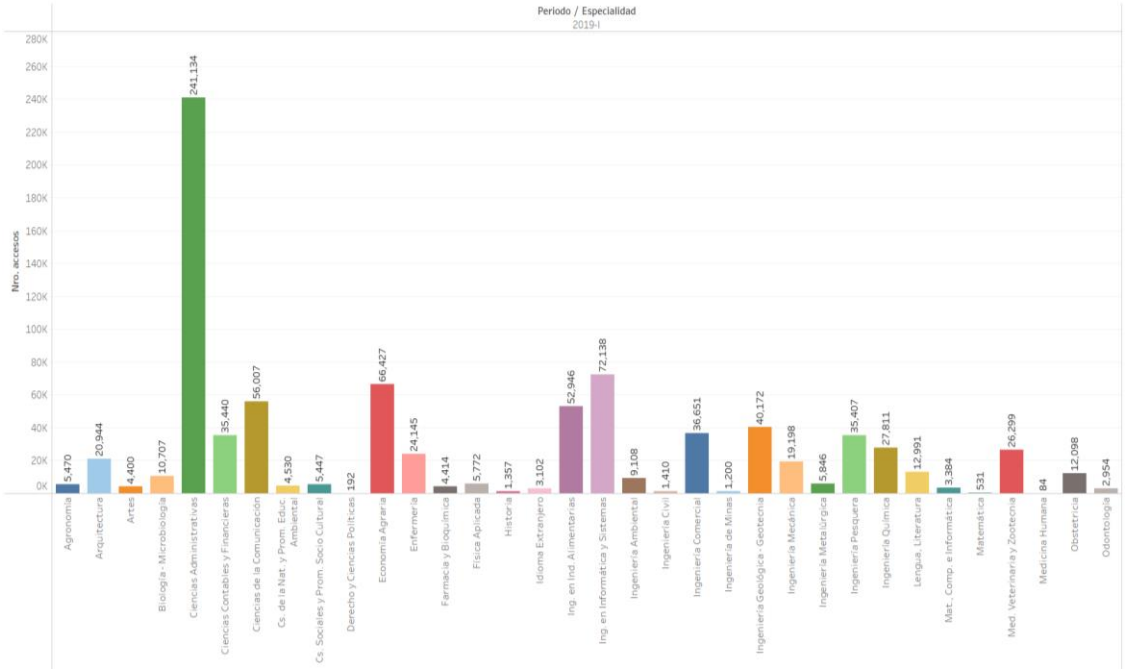


Nota: Se obtuvo utilizando Tableau

De acuerdo con la figura 27, en el período académico 2020-I se pudo notar que la mayor cantidad de accesos fueron hechos por estudiantes de la Facultad de Ciencias Agropecuarias con un total de 3 072 320 accesos, seguido por la Facultad de Ciencias de la Salud con 2 897 235 accesos, luego la Facultad de Ciencias Jurídicas y Empresariales con 2 402 332 accesos entre las principales.

Figura 28

Accesos por Escuela período académico 2019-I



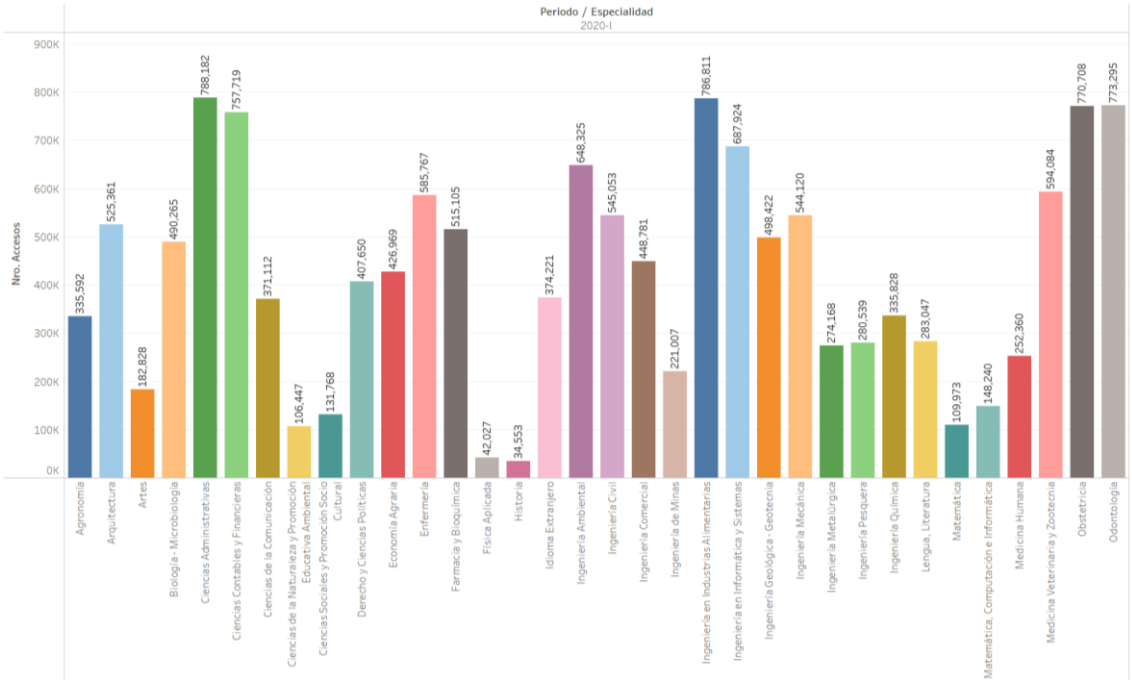
Nota: Se obtuvo utilizando Tableau

En la figura 28 se pudo notar que existe una diferencia muy marcada entre los accesos de la Escuela Profesional de Ciencias Administrativas (241 134) y las demás. La Escuela Profesional que se encuentra en el segundo lugar en términos de accesos es la Escuela Profesional de Ingeniería en Informática y Sistemas tiene 72 138 accesos.

En el período académico 2020-I, se presenta un escenario distinto ya que las diferencias son menores, es así que la Escuela Profesional con más accesos fue la Escuela Profesional de Ciencias Administrativas con un total de 788 182 accesos, pero esta vez seguida de muy cerca por la Escuela Profesional de Ingeniería en Informática y Sistemas con 786 811 accesos de acuerdo con lo que muestra la Figura 29.

Figura 29

Accesos por Escuela Profesional 2020-I

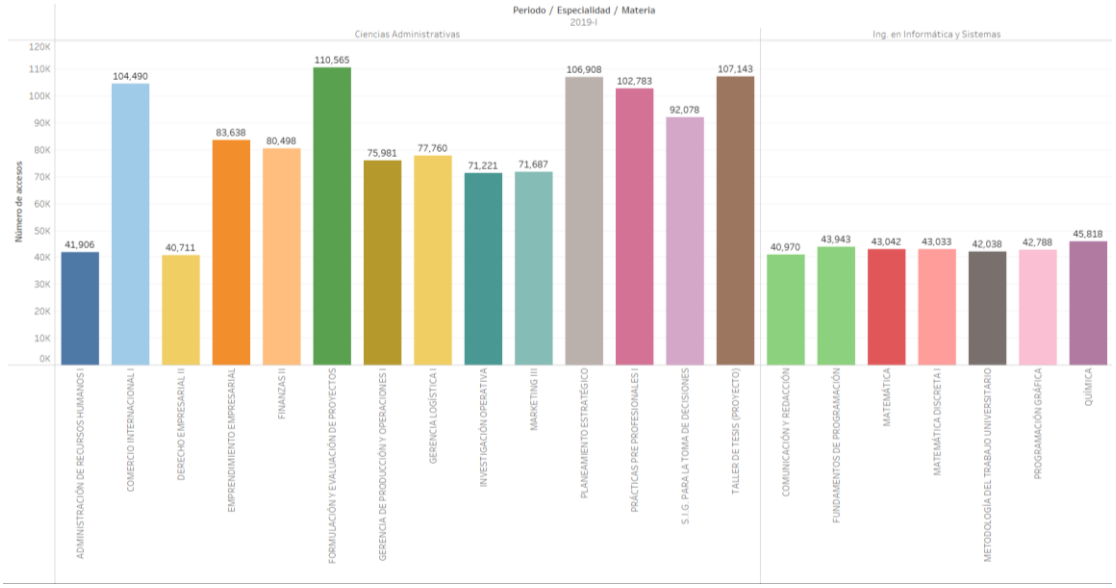


Nota: Se obtuvo utilizando Tableau

En lo que respecta a las asignaturas que más accesos tuvieron en el período académico 2019-I se tuvo que aplicar un filtro de 40 000 accesos como mínimo debido a la gran cantidad de asignaturas que se imparten en la UNJBG. En tal sentido se observó que la asignatura de Formulación y evaluación de Proyectos con 110 565 accesos y Taller de Tesis con 107 143 accesos de la Escuela Profesional de Ciencias Administrativas son las asignaturas que más accesos tuvieron en el aula virtual en el periodo académico 2019-I de acuerdo con la figura 30.

Figura 30

Asignaturas por Escuela con más accesos 2019-I.



Nota: Se obtuvo utilizando Tableau.

En el caso del período académico 2020-I, la asignatura con más accesos fue Química general de la Facultad de Ciencias de Salud con 707 447 accesos, seguida de la asignatura de Química de la Facultad de Ciencias Agropecuarias con 647 413 accesos y la asignatura de Biología general de la Facultad de Ciencias de la Salud con 644 883 accesos de acuerdo con lo mostrado en la figura 31.

Figura 31

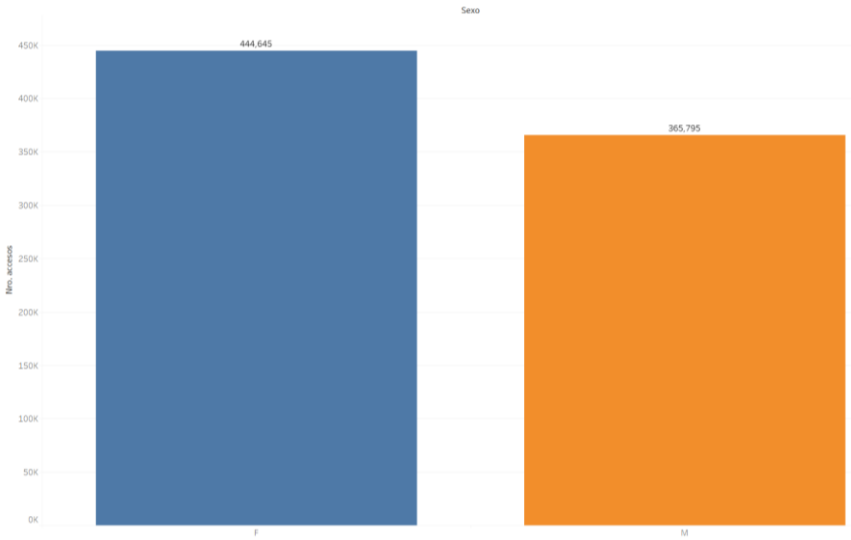
Asignaturas con más accesos en el período académico 2020-I

Nota: Se obtuvo utilizando Tableau

En lo que se refiere a género, hubo más accesos de estudiantes de sexo femenino en el período académico 2019-I con un total de 444 645 accesos frente a 365 795 accesos masculinos, tal como se muestra en la figura 32.

Figura 32

Accesos por sexo en el período académico 2019-I

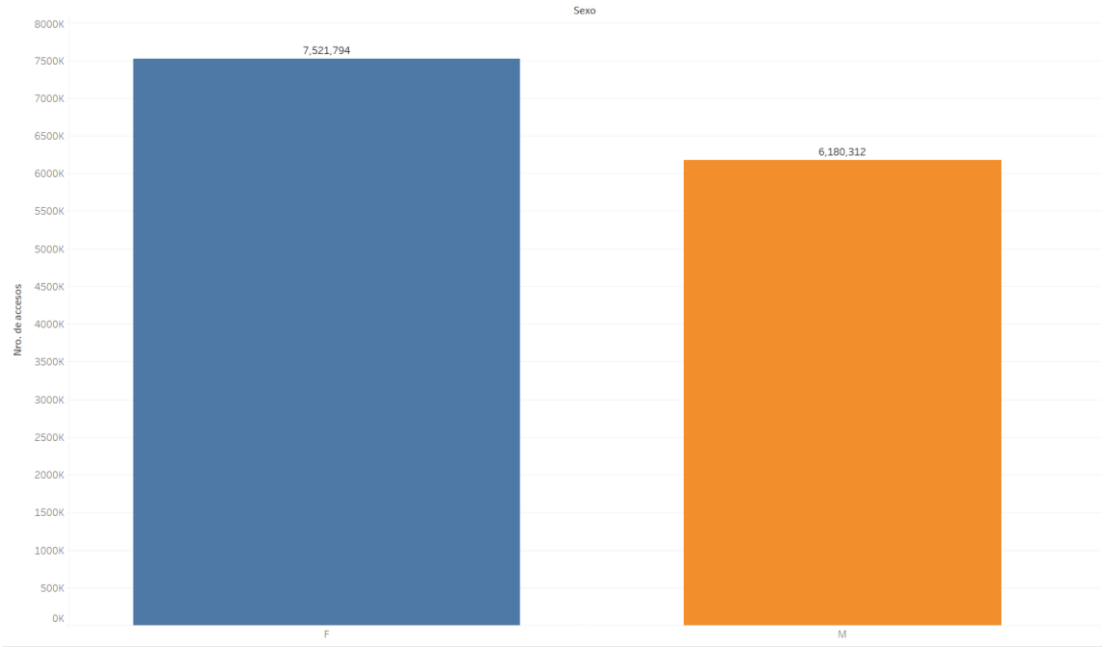


Nota: Se obtuvo utilizando Tableau

Mientras que en el período académico 2020-I, se puede apreciar en la figura 33, que se sigue la misma tendencia, pero con muchos más accesos, las mujeres accedieron en total 7 521 794 veces y los hombres 6 180 312 veces al aula virtual.

Figura 33

Accesos por Sexo período académico 2020-I

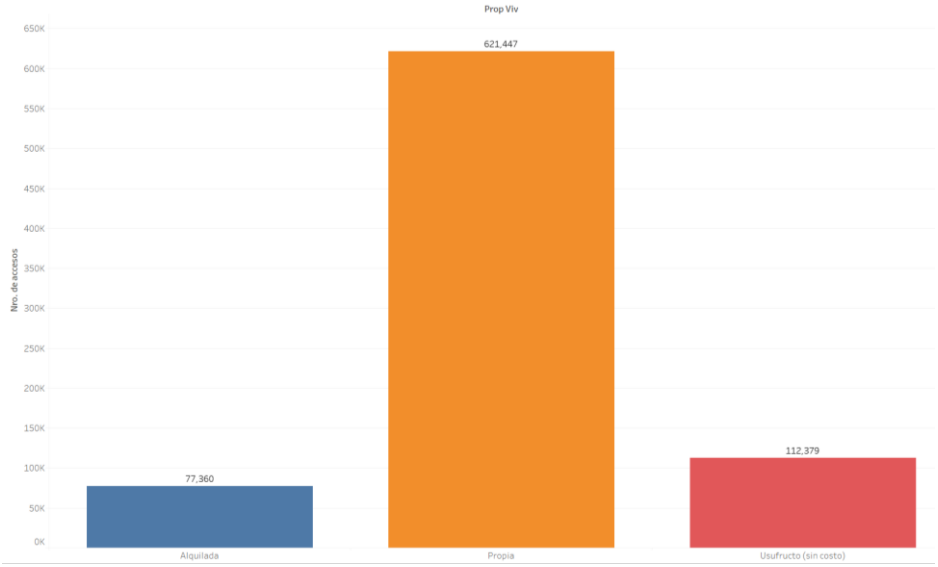


Nota: Se obtuvo utilizando Tableau

En el caso de los accesos según tipo de vivienda, en el período académico 2019-I se aprecia que hubo 621 447 accesos de estudiantes que cuentan con casa propia, marcando una gran diferencia con aquellos estudiantes que viven en casas alquiladas que solo accedieron un total de 77 360, de acuerdo con lo que muestra la figura 34.

Figura 34

Número de accesos por tipo de vivienda en el período académico 2019-I

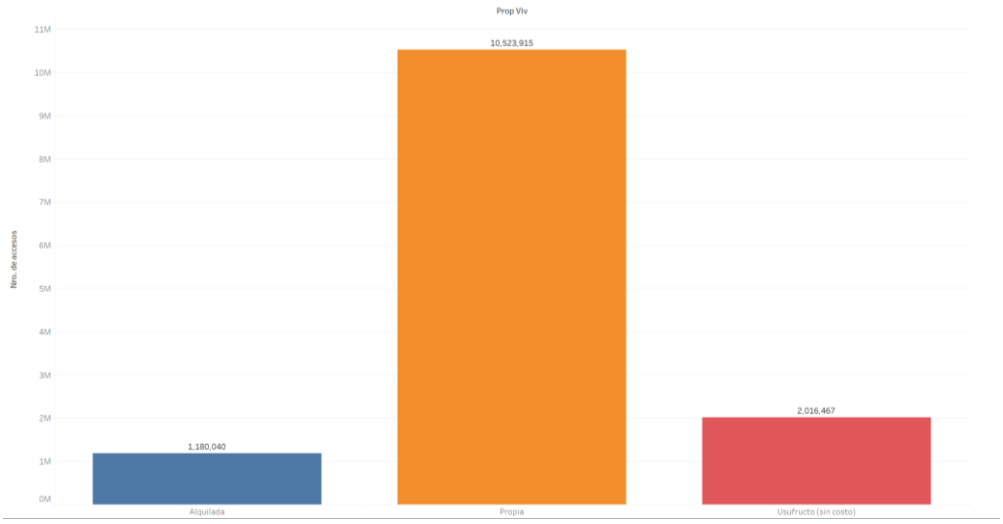


Nota: Elaborado con Tableau

En el caso del período académico 2020-I, se pudo apreciar de acuerdo a la figura 35, que el mayor número de accesos fue hecho por estudiantes que tienen casa propia con un total de 10 523 915 accesos.

Figura 35

Accesos al aula virtual de acuerdo con la propiedad de vivienda 2020-I

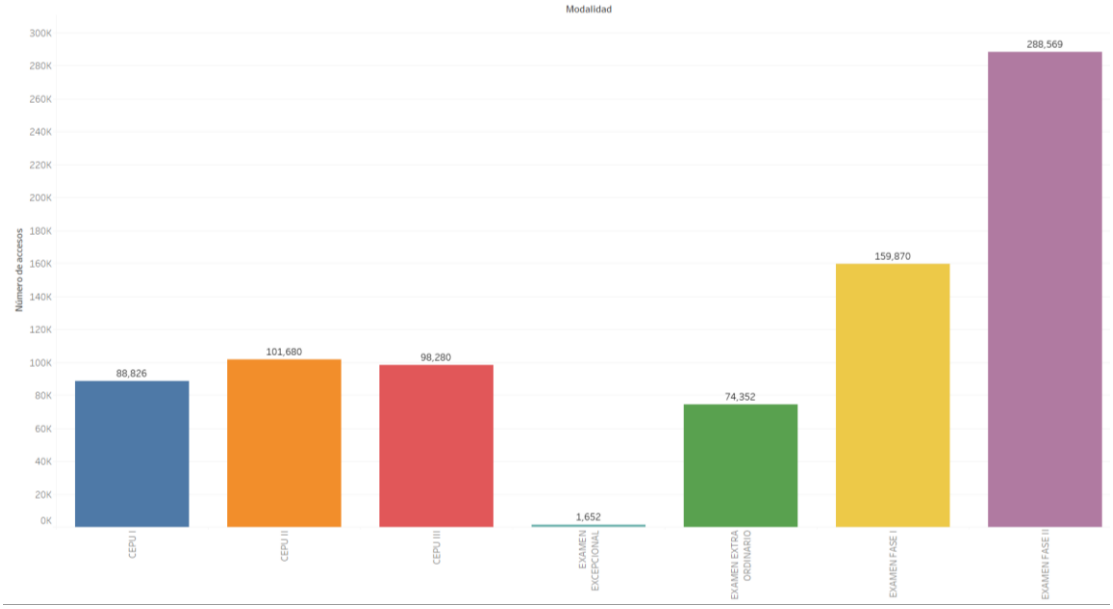


Nota: Se obtuvo utilizando Tableau

Por otro lado, se pudo apreciar de acuerdo con la figura 36, en acceso al aula virtual según su tipo de ingreso a la Universidad en el período 2019-I donde se identifica que el mayor número de accesos al aula virtual proviene de estudiantes que ingresaron bajo la modalidad de Examen Fase II.

Figura 36

Accesos al Aula virtual 2019-I por tipo ingreso a la UNJBG.

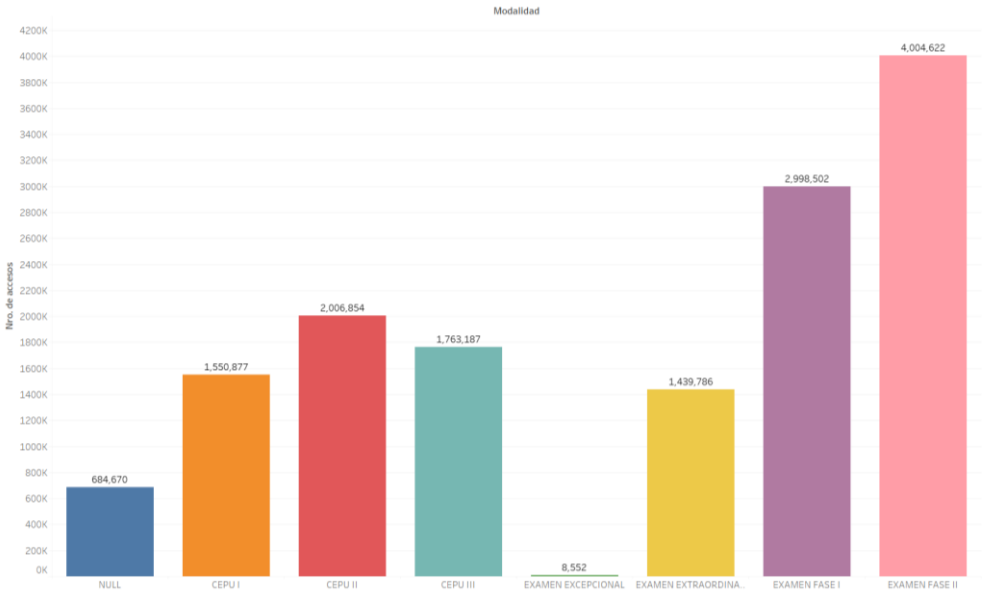


Nota: Se obtuvo utilizando Tableau

En el mismo sentido se pudo notar que para el período académico 2020-I, el mayor número de accesos al aula virtual fue hecho por estudiantes que ingresaron a la universidad bajo la modalidad de examen de admisión Fase II también, con un total de 4 004 622 accesos tal como lo demuestra la figura 37.

Figura 37

Accesos al aula virtual en el período académico 2020-I según modalidad de ingreso a la UNJBG.

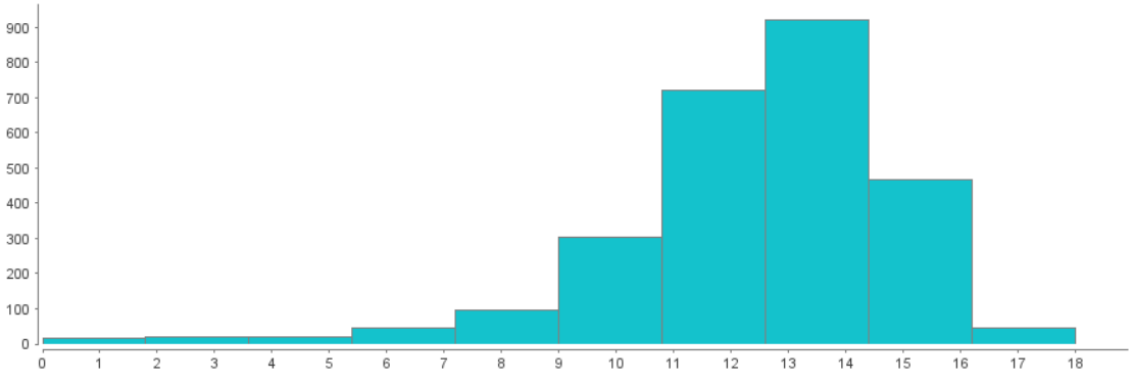


Nota: Se obtuvo utilizando por Tableau

En el caso de la distribución de las calificaciones del periodo académico 2019-I, se pudo notar que la mayor concentración de notas está entre 7 y 16 aproximadamente, de acuerdo con lo que describe la figura 38.

Figura 38

Distribución de las calificaciones en el período académico 2019-I.

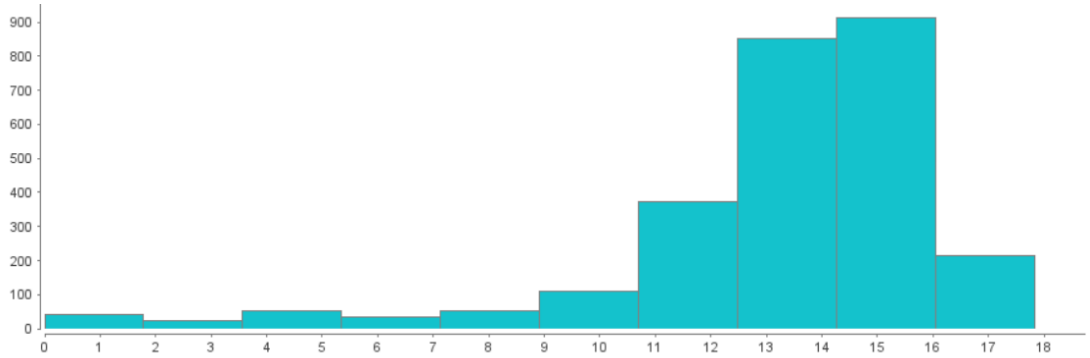


Nota: Se obtuvo utilizando RapidMiner

En el mismo sentido en el caso de las calificaciones del periodo académico 2020-I en tiempos de pandemia los valores de las calificaciones promedio de los estudiantes se encuentran entre 9 y 18 aproximadamente de acuerdo con la figura 39.

Figura 39

Distribución de las calificaciones en el período académico 2020-I.



Nota: Se obtuvo utilizando RapidMiner

5.3. RESULTADOS DE LOS MODELOS APLICADOS

En este apartado de la tesis se describe detalladamente cada uno de los resultados de la aplicación de los algoritmos de clasificación del rendimiento académico en el periodo académico 2020-I que es materia de nuestra investigación todo esto en búsqueda del modelo más preciso y apropiado para las características propias de nuestros estudiantes, y las condiciones de no presencialidad en la Universidad Nacional Jorge Basadre Grohmann de Tacna.

Figura 40

Descripción de las entradas de los modelos

Status	Quality	Name	Correlation ↓	ID-ness	Stability	Missing	Text-ness
●		PROMEDIO_FINAL2019-I	24.32%	?	2.40%	0.00%	0.00%
●		escuela	3.64%	1.13%	9.31%	0.00%	2.15%
●		ACCESOS_TOTALES 2020-I	3.27%	67.72%	0.23%	0.00%	0.00%
●		sexo	1.08%	0.08%	52.44%	0.00%	0.47%
●		ACCESOS_TOTALES 2019-I	0.49%	24.62%	2.52%	0.00%	0.00%
●		deporte	0.38%	0.64%	35.77%	0.00%	6.21%
●		fnacimiento	0.30%	?	0.26%	0.00%	0.00%
●		anio_ingreso	0.29%	0.34%	26.28%	0.00%	0.00%
●		modalidad	0.29%	0.26%	36.56%	0.00%	38.76%
●		peso	0.17%	?	6.19%	0.00%	0.00%
●		mat_viv	0.05%	0.15%	47.11%	0.00%	44.99%
●		facultad	0.04%	0.26%	26.09%	0.00%	1.87%
●		prepa	0.03%	0.15%	40.99%	0.00%	24.91%
●		estado_civil	0.02%	0.08%	99.77%	0.00%	2.69%
●		sit_padres	0.02%	0.26%	39.45%	0.00%	6.58%
●		reli	0.01%	0.38%	72.67%	0.00%	3.96%
●		depende	0.01%	0.19%	44.82%	0.00%	19.74%
●		tipocolegio	0.00%	0.15%	74.59%	0.00%	3.64%
●		prop_viv	0.00%	0.11%	72.64%	0.00%	9.05%
●		talla	0.00%	?	6.42%	0.00%	0.00%

Nota: Se obtuvo utilizando RapidMiner

En la figura 40 se puede notar que las variables de entrada tienen algunas características que se ha podido analizar como la denotada por C (Correlation)

que indica que como era de suponerse las calificaciones del 2019-I, La Escuela de procedencia y el número de accesos son las que más alto nivel de correlación tienen con respecto al valor del rendimiento académico del periodo 2020-I.

Esto sirvió como punto de partida para seleccionar las variables que tienen mayor relevancia en el modelo o sea que tienen mayor impacto en el rendimiento académico del periodo académico 2020-I, que es la variable de interés y la que pretendemos clasificar.

Finalmente se sometió los datos preprocesados a los distintos algoritmos de Machine learning.

Antes de mostrar los resultados, es oportuno recordar que, para poder aplicar los algoritmos, se particionó la data en dos grupos, el primero constituido por el 80 %; es decir alrededor de 2774 registros que son utilizados para el entrenamiento del clasificador y el segundo compuesto por el 20 %; es decir alrededor de 682 registros para testeo, aproximadamente.

Como se puede apreciar en la figura 41, se visualiza la aplicación del algoritmo Random Forest sobre la data preprocesada, si se observa la matriz de confusión este algoritmo tuvo una precisión de 25,37 clasificando rendimientos “DEFICIENTES” y 97,72 clasificando rendimientos “SATISFACTORIOS” de un total de 682 registros testeados clasificó 618 correctamente y 64 de manera incorrecta logrando una precisión general de 90,62 %

Figura 41

Matriz de confusión de Random Forest

Scorer View

Confusion Matrix

	DEFICIENTE (Predicted)	SATISFACTORIO (Predicted)	
DEFICIENTE (Actual)	17	50	25.37%
SATISFACTORIO (Actual)	14	601	97.72%
	54.84%	92.32%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
90.62%	9.38%	0.304	618	64

Nota: Se obtuvo utilizando KNIME

De la misma forma se puede apreciar en la figura 42, que aplicando el algoritmo de árboles de decisión donde tuvo una precisión de 29,23 % clasificando rendimientos académicos “DEFICIENTES” y 95,36 % de precisión clasificando rendimientos académicos “SATISFACTORIOS”, de un total de 669 registros de testeo clasificó correctamente 595 y 74 de manera incorrecta, logrando una precisión general de 88,94%.

Figura 42

Matriz de confusión de Árboles de decisión

Scorer View

Confusion Matrix

	DEFICIENTE (Predicted)	SATISFACTORIO (Predicted)	
DEFICIENTE (Actual)	19	46	29.23%
SATISFACTORIO (Actual)	28	576	95.36%
	40.43%	92.60%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
88.94%	11.06%	0.281	595	74

Nota: Se obtuvo utilizando KNIME

Así mismo en la figura 43, se puede apreciar la matriz de confusión de la aplicación del algoritmo de Gradient Boosted Trees donde muestra un 31,34 % de precisión en la clasificación de rendimientos académicos “DEFICIENTES” y

un 98,37 % en la clasificación de rendimientos académicos “SATISFACTORIOS”, de un total de 682 registros testeados, clasificó correctamente 626 e incorrectamente 56, logrando alcanzar una precisión general de 91,79%.

Figura 43

Matriz de confusión de Gradient Boosted Trees

Scorer View

Confusion Matrix

	DEFICIENTE (Predicted)	SATISFACTORIO (Predicted)	
DEFICIENTE (Actual)	21	46	31.34%
SATISFACTORIO (Actual)	10	605	98.37%
	67.74%	92.93%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
91.79%	8.21%	0.391	626	56

Nota: Se obtuvo utilizando KNIME

Del mismo modo se puede apreciar en la figura 44 la matriz de confusión como resultado de la aplicación del algoritmo de Naive Bayes con un 29,85 de precisión las calificaciones “DEFICIENTES” y con un 97,56 % las calificaciones “SATISFACTORIAS”. Además de los 682 registros se pudo clasificar correctamente 620 de manera incorrecta 62, logrando así una precisión final de 90,91 %.

Figura 44

Matriz de confusión de Naive Bayes

Scorer View

Confusion Matrix

	DEFICIENTE (Predicted)	SATISFACTORIO (Predicted)	
DEFICIENTE (Actual)	20	47	29.85%
SATISFACTORIO (Actual)	15	600	97.56%
	57.14%	92.74%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
90.91%	9.09%	0.348	620	62

Nota: Se obtuvo utilizando KNIME

Sin embargo, en la figura 45, que muestra la matriz de confusión de la aplicación del algoritmo de Regresión logística, se observa que clasificando las calificaciones “DEFICIENTES” tiene una precisión de 53,73 % y 81,30 % de precisión clasificando las calificaciones “SATISFACTORIAS”. Además de los 682 registros utilizados en el testeo el modelo pudo clasificar correctamente 536 e incorrectamente 146, logrando la más baja precisión de todos los modelos aplicados apenas con un 78,59 %.

Figura 45

Matriz de confusión de Regresión logística

Scorer View

Confusion Matrix

	DEFICIENTE (Predicted)	SATISFACTORIO (Predicted)	
DEFICIENTE (Actual)	36	31	53.73%
SATISFACTORIO (Actual)	115	500	81.30%
	23.84%	94.16%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
78.59%	21.41%	0.225	536	146

Nota: Se obtuvo utilizando KNIME

También en la figura 46, se puede observar la matriz de confusión del resultado de la aplicación del algoritmo de Support Vector Machine y con una precisión de 7,46 % en la clasificación automática de las calificaciones “DEFICIENTES” y un 99,51 % de precisión en la clasificación automática de calificaciones “SATISFACTORIAS”. También es importante resaltar que, de los 682 registros utilizados para el testeo del modelo de predicción, se pudo clasificar correctamente 617 y de manera equivocada 65, obteniendo una precisión general de 90,47 %.

Figura 46

Matriz de confusión de SVM.

Scorer View

Confusion Matrix

	DEFICIENTE (Predicted)	SATISFACTORIO (Predicted)	
DEFICIENTE (Actual)	5	62	7.46%
SATISFACTORIO (Actual)	3	612	99.51%
	62.50%	90.80%	

Overall Statistics

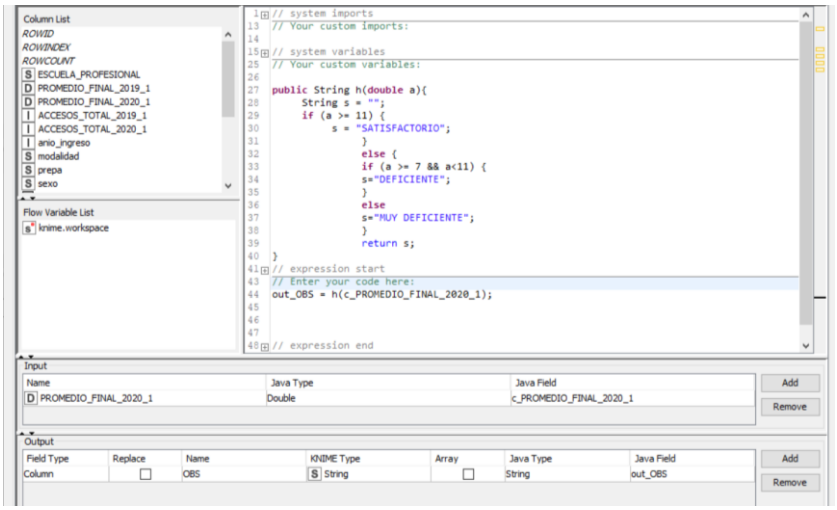
Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
90.47%	9.53%	0.115	617	65

Nota: Elaborado con KNIME

También se pudo experimentar con tres niveles de calificaciones: SATISFACTORIO (nota final mayor o igual que 11), DEFICIENTE (promedio final mayor o igual que 7 y menor que 11) y MUY DEFICIENTE (promedio final menor que 7) tal como lo muestra la figura 47.

Figura 47

Determinación de valores de observación según promedio final



Nota: Se obtuvo utilizando Objeto Java Snippet de KNIME

Esto se hizo para poder determinar si existe alguna variación en las precisiones de los algoritmos, al ampliar los valores de clases a 3 y obtuvimos los siguientes resultados:

Por ejemplo, para el caso del algoritmo de árboles de decisión de acuerdo con la figura 48, se tuvo un 26,15 % de aciertos clasificando calificaciones DEFICIENTES, un 35,14 % de aciertos en la clasificación de calificaciones MUY DEFICIENTES y un 92,56 % de aciertos clasificando calificaciones SATISFACTORIAS. Por otro lado, de un total de 908 registros utilizados para el testeo del modelo generado, se pudo clasificar correctamente 776 registros y 132 de manera errónea, logrando finalmente una precisión general de 85,46 %.

Figura 48

Matriz de confusión del algoritmo árboles de decisión con 3 valores de clase.

Scorer View

Confusion Matrix

	DEFICIENTE (Predicted)	MUY DEFICIENTE (Pr...	SATISFACTORIO (Pre...	
DEFICIENTE (Actual)	17	3	45	26.15%
MUY DEFICIENTE (Ac...	7	13	17	35.14%
SATISFACTORIO (Act...	44	16	746	92.56%
	25.00%	40.63%	92.33%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
85.46%	14.54%	0.285	776	132

Nota: Se obtuvo utilizando KNIME

Así mismo para el caso del algoritmo de Random Forest de acuerdo con la figura 49, se pudo verificar que tuvo un acierto de 4,48 % clasificando calificaciones “DEFICIENTES”, un acierto de 30 % clasificando calificaciones “MUY DEFICIENTES” y un 99,03 % clasificando calificaciones “SATISFACTORIAS”. Además de un total de 931 registros utilizados para el testeo del modelo, este pudo clasificar correctamente 831 y erróneamente 100, obteniendo una precisión total de 89,26 %.

Figura 49

Matriz de confusión del algoritmo Random Forest con 3 valores de clase.

Scorer View

Confusion Matrix

	DEFICIENTE (Predicted)	MUY DEFICIENTE (Pr...	SATISFACTORIO (Pre...	
DEFICIENTE (Actual)	3	0	64	4.48%
MUY DEFICIENTE (Ac...	0	12	28	30.00%
SATISFACTORIO (Act...	4	4	816	99.03%
	42.86%	75.00%	89.87%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
89.26%	10.74%	0.207	831	100

Nota: Se obtuvo utilizando KNIME

Por otro lado, para el algoritmo Gradient Boosted Trees, según se aprecia en la figura 50, se puede verificar que tuvo un acierto de 13,43 % clasificando calificaciones “DEFICIENTES”, 52,50 % clasificando calificaciones “MUY DEFICIENTES” y 96,36 % clasificando calificaciones “SATISFACTORIAS”. Además de un total de 931 registros utilizados para el testeo del modelo obtenido, se pudo clasificar 824 correctamente y 107 de manera incorrecta logrando obtener una precisión general de 88,51 %.

Figura 50

Matriz de confusión del algoritmo Gradient Boosted Trees con 3 valores de clase

Scorer View

Confusion Matrix

	DEFICIENTE (Predicted)	MUY DEFICIENTE (Pr...	SATISFACTORIO (Pre...	
DEFICIENTE (Actual)	9	0	58	13.43%
MUY DEFICIENTE (Ac...	2	21	17	52.50%
SATISFACTORIO (Act...	20	10	794	96.36%
	29.03%	67.74%	91.37%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
88.51%	11.49%	0.324	824	107

Nota: Se obtuvo utilizando KNIME.

Respecto al algoritmo de Naive Bayes, de acuerdo con lo que se aprecia en la figura 51, se pudo obtener una precisión de 8,96 % clasificando calificaciones “DEFICIENTES”, 30% clasificando calificaciones “MUY DEFICIENTES” y 93,81% clasificando calificaciones “SATISFACTORIAS”. Además de los 831 registros utilizados para el testeo, el modelo pudo clasificar correctamente 791 y de manera incorrecta 140 logrando una precisión general de 84,96 %.

Figura 51

Matriz de confusión del algoritmo Naive Bayes con 3 valores de clase

Scorer View

Confusion Matrix

	DEFICIENTE (Predicted)	MUY DEFICIENTE (Pr...	SATISFACTORIO (Pre...	
DEFICIENTE (Actual)	6	8	53	8.96%
MUY DEFICIENTE (Ac...	5	12	23	30.00%
SATISFACTORIO (Act...	31	20	773	93.81%
	14.29%	30.00%	91.05%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
84.96%	15.04%	0.199	791	140

Nota: Se obtuvo utilizando KNIME

5.4. RESULTADOS A NIVEL INFERENCIAL

El objetivo de este apartado de la tesis es mostrar de manera detallada el estudio en el marco de las pruebas de normalidad de cada una de las variables y factores de nuestro trabajo.

a. Para los valores correspondientes al Periodo 2019 – I

   La variable: Accesos al aula virtual 2019-I

- Establecemos las hipótesis.

H_0: El acceso al aula virtual en el periodo académico 2019 - I tiene una distribución normal.

H_1: El acceso al aula virtual en el periodo académico 2019 - I no sigue una distribución normal.

- Se determinó el nivel de significancia con un valor de α = 0,05.
- Por otro lado se utilizó el estadístico de prueba, en nuestro caso se utilizó la prueba de Kolmogorov-Smirnov, ya que el tamaño de muestra es mayor a 50 (Hanusz y Tarasińska, 2015).
- Se procedió a determinar el valor estadístico, los resultados de la prueba de Kolmogorov-Smirnov fueron hechos utilizándola herramienta de software estadístico IBM SPSS Statistics y se muestran en la tabla 7.

Tabla 7

Resultados de la aplicación de la prueba de normalidad el acceso al aula virtual en el periodo académico 2019 - I

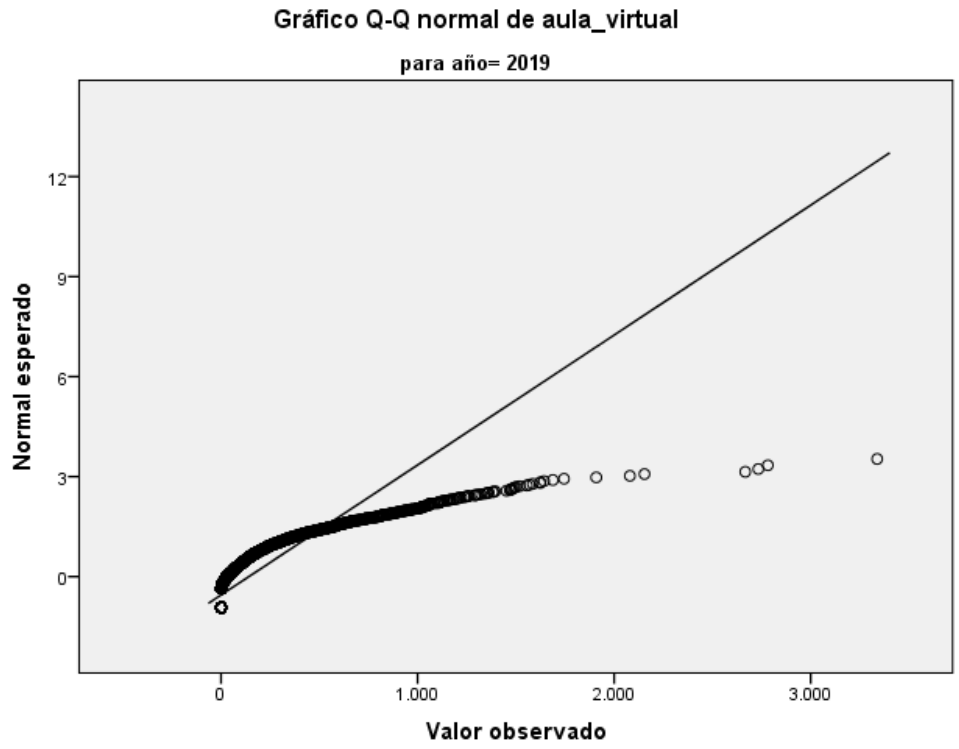
Accesos al aula virtual	Kolmogorov-Smirnov			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
2019 – I	0,291	4799	0,000	0,585	4799	0,000

Nota: Se obtuvo utilizando SPSS.

- Región crítica para el acceso al aula virtual en el periodo académico 2019 - I, el nivel de significancia elegido es de 5 % (α = 0,05), el valor obtenido de significación fue de 0,000, es menor a 0,05, por lo tanto, no sigue una distribución normal. En la figura 52 se puede apreciar el gráfico Q-Q del acceso al aula virtual en el periodo académico 2019 – I.

Figura 52

Gráfico Q-Q de la prueba de normalidad del acceso al aula virtual en el periodo académico 2019 – I.



Nota: Se obtuvo utilizando SPSS

Variable: Promedio final de notas 2019-I

Se establece las hipótesis.

H_0: El promedio final de notas en el periodo académico 2019 - I tiene una distribución normal.

H_1: El promedio final de notas en el periodo académico 2019 - I no sigue una distribución normal.

- Se determina el nivel de significancia con un valor de α = 0,05.

- Se propone el estadístico de prueba, en nuestro caso para la prueba estadística se utilizó la prueba de Kolmogorov-Smirnov, ya que nuestro tamaño de muestra es mayor a 50.
- En lo que respecta a la determinación del valor estadístico, los resultados de aplicar la prueba de Kolmogorov-Smirnov fueron calculados con la ayuda de la herramienta de software estadístico IBM SPSS Statistics y se muestran en la tabla 8.

Tabla 8

Valores de resultado de la aplicación de la prueba de normalidad a la variable promedio final de las notas en el periodo académico 2019 - I

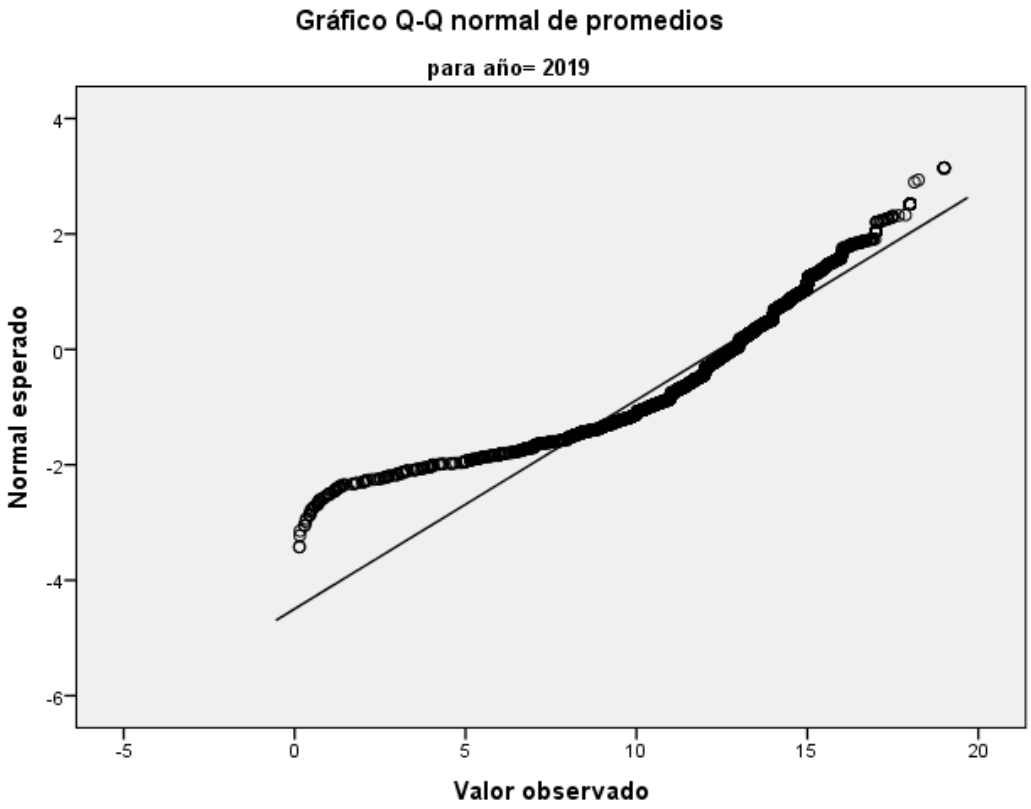
Promedio final de las notas	Kolmogorov-Smirnov			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
2019 – I	0,111	4799	0,000	0,909	4799	0,000

Nota: Se obtuvo utilizando SPSS

- Región crítica para el promedio en el periodo académico 2019 - I, el nivel de significancia elegida es de 5 % (α = 0,05), el valor obtenido de significación fue de 0,000, es menor a 0,05 por lo tanto no siguen una distribución normal. En la figura 53 se puede apreciar el gráfico Q-Q del promedio final de notas en el periodo académico 2019 – I.

Figura 53

Gráfico Q-Q de la prueba de normalidad del promedio final de las notas en el periodo académico 2019 – I



Nota: Se obtuvo utilizando SPSS

Para los valores correspondientes al Periodo 2020 – I

Variable: Accesos al aula virtual 2020-I

- Se establecen las hipótesis.

  H_0: El acceso al aula virtual en el periodo académico 2020 - I tiene una distribución normal.

  H_1: El acceso al aula virtual en el periodo académico 2020 - I no sigue una distribución de carácter normal.

- Se asume un nivel de significancia con el valor de α = 0,05.
- Para nuestro caso se utilizó la prueba de Kolmogorov-Smirnov como estadístico de prueba, ya que el tamaño de muestra es mayor a 50.
- Se determinó del valor estadístico, los resultados de la prueba de Kolmogorov-Smirnov fueron calculados con ayuda del software estadístico IBM SPSS Statistics y se muestran en la tabla 9.

Tabla 9

Valores de los resultados de la aplicación de la prueba de normalidad a la variable: acceso al aula virtual en el periodo académico 2020 - I

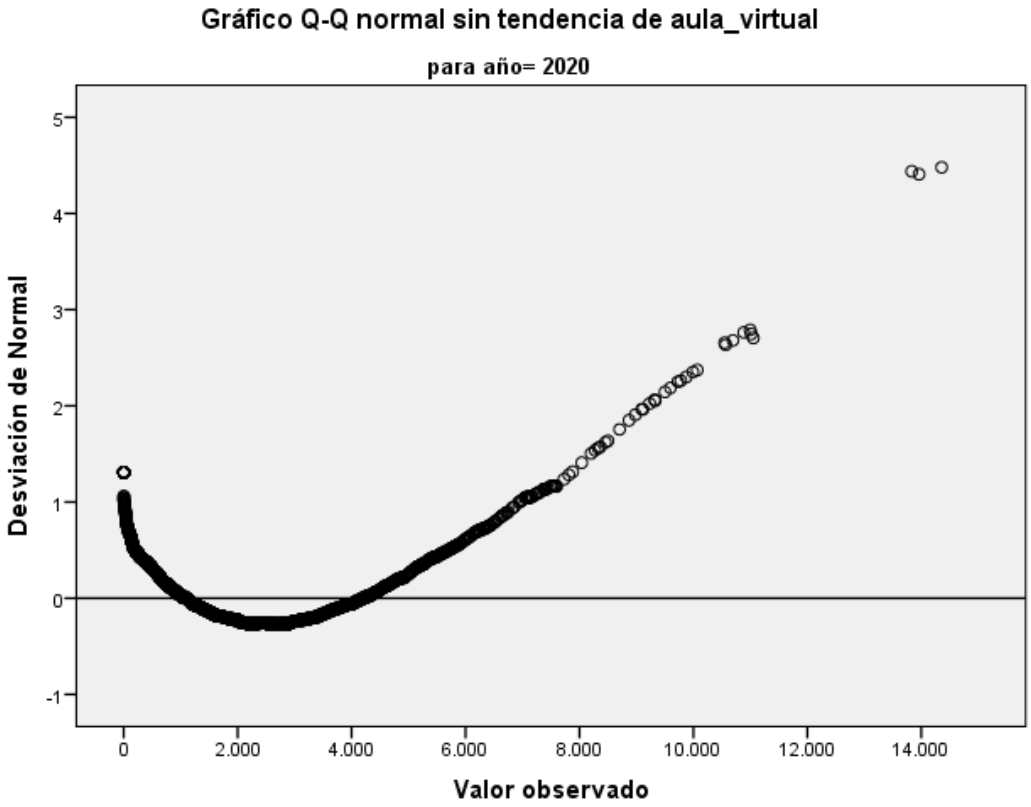
Accesos al aula virtual	Kolmogorov-Smirnov			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	Gl	Sig.
2020 – I	0,103	4799	0,000	0,873	4799	0,000

Nota: Se obtuvo utilizando SPSS

- Región crítica para el acceso al aula virtual en el periodo académico 2020 - I, el nivel de significancia elegida es de 5 % (α = 0,05), el valor obtenido de significación fue de 0,000, es menor a 0,05, por lo tanto, no sigue una distribución normal. En la figura 54 se puede apreciar el gráfico Q-Q del acceso al aula virtual en el periodo académico 2020 – I.

Figura 54

Gráfico Q-Q de la prueba de normalidad del acceso al aula virtual en el periodo académico 2020 – I.



Nota: Se obtuvo utilizando SPSS

Variable: Promedio final de notas 2020-I

- Se establecen las hipótesis.

  H_0: El promedio final de notas en el periodo académico 2020 - I tiene una distribución normal.

  H_1: El promedio final de notas en el periodo académico 2020 - I no sigue una distribución de carácter normal.

- Se asume el nivel de significancia con el valor α = 0,05.

- En nuestro caso se utilizará la prueba de Kolmogorov-Smirnov como estadístico de prueba, ya que nuestro tamaño de muestra es mayor a 50.
- Se determinó del valor estadístico, los resultados de la aplicación de la prueba de Kolmogorov-Smirnov fueron calculados con la ayuda del software IBM SPSS Statistics y se muestran en la tabla 10.

Tabla 10

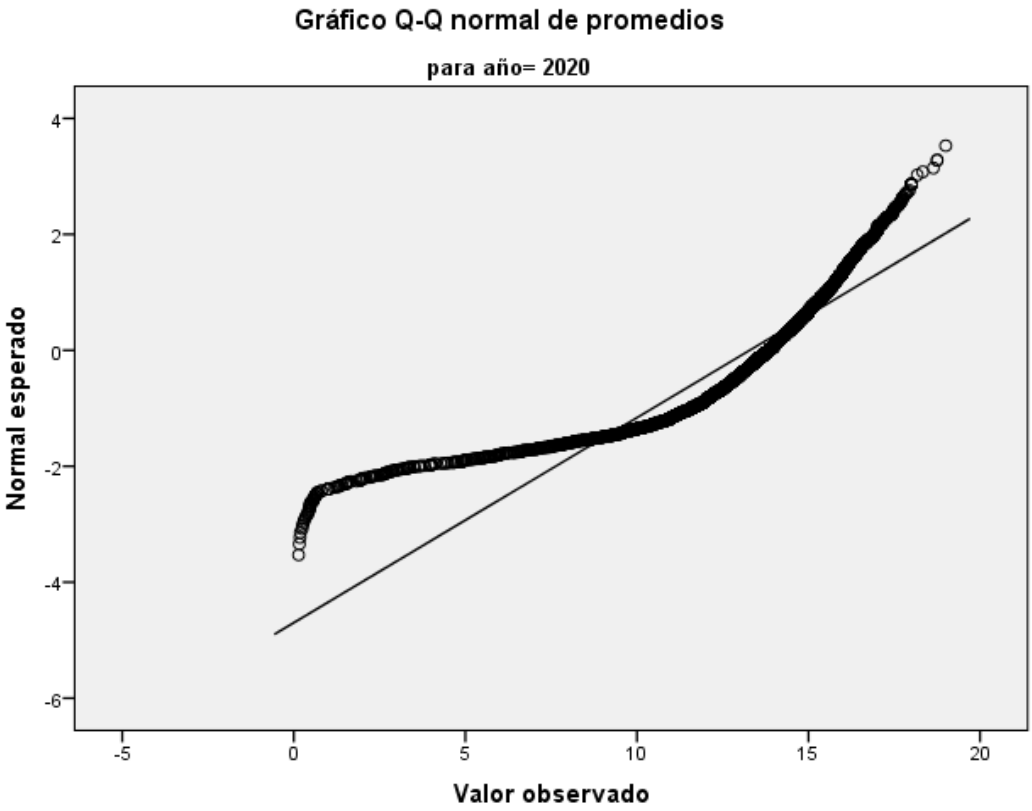
Valores de resultados de la aplicación de la prueba a la variable: promedio final de las notas en el periodo académico 2020 – I.

Promedio final de notas	Kolmogorov-Smirnov			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
2020 – I	0,139	4799	0,000	0,825	4799	0,000

- Región crítica para el promedio en el periodo académico 2020 - I, el nivel de significancia elegida es de 5 % (α = 0,05), el valor obtenido de significación fue de 0,000, es menor a 0,05, por lo tanto, no siguen una distribución normal. En la figura 55 se puede apreciar el gráfico Q-Q del promedio final de notas en el periodo académico 2020 – I.

Figura 55

Gráfico Q-Q de la prueba de normalidad del promedio final de las notas en el periodo académico 2020 – I



Nota: Se obtuvo utilizando SPSS

5.5. ANÁLISIS INFERENCIAL

En este apartado de nuestra tesis se quiere demostrar y probar la hipótesis planteada en nuestro estudio.

a. Subhipótesis 1

H_0: No existe una diferencia significativa entre el acceso al aula virtual en el periodo académico 2019 – I y 2020 - I.

H_1: Existe una diferencia significativa entre el acceso al aula virtual en el periodo académico 2019 – I y 2020 - I

Paso 1: Se establece las hipótesis

$$H_0: \overline{X_1} = \overline{X_2}$$

$$H_0: \overline{X_1} \neq \overline{X_2}$$

Donde:

$\overline{X_1}$ = Acceso al aula virtual en el periodo académico 2019 - I.

$\overline{X_2}$ = Acceso al aula virtual en el periodo académico 2020 - I.

Paso 2: Se determina el valor del nivel de significancia $\alpha = 0{,}05$.

Paso 3: Especificamos la zona de rechazo, todos los valores de $p < 0{,}05$.

Paso 4: En nuestro caso se utilizó la prueba U de Mann-Whitney, ya que como se analizó en la tabla 8 y 10 los datos de ambas virtualizaciones no siguen una distribución normal. Para encontrar el valor p se utilizó el software SPSS. Los resultados se muestran en la tabla 11.

Tabla 11

Resultados de la prueba U de Mann-Whitney para la subhipótesis 1

Acceso al aula virtual	N	Rango promedio	Suma de rangos	U de Mann-Whitney	P valor
2019 – I	4799	2556,17	12267071,50	749471,5	0,000
2020 – I	4799	7042,83	33798529,50		

Nota: Se obtuvo utilizando SPSS.

Decisión: De acuerdo a los datos obtenidos en la tabla 11, el valor estimado de p-valor = 0,000 < 0,05; por lo tanto, se rechaza la hipótesis nula de que no existe una diferencia significativa(Mann y Whitney, 1947), concluyendo que el acceso

al aula virtual en periodo académico 2020 – I fue mayor al acceso del aula virtual en el periodo académico 2019 – I.

b. Subhipótesis 2

H_0: No existe una diferencia significativa entre el promedio de las notas finales en el periodo académico 2019 – I y 2020 - I.

H_1: Existe una diferencia significativa entre el promedio de las notas finales en el periodo académico 2019 – I y 2020 - I

Paso 1: Se establecen las hipótesis

$$H_0: \overline{X_1} = \overline{X_2}$$

$$H_0: \overline{X_1} \neq \overline{X_2}$$

Donde:

$\overline{X_1}$ = Promedio de las notas finales en el periodo académico 2019 - I.

$\overline{X_2}$ = Promedio de las notas finales en el periodo académico 2020 - I.

Paso 2: Se especifica el nivel de significancia $\alpha = 0{,}05$.

Paso 3: Se establece también la zona de rechazo, todos los valores de $p < 0{,}05$.

Paso 4: Se utilizó la prueba U de Mann-Whitney, ya que como se analizó en la tabla 8 y 10 los datos de ambas virtualizaciones no siguen una distribución normal. Para encontrar el valor p se utilizó el software SPSS. Los resultados se muestran en la tabla 12.

Tabla 12

Resultados de la prueba U de Mann-Whitney para la subhipótesis 1

Promedios de las notas finales	N	Rango promedio	Suma de rangos	U de Mann-Whitney	P valor
2019 – I	4799	4180,44	20061940,50	8544340,5	0,000
2020 – I	4799	5418,56	26003660,50		

Nota: Se obtuvo utilizando SPSS.

Decisión: En función a los resultados obtenidos en la tabla 12, el valor estimado de p-valor = 0,000 < 0,05; en consecuencia, se puede rechazar la hipótesis nula de que no existe una diferencia significativa, concluyendo que el promedio de las notas finales en el periodo académico 2020 – I fue mayor al promedio de notas finales en el periodo académico 2019 – I.

CAPÍTULO VI

ANÁLISIS Y DISCUSIÓN

En lo que concierne al análisis de datos utilizando minería de datos respecto al uso del aula virtual de los estudiantes de la Facultad de Ingeniería-FAIN de la UNJBG en el período académico 2020-I se logró determinar que el número de accesos al aula virtual de los estudiantes de la Facultad de Ingeniería es de : 2 063 047 y se constituyen en la cuarta facultad con más accesos, después de la Facultad de Ciencias Agropecuarias-FCAG con 3 072 320 accesos, Facultad de Ciencias de la Salud-FACS con 2 897 235 accesos y Facultad de Ciencias Jurídicas y Empresariales-FCJE con 2 402 332 accesos, a diferencia del período 2019-I donde la Facultad de Ingeniería apenas contó con 126 193 accesos, evidenciándose un aumento exponencial de 1 534,83 % entre el 2019-I y el 2020-I en lo que se refiere a accesos al aula virtual.

Por otro lado, en el análisis se utilizó diversas técnicas de minería de datos haciendo uso de diversas herramientas como: Python, KNIME, Tableau, RapidMiner que fueron de mucha ayuda. En lo que respecta al rendimiento académico del período 2020-I se ha podido evidenciar que la Facultad de Ingeniería posee el penúltimo promedio en calificaciones alcanzando 11,918 a diferencia de la Facultad de Ciencias Jurídicas y Empresariales que con un promedio de 14,584 se encuentra en el primer lugar de calificaciones de la UNJBG. Y en el período 2019-I se pudo evidenciar que la Facultad de Ingeniería alcanzó un 10,903 de promedio y la Facultad de Ciencias Jurídicas y Empresariales alcanzó un 14,584 notándose un incremento en el promedio de la Facultad de Ingeniería en el período académico 2020-I de 1,015 puntos lo que representa un incremento del 9,31 % respecto al 2019-I y una disminución a las notas en la Facultad de Ciencias Jurídicas y empresariales. Esto se debe a que en las Carreras de Letras y Negocios de la Facultad de Ciencias Jurídicas y

empresariales no tuvo mayor efecto el uso del aula virtual en las notas sin embargo en la Facultad de Ingeniería si, ya que la mayoría de los cursos utilizan laboratorios, pensamos que restringió a los profesores ser más exigentes en las evaluaciones y en todo caso fueron más condescendientes a la hora de calificar.

En lo que respecta al modelo de predicción basado en minería de datos, se logró implementar un modelo basado en el algoritmo de Gradient Boosted Trees con la mayor precisión general llegando a predecir correctamente las calificaciones de los estudiantes en función principalmente a los accesos a el aula virtual, con una precisión de 91,79 % para dos valores de clase (SATISFACTORIO Y DEFICIENTE) y el modelo que mayor precisión arrojó en la clasificación de las calificaciones con tres valores de clase (SATISFACTORIO, DEFICIENTE Y MUY DEFICIENTE) fue el modelo generado a partir del algoritmo de Random Forest con una precisión de 89,26 %, creemos que esta diferencia en la precisión se debe a la segmentación de valores de clase, ya que las clases MUY DEFICIENTE y DEFICIENTE son considerablemente más pequeñas que la clase SATISFACTORIA.

Finalmente, desde el punto de vista inferencial se ha demostrado estadísticamente que los accesos al aula virtual en el período académico 2019-I y 2020-I no siguen una distribución normal y aplicando la prueba Mann Whitney, hemos evidenciado que el acceso al aula virtual en periodo académico 2020 – I fue mayor al acceso del aula virtual en el periodo académico 2019 – I. De la misma forma utilizando la misma prueba, se ha evidenciado que las calificaciones tanto del periodo 2019-I y 2020-I no se ajustaban a una distribución normal y pudimos demostrar que las calificaciones en el período académico 2020-I fueron mayores que las calificaciones del período académico 2019-I, lo que permite afirmar que el nivel de acceso al aula virtual si tuvo un efecto significativo positivo en las calificaciones de los estudiantes.

La presente investigación tiene la particularidad de mostrarnos diversos resultados al aplicar distintos algoritmos, ya que los modelos conseguidos son

aplicables a nuestra realidad. Se logró entrenar con 2724 registros de estudiantes y realizar el testeo con 682 registros de estudiantes para la clasificación de los rendimientos académicos en función a diversas variables, pero fundamentalmente los accesos al aula virtual en el período 2020-I, logrando determinar que los accesos al aula virtual tuvieron un efecto significativamente alto en la mejora de los rendimientos académicos.

En contraste a la investigación titulada “Clarify of the Random Forest Algorithm in an Educational Field”(Ahmed y Hikmat Sadiq, 2018) revisada en nuestros antecedentes, donde logran generar un modelo a partir del algoritmo Random Forest obteniendo una precisión de 83,56 %, en la presente investigación logramos alcanzar los 89,26% de precisión que se considera que se debe al análisis más minucioso de los datos, consideración de las calificaciones anteriores específicamente del 2019-I también y de la limpieza previa de ellos.

Por otro lado de acuerdo al trabajo titulado “Predicting Student Academic Performance using Support Vector Machine and Random Forest”(Alamri et al., 2020) también referenciado en nuestros antecedentes, los investigadores logran obtener una precisión de 94,43 % utilizando un modelo basado en el algoritmo SVM y una precisión de 91,59 % con un modelo basado en el algoritmo de Random Forest en contraste con nuestros resultados de 90,47 % utilizando SVM y 90,62 % utilizando Random Forest. En este caso se considera que esta diferencia se debe a que Alamri enfocó su estudio solo a dos asignaturas (Matemática y Portugués) a diferencia nuestra que aplicamos a todas las asignaturas en consecuencia trabajamos con muchos más datos lo que genera más ruido y aparición de outliers.

CONCLUSIONES

1. Aplicando técnicas de minería de datos, se logró generar un modelo que permita determinar el efecto positivo que tuvo el acceso al aula virtual sobre el rendimiento académico de los estudiantes de la Facultad de Ingeniería en el período académico 2020-I en la época más crítica de la pandemia.

2. En lo que se refiera al análisis de datos de accesos al aula virtual utilizando técnicas de preprocesado de datos en el marco de trabajo de KDD, se logró evidenciar que los accesos tuvieron un incremento impresionante de cerca del 1 534 %, esto considerando que en el 2020-I la FAIN tuvo 2 063 047 accesos en total a causa de las restricciones sanitarias de la pandemia y en el 2019-I tuvo apenas 126 193 accesos.

3. En lo que se refiere al análisis de los datos de rendimiento académico se utilizaron técnicas de minería de datos para concluir que la Facultad de Ingeniería logra en el período académico 2020-I un incremento de 9,31 % en el promedio final de calificaciones de sus estudiantes respecto al periodo 2019-I.

4. En cuanto a los modelos de clasificación del rendimiento académico se concluye que, para los datos de la Universidad Nacional Jorge Basadre Grohmann, con un escenario pandémico y con una partición de entrenamiento del 80 % y testeo del 20 %, con muestreo estratificado y con valor semilla de 13, el modelo basado en el algoritmo Gradient Boosted Trees fue el más preciso logrando obtener una precisión de 91,79 % en la clasificación del rendimiento académico con dos valores: SATISFACTORIO y DEFICIENTE. En tanto para una clase de tres valores: SATISFACTORIO,

DEFICIENTE y MUY DECIEINTE fue el modelo basado en el algoritmo Random Forest el que me mejor precisión arrojó alcanzando un 89,26 %.

5. El preprocesamiento de datos fue una tarea muy importante y determinante en el logro de los buenos resultados de los algoritmos aplicados. Se concluye que debemos invertir mucho tiempo en estudiar bien los datos, sus características, sus formas, sus tendencias, comportamientos y aplicar todas las técnicas de preprocesamiento de datos existentes, principalmente aquellas relacionadas a los outliers y tratamiento de valores perdidos.

6. Se concluye también que no existen políticas universitarias para poder integrar la analítica de datos en los procesos académicos y administrativos. Lo que no permite ni permitirá un acompañamiento más efectivo a todos los procesos en función a los datos que se van generando a cada momento.

RECOMENDACIONES

1. Se podría usar también redes neuronales como MultiLayer Perceptron para poder predecir las calificaciones, en nuestra tesis lo intentamos, pero obtuvimos resultados con muy baja precisión, probablemente problemas en la normalización y la naturaleza tan diversa de los datos contribuyan a estos bajos índices de precisión.

2. Se recomienda, aplicar técnicas de detección de Outliers más avanzadas como: Numeric Outlier,Z-Score, DBScan o Insolation Forest lo que nos permitiría mejorar las precisiones.

3. Sería recomendable implementar modelos de clustering, para poder segmentar la data y descubrir nuevos grupos de estudiantes o docentes con características comunes para aplicar estrategias didácticas distintas.

4. Se sugiere generar una conciencia institucional de Education Data Mining, es decir integrar todos los procesos a las tareas de analítica de datos, desde su captura, almacenamiento y procesamiento.

5. Los datos en las instituciones son fundamentales considerarlos como activos importantes, para analizarlos y tomar decisiones correctas.

6. Se recomienda implementar protocolos de tratamiento de datos en la Universidad Nacional Jorge Basadre Grohmann.

7. Sería recomendable predecir otras variables como la motivación, ausentismo y deserción de los estudiantes.

8. Sería importante y recomendable obtener a través de las Instituciones educativas de los estudiantes ingresantes los datos de su rendimiento académico escolar para ampliar el trabajo.

9. Este modelo podría ser replicado fácilmente para las especialidades de la UNJBG, Escuela de Posgrado, ITEL, Centro de Idiomas, CEPU, etc. siempre y cuando se tenga acceso a los datos y estos sean de calidad.

REFERENCIAS BIBLIOGRÁFICAS

Adams Becker, S., Cummins, M., Freeman, A., Ifenthaler, D., Johnson, L., Vardaxis, N. J., y Australia, O. U. (2013). *Technology Outlook for Australian Tertiary Education 2013-2018:* An NMC Horizon Project Regional Analysis. In *New Media Consortium*.

Ahmed, N. S., y Hikmat Sadiq, M. (2018). Clarify of the Random Forest Algorithm in an Educational Field. *ICOASE 2018 - International Conference on Advanced Science and Engineering*, 179–184. https://doi.org/10.1109/ICOASE.2018.8548804

Alamri, L. H., Almuslim, R. S., Alotibi, M. S., Alkadi, D. K., Ullah Khan, I., y Aslam, N. (2020). Predicting Student Academic Performance using Support Vector Machine and Random Forest. *2020 3rd International Conference on Education Technology Management*, 100–107. https://doi.org/10.1145/3446590.3446607

Alloghani, M., Al-Jumeily, D., Hussain, A., Aljaaf, A. J., Mustafina, J., y Petrov, E. (2019). Application of machine learning on student data for the appraisal of academic performance. *Proceedings - International Conference on Developments in ESystems Engineering, DeSE*, *2018-Septe*, 157–162. https://doi.org/10.1109/DeSE.2018.00038

Altaf, S., Soomro, W., y Rawi, M. I. M. (2019). Student Performance Prediction using Multi-Layers Artificial Neural Networks: A case study on educational data mining. *ACM International Conference Proceeding Series*, 59–64. https://doi.org/10.1145/3325917.3325919

Arnold, K. E., y Pistilli, M. D. (2012). Course Signals at Purdue: Using Learning Analytics to Increase Student Success. In LAK ’12 (Ed.), *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (Vol. 23, Issue 14, pp. 267–270). Association for Computing Machinery. https://doi.org/10.1145/2330601.2330666

Asif, R., Merceron, A., Ali, S. A., y Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers and Education*, *113*, 177–194. https://doi.org/10.1016/j.compedu.2017.05.007

Athani, S. S., Kodli, S. A., Banavasi, M. N., y Hiremath, P. G. S. (2018). Student performance predictor using multiclass support vector classification algorithm. *Proceedings of IEEE International Conference on Signal Processing and Communication, ICSPC 2017*, *2018-Janua*(July), 341–346. https://doi.org/10.1109/CSPC.2017.8305866

Baker, R. (2010). Data Mining. In P. Peterson, E. Baker, y B. McGaw (Eds.), *International Encyclopedia of Education (Third Edition)* (Third Edit, pp. 112–118). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-08-044894-7.01318-X

Baker, R., y Yacef, K. (2009). The State of Educational Data Mining in 2009 : A Review and Future Visions. *Journal of Educational Data Mining*, *1*(1), 3–16.

Baranyi, M., Nagy, M., y Molontay, R. (2020). Interpretable Deep Learning for University Dropout Prediction. *SIGITE 2020 - Proceedings of the 21st Annual Conference on Information Technology Education*, 13–19. https://doi.org/10.1145/3368308.3415382

Bouchard, K., Gonzales, L., Maitre, J., y Gaboury, S. (2020). Features Exploration for Grades Prediction using Machine Learning. *ACM International Conference Proceeding Series*, 78–83. https://doi.org/10.1145/3411170.3411232

Bowers, A. J. (2010). Analyzing the longitudinal K-12 grading histories of entire cohorts of students: Grades, data driven decision making, dropping out and hierarchical cluster analysis. *Practical Assessment, Research and Evaluation*, *15*(7). https://doi.org/10.7916/D8QC02TX

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Caraballo, E. (2020). *EDUCACION Y MACHINE LEARNING: La puerta de entrada a un nuevo paradigma*. *35*. http://www.educa.org.do/wp-content/uploads/2017/07/Nota-de-trabajo-EDUCA-No-35.pdf

Cavazos, R., y Garza, S. (2018). *Learning Models for Student Performance Prediction*. Springer International Publishing. https://doi.org/10.1007/978-3-030-02840-4

Chakraborty, A., Goswami, D., y Hassanien, A. E. (2017). Studies in Computational Intelligence 912 Artificial Intelligence for Sustainable Development : Theory , Practice and Future Applications. In *Arabian Journal of Geosciences* (Vol. 10, Issue 17).

Chango, W., Cerezo, R., y Romero, C. (2019). Predicting academic performance of university students from multi-sources data in blended learning. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3368691.3368694

Chanlekha, H., y Niramitranon, J. (2018). Student performance prediction model for early-identification of at-risk students in traditional classroom settings. *MEDES 2018 - 10th International Conference on Management of Digital EcoSystems*, 239–245. https://doi.org/10.1145/3281375.3281403

Chatti, M. A., Dyckhoff, A. L., Schroeder, U., y Thüs, H. (2012). A reference model for learning analytics Mohamed Amine Chatti *, Anna Lea Dyckhoff ,. *Int. J. Technology Enhanced Learning*, *4*(CiL), 318–331.

Chen, H., y Ward, P. A. S. (2020). Predicting student performance using data from an auto-grading system. *CASCON 2019 Proceedings - Conference of the Centre for Advanced Studies on Collaborative Research - Proceedings of the 29th Annual International Conference on Computer Science and Software Engineering*, 234–243.

Chen, Y., Wang, Y., y Xiao, X. (2011). *Knowledge Discovery Technology Based on Access Information Mining on Knowledge Warehouse*. 1285–1288.

Cobo, G., García-Solórzano, D., Morán, J. A., Santamaría, E., Monzo, C., y Melenchón, J. (2012). Using agglomerative hierarchical clustering to model learner participation profiles in online discussion forums. *ACM International Conference Proceeding Series*, *May*, 248–251. https://doi.org/10.1145/2330601.2330660

Cortez, P., y Silva, A. (2008). Using data mining to predict secondary school

student performance. In eds. lit. BRITO, A.; TEIXEIRA, J. (Ed.), *Proceedings of 5th Annual Future Business Technology Conference, Porto, 2008* (pp. 5–12).

Davalbhakta, S., Advani, S., Kumar, S., Agarwal, V., Bhoyar, S., Fedirko, E., Misra, D. P., Goel, A., Gupta, L., y Agarwal, V. (2020). A systematic review of the smartphone applications available for corona virus disease 2019 (COVID19) and their assessment using the mobile app rating scale (MARS). *MedRxiv*, *2019*. https://doi.org/10.1101/2020.07.02.20144964

Devasia, T., Vinushree, T. P., y Hegde, V. (2016). Prediction of students performance using Educational Data Mining. *Proceedings of 2016 International Conference on Data Mining and Advanced Computing, SAPIENCE 2016*, 91–95. https://doi.org/10.1109/SAPIENCE.2016.7684167

Dipert, R. R. (2002). The substantive impact of computers on philosophy prolegomena to a computational and information-theoretic metaphysics. *Metaphilosophy*, *33*(1–2), 146–157. https://doi.org/10.1111/1467-9973.00222

ElGamal, A. F. (2013). An Educational Data Mining Model for Predicting Student Performance in Programming Course. *International Journal of Computer Applications*, *70*(17), 22–28. https://doi.org/10.5120/12160-8163

Escanés, G., Herrero, V., Merlino, A., y Ayllón, S. (2014). Deserción en educación a distancia: factores asociados a la elección de modalidad como desencadenantes del abandono universitario. *Virtualidad, Educación y Ciencia*, *5*(9), 45–55.

Feng, M., Heffernan, N. T., y Koedinger, K. R. (2006). Predicting State Test Scores Better with Intelligent Tutoring Systems: Developing Metrics to Measure Assistance Required. In M. Ikeda, K. D. Ashley, y T.-W. Chan (Eds.), *8th International Conference, ITS 2006: Vol. 3744 LNCS* (pp. 31–40). Springer.

Floridi, L. (2003). Two Approaches to the Philosophy of Information. *Minds and Machines*, *13*(4), 459–469. https://doi.org/10.1023/A:1026241332041

Formia, S. (2012). *Evaluación de técnicas de Extracción de Conocimiento en*

Bases de Datos y su aplicación a la deserción de alumnos universitarios. 80. http://sedici.unlp.edu.ar/bitstream/handle/10915/26772/Documento_completo.pdf?sequence=1&isAllowed=y

García-Balboa, J. L., Alba-Fernández, M. V, Ariza-López, F. J., y Rodríguez-Avi, J. (2018). H*omogeneity test for confusion matrices : a method and an example*. 1203–1205.

Gašević, D., Dawson, S., y Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*, *59*(1), 64–71. https://doi.org/10.1007/s11528-014-0822-x

Guevara, P., Verdesoto, A., y Castro, N. (2020). *Metodologías de investigación educativa (descriptivas, experimentales, participativas, y de investigación-acción)*. *3*, 163–173. https://doi.org/10.26820/recimundo/4.(3).julio.2020.163-173

Han, J., Kamber, M., y Pei, J. (2012). Data Mining: Concepts and Techniques. In *Journal of Chemical Information and Modeling* (Third Edit, Vol. 53, Issue 9). http://library.books24x7.com/toc.aspx?bkid=44712

Hanusz, Z., y Tarasińska, J. (2015). Normalization of the Kolmogorov–Smirnov and Shapiro–Wilk tests of normality. *Biometrical Letters*, *52*(2), 85–93. https://doi.org/10.1515/bile-2015-0008

Hardman, J., Paucar-Caceres, A., y Fielding, A. (2013). Predicting Students' Progression in Higher Education by Using the Random Forest Algorithm. *Systems Research and Behavioral Science*, *30*(2), 194–203. https://doi.org/10.1002/sres.2130

Hecht-Nielsen, R. (1989). Theory of the Backpropagation Neural Network. *International 1989 Joint Conference on Neural Networks*, 593–605. https://doi.org/10.1109/IJCNN.1989.118638

Huang, S., y Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers and Education*, *61*(1), 133–145. https://doi.org/10.1016/j.compedu.2012.08.015

Ingrassia, S., y Morlini, I. (2005). Neural network modeling for small datasets.

Technometrics, *47*(3), 297–311. https://doi.org/10.1198/004017005000000058

Jalota, C., y Agrawal, R. (2019). Analysis of Educational Data Mining using Classification. *Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Prespectives and Prospects, COMITCon 2019*, 243–247. https://doi.org/10.1109/COMITCon.2019.8862214

Johnson, L., Adams, S., y Cummins, M. (2012). *The NMC Horizon Report: 2012 Higher Education Edition*.

Jones, K. M. L. (2019). Learning analytics and higher education: a proposed model for establishing informed consent mechanisms to promote student privacy and autonomy. *International Journal of Educational Technology in Higher Education*, *16*(1). https://doi.org/10.1186/s41239-019-0155-0

José, M. N. J. (2020). Sociedad digital: Gestión organizacional tras el covid-19. *Revista Venezolana de Gerencia*, *25*(90), 394–401. https://doi.org/10.37960/rvg.v25i90.32383

Long, P. D., y Siemens, G. (2014). Penetrare la nebbia: tecniche di analisi per l'apprendimento. *Italian Journal of Educational Technology*, *22*(3), 132–137. https://doi.org/10.17471/2499-4324/195

Mahboob, T., Irfan, S., y Karamat, A. (2017). A machine learning approach for student assessment in E-learning using Quinlan's C4.5, Naive Bayes and Random Forest algorithms. *Proceedings of the 2016 19th International Multi-Topic Conference, INMIC 2016*. https://doi.org/10.1109/INMIC.2016.7840094

Mann, H. B., y Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, *18*(1), 50–60. https://doi.org/10.1214/aoms/1177730491

Martínez-Ortega, R., y Tuya-Pendás, L. (2009). Rev haban cienc méd La Habana, Vol. VIII No.2, abr-jun 2009. *Revista Habanera de Ciencias Médicas*, *VIII*(2).

Martínez, V. (2017). Educación presencial versus educación a distancia. *Polired.Upm.Es*, *9*, 108–116. http://webcast.berkeley.edu

Mishra, T., Kumar, D., y Gupta, S. (2014). Mining students' data for prediction performance. *International Conference on Advanced Computing and Communication Technologies, ACCT*, 255–262. https://doi.org/10.1109/ACCT.2014.105

Moulet, M., y Kodratoff, Y. (1995). From machine learning towards knowledge discovery in databases. *IEE Colloquium on Knowledge Discovery in Databases (Digest No. 1995/021 (A))*, 5/1-5/3. https://doi.org/10.1049/ic:19950116

Opitz, D., y Maclin, R. (1999). Popular Ensemble Methods: An Empirical Study. *Journal of Artificial Intelligence Research*, *11*, 169–198. https://doi.org/10.1613/jair.614

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, *26*(1), 217–222. https://doi.org/10.1080/01431160412331269698

Pardos, Z. A., Heffernan, N. T., Anderson, B., y Heffernan, C. L. (2007). The Effect of Model Granularity on Student Performance Prediction Using Bayesian Networks. In C. Conati, K. McCoy, y G. Paliouras (Eds.), *User Modeling 2007: Vol. 4511 LNCS* (pp. 435–439). Springer Berlin Heidelberg.

Pasini, A. (2015). Artificial neural networks for small dataset analysis. *Journal of Thoracic Disease*, *7*(5), 953–960. https://doi.org/10.3978/j.issn.2072-1439.2015.04.61

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, *41*(4 PART 1), 1432–1462. https://doi.org/10.1016/j.eswa.2013.08.042

Poole, D. (1998). Computational intelligence: a logical approach. *Choice Reviews Online*, *35*(10), 35-5701-35–5701. https://doi.org/10.5860/choice.35-5701

Prenkaj, B., Velardi, P., Stilo, G., Distante, D., y Faralli, S. (2020a). A Survey of Machine Learning Approaches for Student Dropout Prediction in Online Courses. *ACM Comput. Surv.*, *53*(3), 1–14. https://doi.org/10.1145/3388792

Prenkaj, B., Velardi, P., Stilo, G., Distante, D., y Faralli, S. (2020b). A Survey of Machine Learning Approaches for Student Dropout Prediction in Online Courses. *ACM Computing Surveys*, *53*(3). https://doi.org/10.1145/3388792

Rapaport, W. J. (1986). Philosophy of Artificial Intelligence. *Teaching Philosophy*, *9*(2), 103–120. https://doi.org/10.5840/teachphil19869220

Riquelme, J., Ruiz, R., y Gilbert, K. (2006). Minería de Datos: Conceptos y Tendencias. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, *10*(29), 11–18. https://www.redalyc.org/pdf/925/92502902.pdf

Rodríguez, M., y Mendivelso, F. (2018). Diseño de investigación de Corte Transversal. *Revista Médica Sanitas*, *21*(3), 141–146. https://doi.org/10.26852/01234250.20

Romero, C., López, M. I., Luna, J. M., y Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers and Education*, *68*, 458–472. https://doi.org/10.1016/j.compedu.2013.06.009

Serna Alcantara, G. (2007). Misión social y modelos de extensión universitaria: del entusiasmo al desdén. *Revista Iberoamericana de Educación*, *43*(3), 1–7. https://doi.org/10.35362/rie4332324

Shi, J., Li, W., Yang, Y., Yao, N., Bai, Q., Yongchareon, S., y Yu, J. (2021). Automated Concern Exploration in Pandemic Situations - COVID-19 as a Use Case. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 12280 LNAI*. Springer International Publishing. https://doi.org/10.1007/978-3-030-69886-7_15

Sohrabi, C., Alsafi, Z., O'Neill, N., Khan, M., Kerwan, A., Al-Jabir, A., Iosifidis, C., y Agha, R. (2020). World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). *International Journal of Surgery*, *76*(February), 71–76. https://doi.org/10.1016/j.ijsu.2020.02.034

Strecht, P., Cruz, L., Soares, C., Mendes-Moreira, J., y Abreu, R. (2015). A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance. *International Educational Data*

Mining Society, Paper Presented at the International Conference on Educational Data Mining (EDM) (8th, Madrid, Spain, Jun 26-29, 2015), 392–395. http://www.educationaldatamining.org/EDM2015/proceedings/short392-395.pdf

Talavera, L., y Gaudioso, E. (2004). Mining student data to characterize similar behavior groups in unstructured collaboration spaces. *Proceedings of Workshop on Artificial Intelligence in CSCL*, 17–23.

Velazque Rojas, L., Valenzuela Huamán, C. J., y Murillo Salazar, F. (2020). Pandemia COVID-19: repercusiones en la educación universitaria. *Odontología Sanmarquina*, *23*(2), 203–205. https://doi.org/10.15381/os.v23i2.17766